

SceneFormer: Indoor Scene Generation with Transformers

Xinpeng Wang* Chandan Yeshwanth* Matthias Nießner
 Technical University of Munich



Figure 1: Shape-conditioned scenes generated by our bedroom and living room models. In each pair, the image on the left is a visualization of the input to the model that includes the shape of the floor of the room, and locations of the doors and windows. The image on the right is the generated scene, with objects added. Our models learn the desired locations of salient objects in each task - the bed and sofa, and then add other objects that augment the scene in a realistic manner. Importantly, we can generate objects on the walls and ceilings for overall realism, as compared to only objects on the floor in existing works.

Abstract

The task of indoor scene generation is to generate a sequence of objects, their locations and orientations conditioned on the shape and size of a room. Large scale indoor scene datasets allow us to extract patterns from user-designed indoor scenes and then generate new scenes based on these patterns. Existing methods rely on the 2D or 3D appearance of these scenes in addition to object positions, and make assumptions about the possible relations between objects. In contrast, we do not use any appearance information, and learn relations between objects using the self attention mechanism of transformers. We show that this leads to faster scene generation compared to existing methods with the same or better levels of realism. We build simple and effective generative models conditioned on the room shape, and on text descriptions of the room using only the cross-attention mechanism of transformers. We carried out a user study showing that our generated scenes are preferred over DeepSynth scenes 57.7% of the time for bedroom scenes, and 63.3% for living room scenes. In addition, we generate a scene in 1.48 seconds on average, 20% faster

than the state of the art method Fast & Flexible, allowing interactive scene generation.

1. Introduction

Generating virtual indoor scenes is of commercial interest to real estate and interior furnishing companies, as they can quickly visualize a furnished room and its contents without moving any physical objects. Such a room can be presented on augmented or virtual reality platforms such as a headset, allowing a user to walk through their future home and interactively modify it.

The scene generation task is: given the shape of an empty room and the locations of doors and windows, generate a set of objects and place them in the room. Each object has a category such as ‘table’ or ‘chair’, a location in the room (x, y, z), its orientation θ and a size (l, w, h). Once this sequence is generated, for each object the most relevant CAD model is retrieved from a database and placed in the scene at the predicted location. The relevance of a CAD model can be predicted based only on size, a shape descriptor [29], or other heuristics such as texture. CAD model selection reduces object collisions, accommodates special object prop-

* Equal contribution

erties such as symmetry and can ensure style-consistency across objects.

Existing methods operate on internal representations of the scene such as 2D images [27], graphs [26], and matrices [29]. Many of these works generate objects in an autoregressive manner - the properties of the $(n + 1)^{th}$ object are conditioned on the properties of the first n objects. We adopt a similar autoregressive prediction. The extraction of object and scene patterns is enabled by large object and scene datasets such as ModelNet [28], ShapeNet [8] and other human-annotated scene datasets with synthetic objects [23]. Some existing methods which require object relations to be annotated assume a fixed and manually created set of possible relations [30, 26], instead of operating on the raw scene data. Our method differs from these in that we operate directly on the raw locations and orientations of the objects without any additional information. In this way, we avoid any bias introduced by the manual selection of relations, or the heuristics used to create these relations.

Transformers [25] perform well on a variety of natural language processing tasks by treating a sentence as a *sequence* of words, and recently on images [15] and 3D meshes [17] as well. Based on the idea that a scene can be treated as a *sequence* of objects, we propose the *Scene-Former*, a series of transformers that autoregressively predict the category, location, orientation and the size of each object in a scene. We show that such a model generates realistic and diverse scenes, while requiring little domain knowledge or data preparation. In addition, we do not use any visual information such as a 2D rendering of the scene either as input or as an internal representation.

We use the cross-attention mechanism of the Transformer decoder to build *conditional* models. *Conditioning* differs from a *translation* task – in translation the output is expected to be aligned with the input as closely as possible according to a given metric, while in a conditioning task the input only guides the scene generation. We expect the output to mostly adhere to the input, in addition it can contain some information not indicated by the input. We condition separately on two kinds of user inputs. The first is the shape of the room, including the positions of doors and windows. Our shape-conditioned scenes are preferred over DeepSynth 57.7% of the time for bedroom scenes, and 63.3% for living room scenes. The second possible input is a text description of the room, such as “*There is a room with a bed and a wardrobe. There is a table lamp next to the wardrobe.*”.

Contributions

- We represent an indoor scene as a *sequence* of object properties. This converts the scene generation task into a sequence generation task.

- We leverage the self-attention mechanism of Transformers to learn relations between objects in a scene. Hence we eliminate the need for annotated relations such as “X is on Y”. Furthermore, we eliminate the need for visual information such as color, texture or object meshes.
- We generate complex scenes with objects on the wall and ceiling, and any other spatial relations in the dataset. We do this by generating the 3D object coordinates directly as integer, as opposed to earlier methods which generate coordinates in a 2D grid.
- We generate indoor scenes conditioned on unrestricted room shapes using a simple floor plan image, and then using the cross attention mechanism of the transformer on an embedding of this image. Doors and windows are represented as special tokens in the beginning of the object sequence.
- We generate realistic indoor scenes conditioned on textual descriptions of the room, and show both quantitatively and qualitatively that the generated scene respects the objects and relations mentioned in the description.

2. Related Work

Room-shape Conditional Scene Generation Previous works employ various scene representations - graphs, images, and other structures such as matrices.

Graph-based methods A natural representation of a scene is a graph [30, 15, 20], where each object is a node, and an edge is a relation between objects such as ‘is next to’ or ‘is on top of’. Features of the room such as walls, doors and windows can be represented as nodes of the graph [26]. This gives a simple method for conditioning – when the model is autoregressive on the graph and is able to generate one node at a time, the input is initialized with the required object nodes and then repeatedly expanded using the model. Such a representation lends itself well to processing with graph convolutional networks [13].

