

# Learning Object Placement Programs for Indoor Scene Synthesis with Iterative Self Training

ADRIAN CHANG, Vision Systems Inc., USA and Brown University, USA

KAI WANG, Brown University, USA

YUANBO LI, Brown University, USA

MANOLIS SAVVA, Simon Fraser University, Canada

ANGEL X. CHANG, Simon Fraser University, Canada

DANIEL RITCHIE, Brown University, USA

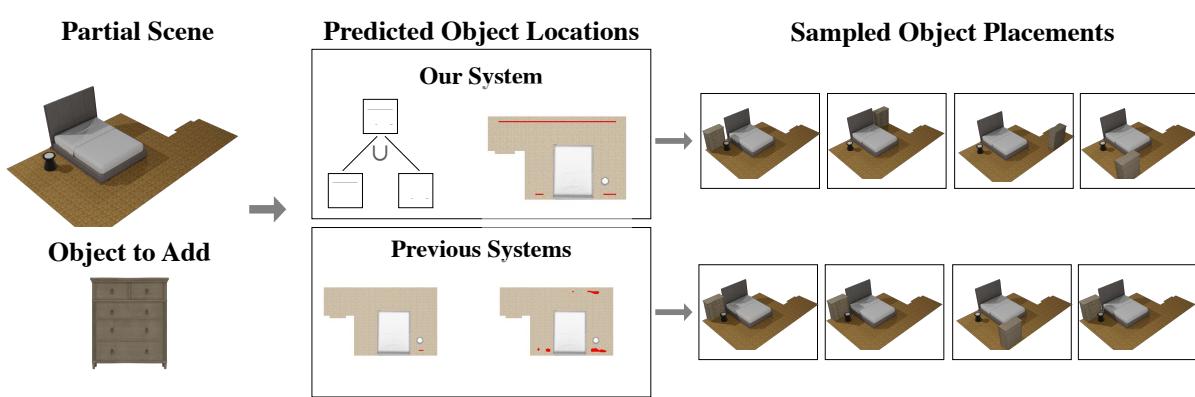


Fig. 1. Autoregressive indoor scene synthesis systems perform scene synthesis by iteratively placing objects. Given a partial scene and an object to add, they predict possible places that object can go. We visualize the placement options produced by our and previous systems for the inputs shown on the left. Note how previous systems predict incomplete distributions while our system produces a variety of placement options.

Data driven and autoregressive indoor scene synthesis systems generate indoor scenes automatically by suggesting and then placing objects one at a time. Empirical observations show that current systems tend to produce incomplete next object location distributions. We introduce a system which addresses this problem. We design a Domain Specific Language (DSL) that specifies functional constraints. Programs from our language take as input a partial scene and object to place. Upon execution they predict possible object placements. We design a generative model which writes these programs automatically. Available 3D scene datasets do not contain programs to train on, so we build upon previous work in unsupervised program induction to introduce a new program bootstrapping algorithm. In order to quantify

our empirical observations we introduce a new evaluation procedure which captures how well a system models per-object location distributions. We ask human annotators to label all the possible places an object can go in a scene and show that our system produces per-object location distributions more consistent with human annotators. Our system also generates indoor scenes of comparable quality to previous systems and while previous systems degrade in performance when training data is sparse, our system does not degrade to the same degree.

**CCS Concepts:** • Computing methodologies → Scene understanding; Unsupervised learning; Neural networks; Statistical relational learning.

**Additional Key Words and Phrases:** Indoor Scene Synthesis, Program Induction, Domain Specific Language, Transformers,

## ACM Reference Format:

Adrian Chang, Kai Wang, Yuanbo Li, Manolis Savva, Angel X. Chang, and Daniel Ritchie. 2018. Learning Object Placement Programs for Indoor Scene Synthesis with Iterative Self Training. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 21 pages. <https://doi.org/XXXXXX.XXXXXXX>

Authors' Contact Information: Adrian Chang, anchang2001@gmail.com, Vision Systems Inc., Providence, Rhode Island, USA and Brown University, Providence, Rhode Island, USA; Kai Wang, kwang.ether@gmail.com, Brown University, Providence, Rhode Island, USA; Yuanbo Li, yuanbo\_li@brown.edu, Brown University, Providence, Rhode Island, USA; Manolis Savva, msavva@sfsu.ca, Simon Fraser University, Burnaby, British Columbia, Canada; Angel X. Chang, angelx@sfsu.ca, Simon Fraser University, Burnaby, British Columbia, Canada; Daniel Ritchie, daniel\_ritchie@brown.edu, Brown University, Providence, Rhode Island, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM 1557-7368/2018/8-ART111  
<https://doi.org/XXXXXX.XXXXXXX>

## 1 Introduction

Humans spend a majority of their time indoors, and our media and technical applications reflect this lived experience. Augmented / virtual reality (AR/VR), movies, and video games all rely on virtual indoor spaces. Computer vision and robotics researchers working on tasks such as scene understanding and robotic navigation also rely on 3D scenes [9, 21]. Manually authoring indoor scenes however is

laborious, time consuming, and often requires expert knowledge – a process which does not scale.