Image-based Methods A scene can be represented by a top down view of the objects, walls and floor [21, 27]. The walls and floor can be represented by binary images. Object categories, orientations and other properties can be represented by continuous or discrete numbers. This method can be used to represent arbitrary room shapes and sizes, by normalizing the room dimensions to a known maximum dimension along each axis. An image representation can take advantage of modern CNN architectures such as the ResNet [12].

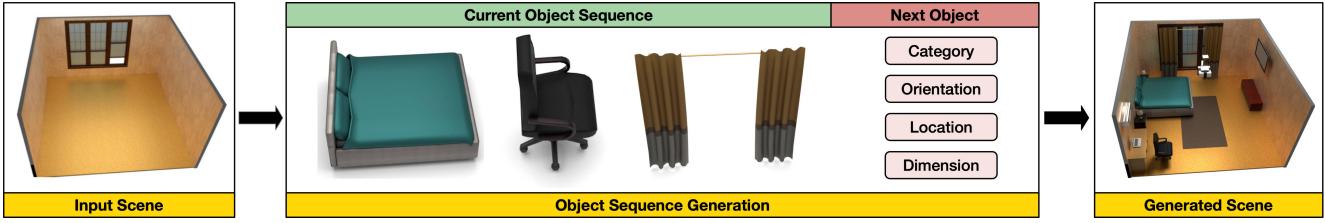


Figure 2: Shape-conditional SceneFormer: Our method takes as input a shape-condition (left). The input to the network is a binary image representing the floor, which is visualized here with the corresponding walls. Doors and windows, when present are inserted as tokens in the *Current Object Sequence*. The network sequentially generates the properties of the *Next Object*, appending these to the *Current Object Sequence*. On the right, the *Generated Scene* is visualized.

Other methods A scene can also be represented as a matrix [29] of objects in each category or a hierarchy [14] of objects. Next, an appropriate encoder-decoder or other model is used to generate an output scene in the same format. However, these methods require complex optimization and post-processing steps to obtain realistic scenes. This may be in the form of a discriminative loss, linear programming or other heuristics that are specific to the task. In contrast, we use only the cross-entropy loss for classification that is both conceptually simpler and easier to optimize.

Text Conditional Scene Generation Several previous works have addressed the task of text-conditional generation or text-to-scene translation. Text inputs have been used to create detailed partial scenes [16, 6], such as a table with several small objects on it, and then insert this into the rest of the scene. User input is required for refinement at every step, making it semi-automatic. Similarly, SceneSeer [7] and related methods [4, 22] rely on interactive user inputs. Other related works [3, 5] generate simple scenes with few objects in them, by inferring rules from a smaller scale human annotated dataset. Text2Scene [24] solves a similar task in 2 dimensions, by iteratively placing objects into an 2D image and then ensuring consistency. Intelligent Home 3D [9] tackles the related task of generating the full room layout of a house from text, and proposes a new dataset for this task. Our method differs from these in that it can generate high quality complex scenes with a large number of objects, without requiring user inputs for refinement. However, our model is still flexible enough to accept user inputs when required.

3. Method

An overview of our scene generation pipeline is shown in Fig. 2. We first discuss our data preparation steps, and then the *SceneFormer* model and its shape-conditioned and text-conditioned variants. Then we list relevant dataset and training details.

3.1. Data Preparation

We treat each scene as a sequence of objects, ordered by the frequency f_{c_i} of their categories c_i in the training set. This ordering of objects is required to produce a unique representation of each scene, upto the ordering within objects of the same class. Then the location of an object is normalized by the maximum room size, and quantized in the integer range of (0, 255) to give the new coordinates of the object (x, y, z) . Similarly the dimensions of each object; length, width and height or (l, w, h) are scaled and quantized. The orientation of the object in the floor plane of the room θ is also quantized in the range (0, 359). Hence, for each scene with objects o_i we obtain 8 sequences $(\{c_i\}, \{x_i\}, \{y_i\}, \{z_i\}, \{\theta_i\}, \{l_i\}, \{w_i\}, \{h_i\})$. We then add special *start* and *stop* tokens to each sequence to indicate the beginning and end of the sequence. Sequences are finally padded to the maximum length present in the training set.

3.2. Transformer for scene generation

Our model architecture is shown in Figure 3. We use the transformer decoder [25] to generate the sequences of object properties. For each of the four different properties, we train a separate model to predict the corresponding token of the current object. Object properties are predicted in an *autoregressive* manner - first we predict all properties of object 1 in the order category, orientation, location, dimension, then all properties of object 2 and so on.

Hence, the distribution over the sequence $\{c_i\}$ is factorized as

$$p(\{c_i\}; M_c) = \prod_{n=1}^N p(c_n | o_{<n}; M_c)$$

where the category model M_c expresses the distribution over the category of a single object c_n . The factorization of the orientation sequence is

$$p(\{\theta_i\}; M_\theta) = \prod_{n=1}^N p(\theta_n | c_{\leq n}, o_{<n}; M_\theta)$$

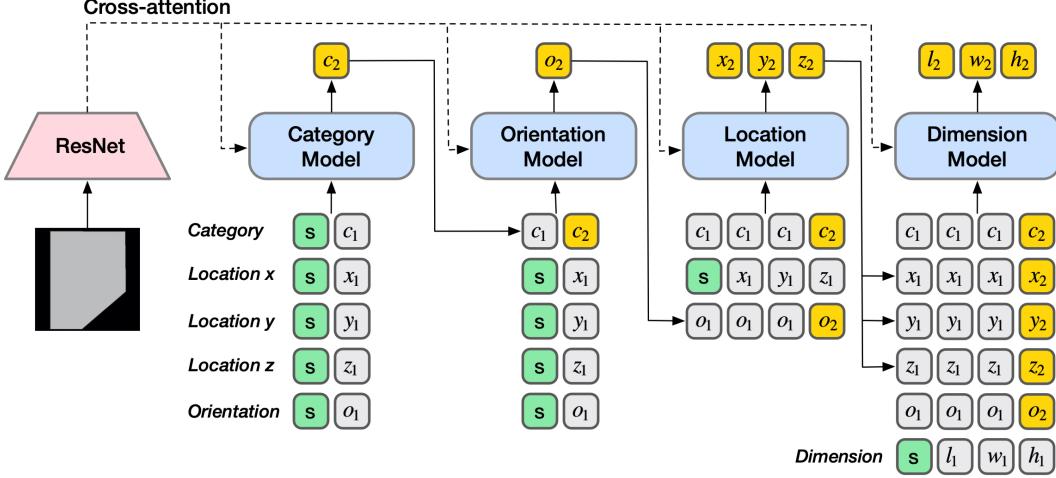


Figure 3: Shape-conditional SceneFormer model. Start tokens are shown in green, tokens from existing objects in gray and tokens corresponding to new objects in yellow. Stop tokens and padding are omitted. There are 4 kinds of sequences as input - category, orientation, location and dimension. The respective outputs of the models are appended to the existing sequence before passing it on to the next model. A model with N output tokens indicates that it is run N times, producing one token from each step.

The location and dimension models M_{loc} and M_{dim} are conditioned on tokens generated so far, for example the factorization for $\{y_i\}$ is

$$p(\{y_i\}; M_{loc}) = \prod_{n=1}^N p(y_n | c_{\leq n}, \theta_{\leq n}, x_{\leq n}, o_{\leq n}; M_{loc})$$

Hence, each model is conditioned on the output of the previous models. The category model is conditioned on all previous objects. The orientation model is conditioned on the category of the current object, as well as all other properties of all previous objects and so on. We find empirically that conditioning on the object dimensions does not improve performance, hence none of models, except the dimension model, take the dimension sequence as input. Similarly, swapping the order of location and orientation models led to unrealistic object locations. Intuitively, learning locations is the hardest task of the 4 properties considered and the location model benefits from more inputs.

Sequence Representation Each model takes multiple sequences - category, location and orientation as input. Since the location (x, y, z) and dimension (l, w, h) of each object have three values each, the input sequences for the location and dimension models are obtained by concatenating tuples of $(x_i, y_i, z_i)_i$ and $(l_i, w_i, h_i)_i$. Therefore, the other input sequences should be repeated 3 times. In order to condition on properties of the current object during training, the corresponding input sequence is shifted to the left by one token. For example, as shown in the Figure 3, the category input sequence for the orientation model is shifted towards

the left by one token, so that the orientation is generated conditioned on the category of the current object.

Embedding We use learned position and value embeddings [10] of the input sequences for each model. The position embedding indicates the position in the sequence the object belongs to. The value embedding indicates the token’s value. For the location and dimension models, we add another embedding layer to indicate whether the token is an x , y or z coordinate for location sequence, and whether the token is l , w or h for the dimension sequence. Then we combine the embeddings of all sequences by adding these together.

The output embedding is then converted to N logits with a linear layer, where N is the number of possible quantized values. Each network is trained independently using a cross entropy loss.

Inference During inference, the properties of the object are generated in the order of category, orientation, location and dimension. Once a new token is generated, the corresponding sequence is appended with the new token and given as input to the next model. The location and dimension models are run three times each, to obtain three different output tokens (x, y, z) and (l, w, h) respectively.

We use probabilistic nucleus sampling (top- p sampling) on the outputs of the category model with $p = 0.9$, and pick the token with the maximum score from the other 3 models. If any model outputs a stop token, the sequence is terminated.

3.3. Room-shape Conditional scene generation

We generate the indoor scene conditioned on an empty room, consisting of its floor shape, windows and doors. The floor shape in turn defines the positions of the walls. We take the binary image of the floor as input and feed this into an encoder consisting of a series of residual blocks [12], shown in Figure 3. We use 1-channel input images of size $[512, 512]$ and obtain feature maps of size $[16, 16, E]$, where E is embedding dimension of the transformer model. A discrete 2D coordinate embedding is added to this feature map, which is then flattened to obtain a $[256, E]$ sequence. The SceneFormer decoder then performs cross-attention on the embedded sequence. To condition generation on the locations of doors and windows, we simply insert these objects at the beginning of the input sequence. Hence, inference starts with a sequence consisting of the start token as well as tokens corresponding to the doors and windows.

We do not use an additional loss to enforce that the generated objects are within the input floor region.

3.4. Text Conditional scene generation

A room is described by a list of sentences. We pick the first 3 sentences, tokenize them and then pad the token sequence to a maximum length of 40 tokens. We then embed each word with an embedding function. In particular, we experiment with GloVe [18], ELMo [19] and BERT [11]. We obtain fixed size word embeddings with dimension d (100, 1024 and 768 respectively) using the Flair library [1], and then use a 2-layer MLP to convert from d dimensions to E dimensions, where E is the dimension of the SceneFormer embedding. For the text-conditional model, we use decoders only for the category and location models. The *decoders* for orientation and dimension models are replaced by *encoders* without cross-attention. We do not use an additional loss to align the transformer output with the input text, this relation is learned implicitly.

3.5. Object Insertion

For each generated object, we find the CAD model that is closest in size using the L2-norm over the dimensions of the object. If this causes a collision with 3D IoU upto 0.05, we reselect the next CAD model, and repeat this upto 20 times. If none of the models fit in the predicted location, we resample the object category.