Indoor scene synthesis systems which can automatically generate indoor scenes have emerged as a solution to these demands. While these methods take on a variety of formulations, we focus on data-driven and autoregressive systems [20, 23]. Data-driven methods learn to generate scenes similar to those in an example dataset. This formulation is useful in instances where a user wants to generate scenes that mimic design patterns present in a set of exemplars. Autoregressive systems perform scene synthesis by either iteratively placing objects in an empty scene until it is complete, or scene completion from a given partial scene via the same methodology. For applications in interactive scene generation [1, 24], these systems can suggest both the next object to place and where in the scene it should be placed.

Empirical observations of next object location distributions predicted by data-driven autoregressive systems reveal their tendency to become incomplete with respect to human-authored placement rules. For example, a distribution which only places a bed in the corner of an empty room is incomplete since one could place it along any of the walls. This flaw both limits the diversity of possible final scenes for scene synthesis and completion tasks, and limits their overall usefulness in an interactive application. Data-driven systems overfit to particular object placements seen during training. When 3D scene data is sparse, this effect is amplified as there are not enough examples to learn placement rules effectively.

Standard solutions to reducing overfitting in scene synthesis systems are unreliable and come at a cost. Scaling the amount of scene data a system trains on can help an autoregressive system learn a more complete distribution, but this solution is expensive. Neural networks are also known to represent the most common inputs over the rare [3] so these models might still miss placement modes. Model regularization and stopping training early is another option, but in practice it is hard to balance mode coverage and specificity. Better mode coverage often comes at the cost of decreased final scene quality.

Humans craft their indoor environments to reflect their needs, desires, and tastes – a tv stand in front of a couch, a nightstand within arms reach of a bed, a wardrobe against a wall. Object-to-object and object-to-room relationships dictate the rules which guide object arrangements. A neural-network based system encodes these rules implicitly making them hard to control and align with our intuitions. Symbolic constructs such as programs can succinctly represent these rules. Their structured representation makes it easier to incorporate prior knowledge, edit the rules they represent, and ensure desirable characteristics such as continuity and completeness upon execution. We hypothesize that learning a symbolic representation, such as a Domain Specific Language (DSL) for specifying functional constraints can produce more complete per-object location distributions.

We propose a new approach to indoor scene synthesis. Our system is autoregressive, placing objects one at a time. Rather than task a generative model with predicting possible object placements however, we instead ask it to predict a DSL program. We design this DSL with functional constraints that explicitly represent human activity and inter-object relationships. Programs from this DSL are

relational layout programs which, given a partial scene and object to place, produce a mask representing all the possible object locations. We incorporate this language into a learning-based framework to learn how to automatically infer programs from partial scenes and query objects. Our system addresses the problem of incomplete next object location distributions that previous systems suffer from.

We use a transformer-based generative model to generate programs automatically. Available 3D scene data contain no “ground truth” programs which can supervise said model so we introduce an iterative self-training scheme inspired by prior work in unsupervised visual program inference [12] that improves the performance of our system.

To quantify the empirical observations mentioned previously we also introduce a new evaluation procedure which captures how well a system models per-object location distributions. We ask human annotators to label all the possible places an object can go in a scene and compare this rule against one produced by the system in question. We show how our system produces per-object location distributions that are more consistent with human annotators. While other systems show consistent degradation in per-object location modeling with less scene data, our system does not degrade in performance to the same degree.

In summary our contributions are:

- A new approach to indoor scene synthesis where we predict a relational layout program from a given partial scene and object to place, and execute that program to predict possible object placement locations
- A bootstrapped self-training algorithm that improves our system’s performance
- A new evaluation procedure that evaluates a system’s ability to model per object location distributions

## 2 Related Work

*Indoor Scene Synthesis.* Before the existence of large indoor scene datasets [7] and 3D deep learning algorithms, researchers approached indoor scene synthesis with explicit rules such as statistical relationships between objects [39], programmatically-defined constraints [38], design principles [18], or heuristics for human activity [6, 8]. Similarly, our DSL also encodes object relationships and human activity explicitly.