3.6. Data and Training Details

We use bedrooms and living rooms from a human-created scene dataset, referred to as *GT* (ground truth) [23] and filter the bad samples as done in earlier works [26, 21] to get a total of 6351 bedrooms and 1099 living rooms. We use 50 object categories for bedroom scenes and 39 object categories for living room scenes. Rooms are augmented

with rotations from the set $(0, 90, 180, 270)$ degrees, and locations sampled uniformly from $(0, 0.5)$. We train with a learning rate of $3e^{-4}$ and apply cosine annealing with restarts after 40,000 iterations, for a maximum of 2000 epochs, using the Adam optimizer with a 0.001 weight decay and a batch size of 128. All experiments are run on an Nvidia RTX 2080 Ti GPU. Training takes approximately 4 hours for each model.

To generate textual scene descriptions, we use a heuristic method. We first extract *relations* between the objects in the scene as done in [2]. All related objects within a threshold distance of $2.5m$ are retained. In addition, objects can only be related to an object that appears earlier in the sequence. Then, we use a set of rules to generate sentences from these filtered relations. The first sentence mentions upto the first 3 objects in the room’s category sequence $\{c_i\}$. Then we iterate over all objects except the first in the sequence, and with probability 0.3 describe an object o_i using its relation to another object $o_j, j < i$ which has already been described. Each relation is a tuple (o_i, rel, o_j) , where rel is the relation type.

4. Results and Evaluation

4.1. Qualitative Results

4.1.1 Room-shape conditional scene generation

Scene Synthesis Comparison Figure 5 shows the bedroom scenes generated by our model compared to the scenes from DeepSynth [27] conditioned on the same room. It shows that we generate more complex scenes in terms of object categories and object relations. Previous works such as DeepSynth [27], FastSynth [21] and PlanIT [26] perform image-based scene generation and therefore can only generate objects on the floor, or objects that are supported by a plane such as laptop. We are able to generate objects on the walls, such as air conditioner, wall lamp, and on the ceiling - chandelier and ceiling fan. Object relations are learnt correctly learnt - the television is placed opposite the bed or sofa, curtains are placed on the window at the correct height, bed stands are placed on either side of the bed, and so on.

Scene Completion We show the effectiveness of our method on the scene completion task in Fig. 6. Given an incomplete scene with a few objects, our model can add relevant and missing objects to the scene to complete it. As a result of training on sorted sequences, the model first generates large and frequent objects before generating small and infrequent objects in a greedy manner. Due to this, stopping generation at an intermediate stage results in a realistic scene as well and the user can interactively choose how complete the scene must be.

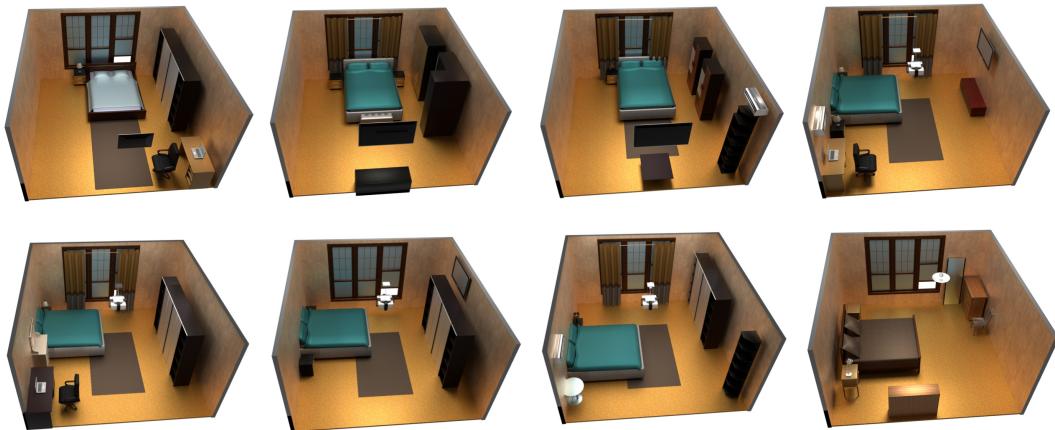


Figure 4: Scene diversity: the ground truth scene is shown in the bottom right. The other 7 scenes are generated conditioned on the same room shape. Combinations such as table-chair are both generated, or not generated at all. Diversity of locations is also obtained as a consequence of rotation and jittering augmentations, from a relatively small dataset.



Figure 5: Bedroom and living room scenes generated by DeepSynth and our method, conditioned on the same room

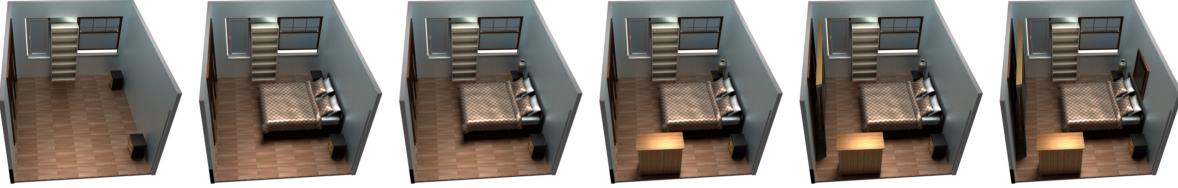


Figure 6: Scene completion sequence: iterative process of adding objects; the leftmost is the input to the model and the others are obtained sequentially by adding objects

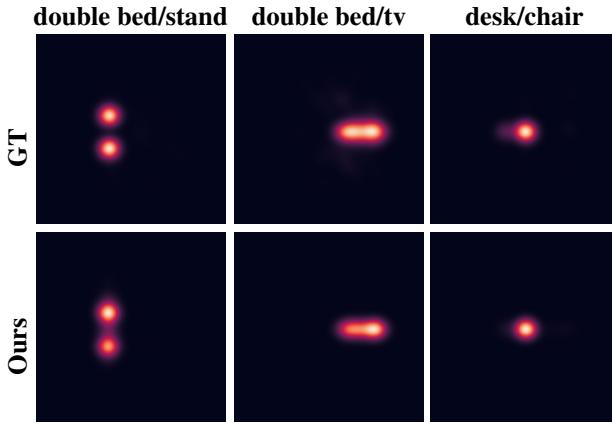


Figure 7: Object location heatmap for pairs of objects in bedroom scenes, ground truth and ours. Common relations are captured, while introducing new relations through augmentation.