Deep learning methods enabled end-to-end and differentiable systems to learn scene priors automatically from large scene datasets. A scene graph is a popular representation with works using a graph neural network [11, 31], recursive neural network [15] or diffusion model [28] to learn these priors. Another line of work takes an image based approach and operates over the top down view of the scene with CNNs [23, 32]. Recent transformer based approaches [19, 20, 33] have also found success working directly with the 3D bounding box information of objects in the scene. Our system leverages recent advances in 3D deep learning, using the transformer architecture as the backbone for scene generation. Instead of having the network directly predict object placements however, our generative model writes programs defined by our DSL.

Another line of work is zero-shot indoor scene synthesis. These methods do not require training on 3D scene datasets and instead

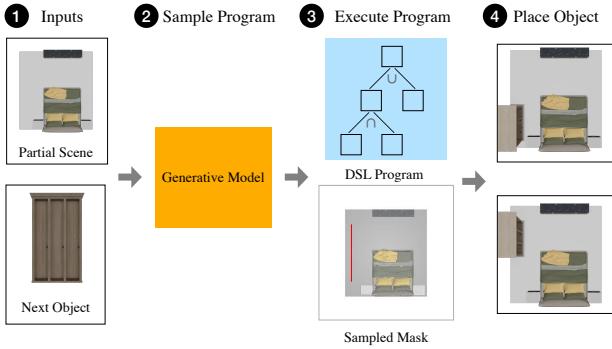


Fig. 2. Inference Pipeline: Our system is autoregressive, placing objects one at a time. Given a (1) partial scene and query object we (2) sample a DSL program from a generative model. (3) Upon execution, this produces a binary mask of possible centroid locations for the query object in the scene. (4) Sampling this mask produces valid object locations of the query object.

rely on latent scene knowledge embedded within Large Language Models (LLMs) to generate scenes. Holodeck [37] generates 3D environments for embodied AI training by sampling spatial relational constraints from an LLM and then optimizing the layout with a physical solver. Another work designs a LLM-based program synthesizer whose programs are specified by a DSL [2]. This DSL is a scene description program that specifies the objects and their spatial layout. We also generate DSL programs whose primitives are relational constraints, but our programs specify possible object placements rather than an entire scene. Our work also requires 3D scene data to train on and aims to match the distribution of input scenes rather than generate any indoor scene zero-shot.

**Visual Program Inference.** Visual Program Inference (VPI) aims to automatically infer programs that explain visual data [22]. If the visual data of interest comes with ground truth programs, supervised learning is an obvious option [34–36]. In most domains however, programs for visual data are not readily accessible. Unsupervised learning, and in particular bootstrapping [5, 16], is one option for extracting and improving programs from visual data.

Bootstrapping methods search for “good” programs, retrain on these programs, and then repeat. PLAD [12] groups different bootstrapping methods under a single conceptual framework and applies it to VPI on 3D and 2D shapes. The method sources programs from both a recognition model that matches input shapes with programs as “pseudo-labels” and pairs of synthetic shapes and their programs for an approximation of the target distribution. Later works which built on the PLAD framework [10, 13] searched for new programs by editing existing ones. SIRI [10] used domain specific operations to edit a subset of programs at a time for distributional stability. Our algorithm is an instance of PLAD and like SIRI we use domain-specific editing operations.

### 3 Method Overview

The following steps constitute our full system. Figure 2 shows our inference pipeline and Figure 3 shows how we train our system.

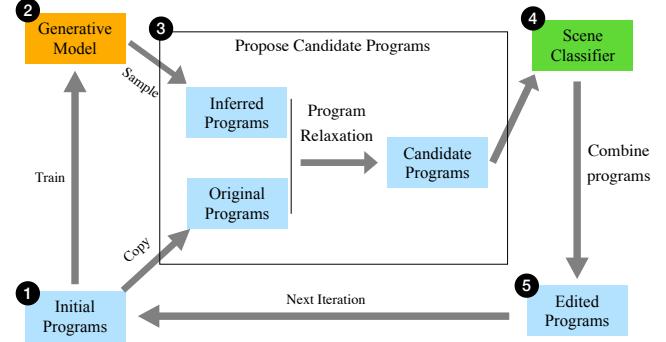


Fig. 3. We develop a program bootstrapping algorithm which discovers programs automatically from scene data. (1) We start by extracting programs with geometric heuristics and then training our (2) model on these initial programs. (3) We propose new programs by deleting constraints from both the inferred and original programs. (4) These candidate programs are then filtered with a scene real fake classifier to remove “bad” programs. (5) Domain specific operations combine “good” candidate programs together, and insert them back into the training set.

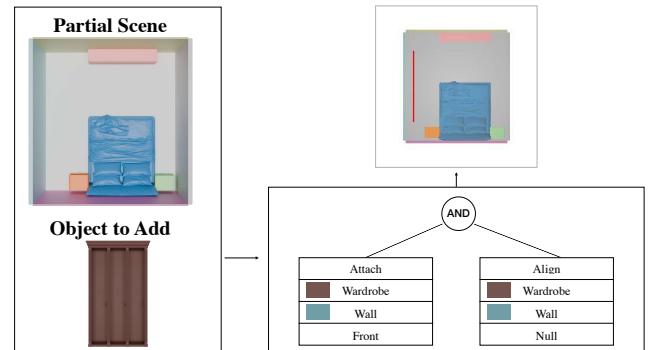


Fig. 4. Program Example: Given a partial scene and object to add, our DSL program outputs a binary mask representing possible placements of that object. Programs take on the structure of Constructive Solid Geometry (CSG) trees where each leaf node is a functional constraint that describes object function. Upon execution, these constraints produce binary masks which are combined according to the structure of the tree.

**Defining a DSL to describe object placement distributions.** To represent object placement distributions in semantically meaningful terms and align placement rules closer to ones a human might write, we introduce a domain specific language (DSL). Section 4 describes this language and its motivations in more depth.

**Automatic scene synthesis with programs.** To perform scene completion and scene synthesis we design a generative model which generates programs automatically. Section 5 describes this model.

**Program bootstrapping.** Current indoor scene datasets do not contain programs to train on, so we introduce a program bootstrapping algorithm to discover these programs and boost system performance. Section 6 describes this algorithm.

## 4 Language Design

Our DSL programs take as input a partial scene and the next object to place. They then output a binary mask representing possible centroid locations of the query object. These programs take on the same structure of a Constructive Solid Geometry (CSG) tree, but instead of operating over a continuous 3D space, the programs operate over 2D masks. Leaf nodes of this tree are functional constraints that explicitly represent human activity and inter-object relationships. When executed these constraints produce binary masks.

Binary masks are discretized along 3 dimensions. The three dimensions represent the width of the room, the height of the room, and the possible orientations of the object to place respectively. This third dimension of the mask is necessary because the validity of an object's centroid position is dependent on its orientation. The orientation of an object dictates the centroid locations the program executor predicts. We represent the orientation of an object as its rotation about the up axis of the room and snap it to one of the cardinal directions (N, E, S, W). Most objects in 3DFRONT [7] are axis aligned, so our language can model most object placements. Incorporating additional degrees of freedom for more complex scenarios is an avenue for future work.

Figure 4 shows an example program, its inputs, and its outputs.

### 4.1 Design Goals

We hypothesize that forcing our system to explain object placements with logical rules and semantically meaningful terms will produce placement rules more consistent with human intuition. Humans would likely describe object placement rules in terms of higher order concepts such as alignment, attachment, or reachability. They might also then use boolean statements to specify how those concepts should interact with each other. For example, one might describe a nightstand as both being **attached** to the side of a bed and **aligned** in its forward facing direction. A wardrobe could be **attached** to the left wall **or** the right wall. Programs excel at representing higher-order concepts and CSG is a straightforward choice for translating boolean statements to the visual domain. We are also inspired by previous work [4, 25] in inferring and editing CSG programs from input geometry as we want a representation conducive to visual program induction.

### 4.2 Constraint Specification

Functional constraints fall under two categories. Location constraints (**attach** and **reachable\_by\_arm**) assume nothing about an object's orientation and predict an object's possible centroid locations. Orientation constraints (**align** and **face**) constrain the possible orientations of the object within a scene in addition to the location.

5 total directions are specified in the language (*Up, Down, Left, Right, Null*) and all directions are specified within the local coordinate frame of the reference object. The *Null* direction exists because orientation constraints do not need a direction specification.

Visualizations of each constraint alongside further implementation details are shown in the supplemental materials.