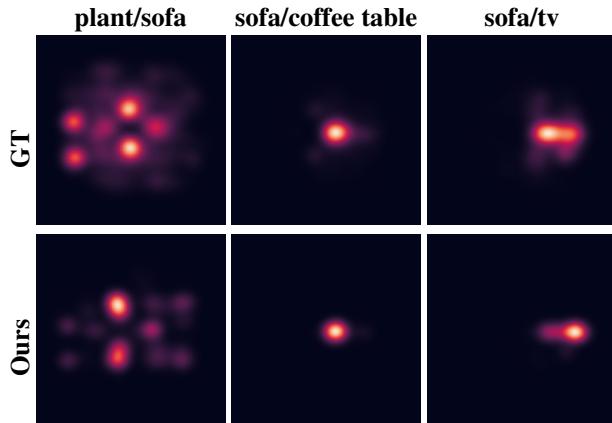


Figure 8: Object location heatmap for pairs of objects in living room scenes, ground truth and ours

Scene Diversity Various generated scenes conditioned on the same input room shape are shown in Fig. 4. We are able to generate different sets of objects and object orientations that are consistent with each other.

Object Category Heatmap Figure 7 and Figure 8 show the heat-map of relative location between objects. It evaluates whether the model learns object relation correctly from the data.

4.1.2 Text conditional scene generation

Text-conditioned scenes generated by our model and the corresponding descriptions are shown in Figure 13. Generated scenes largely respect the input sentence in terms of object categories and relations, and have additional objects to improve the completeness of the scene.

4.2. Quantitative results

4.2.1 Room-shape conditional generation

Perceptual study We conducted a perceptual study as done by earlier works [26, 21] where users are shown 24 pairs of images. Each pair contains one image generated using DeepSynth [21] or from *GT* [23], and one scene generated by our method. The user has to select the one which looks more realistic. We use 4 pairs of images as vigilance tests. The responses of users who do not pass all the tests are discarded. The study was taken up by a total of 30 users through Amazon Mechanical Turk. Table 1 shows the results of this study.

Room Type	Ours vs. DeepSynth	Ours vs. GT
Bedroom	57.7 ± 5.6	50.2 ± 6.6
Living	63.3 ± 5.3	48.5 ± 6.6

Table 1: Perceptual User Study of our generated images vs. the *GT* dataset and DeepSynth. The metric is the percentage of image pairs where our image is preferred.

Speed comparisons We measure the inference time of our shape-conditioned model, and compare with existing models in Table 2. Our model is an order of magnitude faster than PlanIT and Deep Synth, and 20% faster than Fast & Flexible.

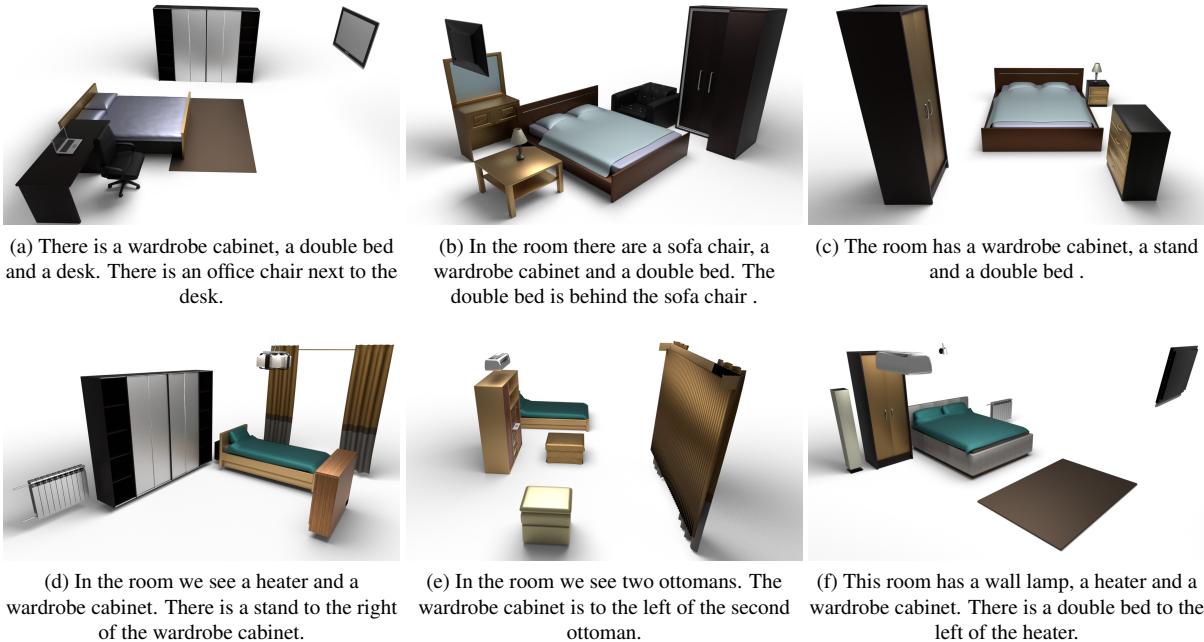


Figure 9: Text-conditioned bedroom scenes shown with the corresponding text inputs. Since the room shape is not given as input, objects are placed within a room-shape prior learnt from the ground truth scenes. Important relations such as table-chair, bed-television and sofa-television are captured by the model.

Method	Time (s)
DeepSynth [27]	240
Fast & Flexible [21]	1.85
PlanIT [26]	11.7
Ours	1.48

Table 2: Comparison of Inference Time

4.2.2 Text-conditional Generation

Perceptual study We performed a perceptual study, showing 40 users 34 pairs of text and images each, asking them to answer two questions for each pair - how realistic the generated scene is, and how closely the generated scene matches the input text. Responses are given on a scale of 1 to 7, with 1 being being a poor match and 7 being a good match. Our bedroom scenes obtained a realism score of 4.61 ± 1.84 , and a match score of 4.38 ± 1.73 , indicating that our text conditional model can both generate realistic scenes and capture the objects and relations mentioned in the text description.

Category Accuracy We measured the fraction of objects mentioned in the scene description that are presented in the generated scene, shown in Table 3. We compare against 2 baselines - *Uniform*, by sampling uniformly from all categories and the stop token, and *GT* by sampling from the

ground truth object category distribution, with the stop token having the average frequency of all objects.

Model	Bedroom	Living Room
Uniform	23.70	27.89
GT	38.64	44.11
Ours (GloVe)	58.10	59.35
Ours (ELMo)	76.59	63.31
Ours (BERT)	84.83	68.81

Table 3: Category accuracy of text-conditioned scene generation (percentage)

5. Conclusion and Future Work

We presented the SceneFormer, a combination of transformer models that generates realistic indoor scene layouts with minimal data preparation, and performs fast inference. This can enable interactive scene generation from partial scenes. We eliminate the need for expensive visual information such as an image or mesh. Our model can serve as a general framework for scene generation: a different task can be solved by changing the set of object properties or conditioning inputs. Future work can consider the 3D mesh of each object to obtain global style consistency. Our model can be used for 3D reconstruction of an indoor scene given

its 3D scan as a prior. Since we perform room shape and text conditioning separately, it would be interesting to see scene generation jointly conditioned on both inputs.

6. Acknowledgement

This work was supported by a TUM-IAS Rudolf Mößbauer Fellowship, the ERC Starting Grant Scan2CAD (804724), and the German Research Foundation (DFG) Grant Making Machine Learning on Static and Dynamic 3D Data Practical.

References

- [1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018. 5
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 5
- [3] Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. Text to 3D scene generation with rich lexical grounding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 53–62, Beijing, China, July 2015. Association for Computational Linguistics. 3
- [4] Angel Chang, Manolis Savva, and Christopher D Manning. Interactive learning of spatial knowledge for text to 3d scene generation. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 14–21, 2014. 3
- [5] Angel Chang, Manolis Savva, and Christopher D Manning. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2028–2038, 2014. 3
- [6] Angel Chang, Manolis Savva, and Christopher D Manning. Semantic parsing for text to 3d scene generation. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 17–21, 2014. 3
- [7] Angel X Chang, Mihail Eric, Manolis Savva, and Christopher D Manning. Sceneseer: 3d scene design with natural language. *arXiv preprint arXiv:1703.00050*, 2017. 3
- [8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2
- [9] Qi Chen, Qi Wu, Rui Tang, Yuhang Wang, Shuai Wang, and Mingkui Tan. Intelligent home 3d: Automatic 3d-house design from linguistic descriptions only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12625–12634, 2020. 3
- [10] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 4
- [11] 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 5
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5
- [13] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 2
- [14] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019. 3
- [15] Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B Tenenbaum. End-to-end optimization of scene layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3754–3763, 2020. 2
- [16] Rui Ma, Akshay Gadi Patil, Matthew Fisher, Manyi Li, Sören Pirk, Binh-Son Hua, Sai-Kit Yeung, Xin Tong, Leonidas Guibas, and Hao Zhang. Language-driven synthesis of 3d scenes from scene databases. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018. 3
- [17] Charlie Nash, Yaroslav Ganin, SM Eslami, and Peter W Battaglia. Polyscen: An autoregressive generative model of 3d meshes. *arXiv preprint arXiv:2002.10880*, 2020. 2
- [18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5
- [19] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 5
- [20] Pulak Purkait, Christopher Zach, and Ian Reid. Learning to generate new indoor scenes. *arXiv preprint arXiv:1912.04554*, 2019. 2
- [21] Daniel Ritchie, Kai Wang, and Yu-an Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6182–6190, 2019. 2, 5, 7, 8

- [22] Manolis Savva, Angel X Chang, and Maneesh Agrawala. Scenesuggest: Context-driven 3d scene design. *arXiv preprint arXiv:1703.00061*, 2017. 3
- [23] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 2, 5, 7
- [24] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019. 3
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3
- [26] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2, 5, 7, 8
- [27] Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 5, 8
- [28] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2
- [29] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020. 1, 2, 3
- [30] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7384–7392, 2019. 2

Appendix

A. Model details

Further details on our model are given here.

Transformers We use these common hyperparameters across all 4 models, unless otherwise specified.

- Maximum number of objects in the scene: 50
- Embedding dimension: Dimension of input and output of the transformer, 256 (1024 in location model)
- Dimension of transformer activations: 256 (1024 location model)
- Number of transformer attention heads: 8
- Number of transformer blocks: 8

ResNet for Shape Conditioning We use a ResNet with

- 1 input channel
- 3 blocks, having 3, 3, and 4 layers respectively
- output dimension 256

B. Shape-conditioned model

B.1. Results

Additional scenes generated by our shape-conditioned model are shown in Fig. 10

B.2. Object category heatmaps

Additional object pair heatmaps from our generated scenes and the ground truth are shown in Fig. 11

B.3. Survey interface

The web interface used to conduct the survey on user preference between our scenes, and scenes from DeepSynth or from the dataset, is shown in Fig. 12.

C. Text-conditioned model

C.1. Scene description generation

We describe the method used to generate scene descriptions.

Relation Generation First we generate a set of relations between objects in the scene. Given a sequence of objects $\{o_1, o_2, \dots\}$, relations r_{ij} are generated such that $i < j$. Based on the bounding boxes of o_i and o_j , the relation type is classified as one of these - *on*, *above*, *surrounding*, *inside*, *right of*, *left of*, *behind*, *in front of*. At the same time, the distance between the bounding box centers is computed.

Description Generation The first 2 or 3 objects in the sequence are described in the first sentence, this choice is made uniformly. We use a fixed set of starting phrases for the sentence – *The room has*, *In the room there are*, *The room contains*, *This room has*, *There are*, *In the room we see*, and then add a list of the first 2 or 3 objects. Repeated objects are mentioned as the count followed by the object category.

The remaining sentences describe objects in relation to any object that has already been described. We iterate through objects that have not been described, and choose to describe any of them with a probability of 0.7. Then, we pick one of the relations of this object to an object that has already been described, such that the distance between them is less than the threshold of 2.5 meters. This choice of relation is done uniformly.

Similar to the earlier procedure, ordinal prefixes such as *second* and *third* are added when the new object is not the first in its category to be described. In addition, objects of the same category are not related together; we reject sentences such as “There is a second table next to the first table”.

Next, we choose the appropriate article for each object (*a*, *an*, *the*), and choose a sentence template based on the relation between the objects, such as – *There is a table next to the chair*, *There is a sofa to the right of the wardrobe cabinet*. These sentences are appended together to form the final scene description.

C.2. Results

C.3. Survey interface

The web interface used to conduct the survey of generated scene quality is shown in Fig. 14 and Fig. 15. The introduction explains the task to the user, and each scene has 2 questions, with answers in the range of 1–7.



Figure 10: Bedrooms and living room scenes generated by our method. In each pair, the left image is the conditioning input with the shape of the room and locations of doors and windows. The right image is the generated scene. All objects are generated within the boundaries of the room, and such that they do not overlap doors and windows. In addition, meaningful relations between different object categories are learnt by the model.

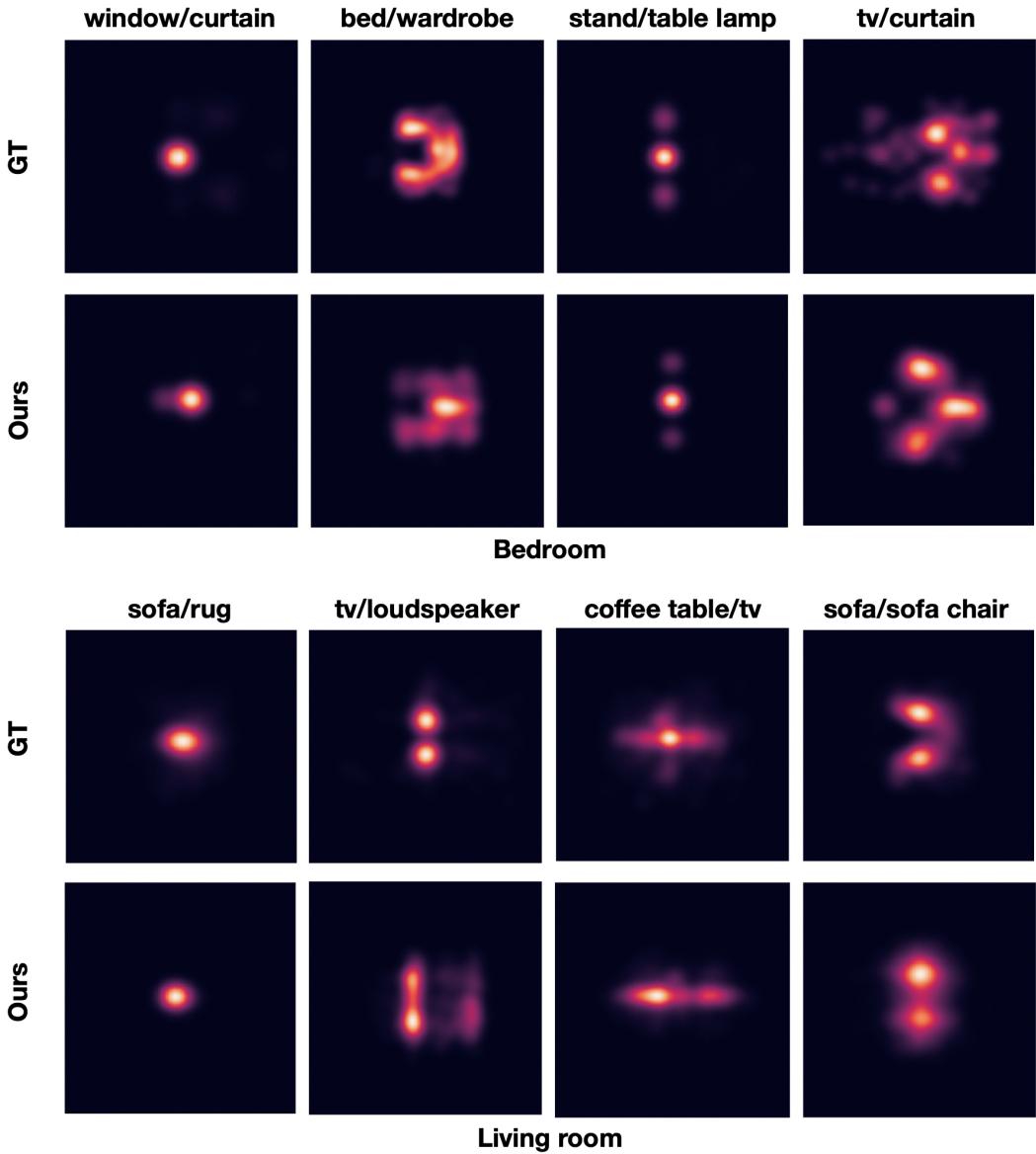


Figure 11: Additional heatmaps comparing our generated scenes with those from the ground truth. On top are displayed the pairs of object categories in the heatmap. We see that our model captures the GT distribution well, while adding novel relations in some cases. This is crucial for generating diverse scenes and avoiding overfitting.

* Please choose the scene that is more realistic to you:

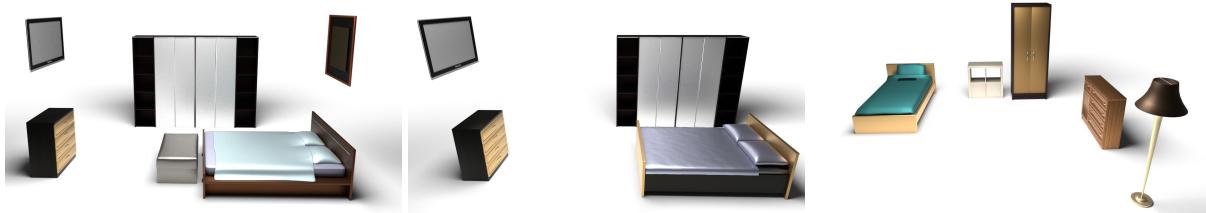


1a



1b

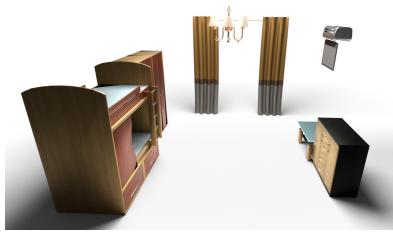
Figure 12: A representative survey question from the survey for shape-conditioned images. Each question contains a pair of images. One is from our method, the other is from the dataset or generated by DeepSynth. The user has to choose the one which looks more realistic. Here we use the top down view of the scene as down in earlier works.



(a) The room has an ottoman and a wardrobe cabinet.

(b) In the room we see a wardrobe cabinet, a double bed and a television. There is a dresser to the left of the television.

(c) The room has a shelving and two wardrobe cabinets.



(d) The room contains a bunker bed and a wardrobe cabinet.

(e) The room has two stands and a double bed. The double bed is to the right of the second stand. There is a dresser to the right of the double bed.

(f) This room has a wardrobe cabinet and a double bed. There is a desk to the left of the double bed.

Figure 13: Text-conditioned bedroom scenes shown with the corresponding text inputs. Since the room shape is not given as input, objects are placed within a room-shape prior learnt from the ground truth scenes. Important relations such as table-chair, bed-television and sofa-television are captured by the model.

Bedroom scenes and their text descriptions

E

Hello!

In this survey you will be shown 35 pairs of images and sentences. The sentence describes the contents of a bedroom, and the image shows the contents of the same room.

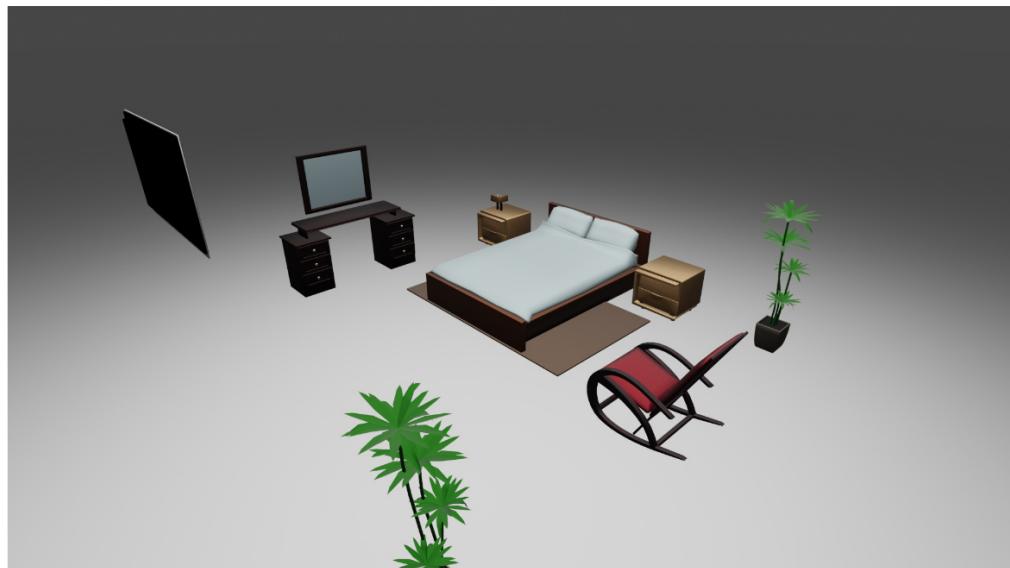
For each pair, you have to answer 2 questions by rating the scene

1. How realistic is the scene in the image?
2. How closely does the scene match the text description, by considering
 - (a) are the objects mentioned in the description included in the scene?
 - (b) are the described relationships between the objects correct in the scene?
 - (c) overall, does the scene fit the description?

Next

Figure 14: Introduction to the perceptual study of text-conditional samples

* In the room there are two plants and an armchair . The second plant is in front of the first plant . The armchair is to the right of the second plant . There is a coffee table surrounding the second plant .



Strongly disagree

1

2

3

4

5

6

7

Strongly agree

The image of the room matches the given description

The image of the room is realistic

Next Question

Figure 15: A representative question from the perceptual study on text-conditional samples