- **attach(query object, reference object, direction)**: Constrain the possible centroid locations of the query object to be

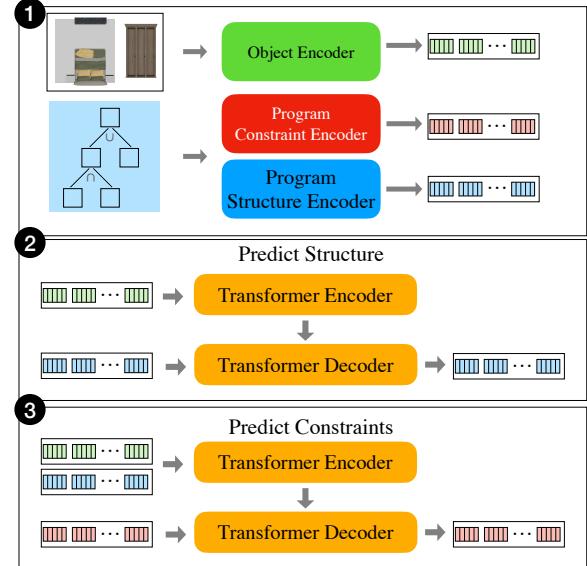


Fig. 5. Our generative model takes as input a partial scene and object and predicts a program written in our DSL. We formulate this modeling task as a seq-2-seq problem. (1) We first vectorize and then embed both the input objects and program. The structure of the program tree and the constraint attributes are embedded as separate sequences. (2) Our first transformer encoder decoder pair predicts the structure of the program from the input objects. (3) Our second transformer encoder decoder pair predicts the constraint attributes from the object and structure embeddings.

within 15 centimeters of the reference object in the direction specified

- **reachable\_by\_arm(query object reference object, direction)**: Constrain the possible centroid locations of the query object to be between 15-60 centimeters of the reference object in the direction specified. The reference object must also hold humans (i.e. bed, chair).
- **align(query object, reference object)**: Constrain the possible orientation of the query object such that it points in the same direction as the reference object.
- **face(query object, reference object)**: Constrain the possible locations of the query object such that it points toward the reference object. Evaluate this for every possible orientation.

Executing a program will execute each constraint in the tree and then combine the masks accordingly. We apply a post-processing step that removes placements of the query object which intersect with other objects in the scene beyond a specified threshold.

## 5 Generating programs

The design of our generative model takes inspiration from previous work which use partial program intermediates to aid in program synthesis [5, 26]. These works use a partial programs to express the high-level structure, but leaves holes for low level implementation details. In similar fashion, we predict the high level topological structure of our program first with one network, and then ask a second network to fill in the low level instantiation details. This

approach helps reduce overfitting to particular structure and parameter combinations.

### 5.1 Overview

We treat program synthesis as a sequence to sequence (seq-2-seq) translation task. The input sequence are the objects in the room and an object to place. The output or target sequence is the program. We train two transformer [29] encoder-decoder pairs in a two-pass approach. The first encoder-decoder pair takes in object encodings and outputs the structure or topology of the target program. The second encoder-decoder pair takes as input the source and target sequences of the first pass concatenated together and outputs the attributes of all the program’s constraints as a flattened list. Figure 5 shows the network architecture.

### 5.2 Object Encoding

Objects are represented by their bounding box with attributes category, size, position, orientation, and whether it holds humans  $o_i = \{t_i, s_i, p_i, o_i, h_i\}$ . The category  $t_i$  is an integer id.  $s_i, p_i \in \mathbb{R}^2$  (height of objects are not considered and all objects are considered grounded). The orientation  $o_i \in \mathbb{R}$  of the object is the rotation of the object about the up vector. Holds\_humans  $h_i \in [0, 1]$  is a binary flag of whether the object’s purpose is to hold a human.

The object encoder encodes object bounding boxes into an embedding vector  $\in \mathbb{R}^d$ . A learned embedding of the object category is concatenated to the raw values of the other attributes and passed through an MLP. Instead of encoding the floor plan, and implicitly the walls, as a single object, we encode each wall segment as its own object. This is due to the difficulty of encoding all the pertinent information about a floor plan into a single feature vector, and giving the model access to that information in a salient manner. The model struggles to infer algebraic quantities such as object to wall distances, especially for floor plans with non-convex geometry.

### 5.3 Program Encoding and Decoding

We represent programs as two separate sequences. One sequence represents the topology or structure of the program tree, and the other represents the attributes of each constraint in the program. We flatten out the program’s tree structure with prefix notation. We embed this “structure” sequence with per token learnable embeddings.

Each constraint takes the form (constraint type, query object index, reference object index, direction) or  $(c_j, q_j, r_j, d_j)$ . Constraint attributes are concatenated into a sequence following inorder traversal (left subtree to root to right subtree). The constraint type and direction receive per token learnable embeddings. Tokens which represent the query or reference object index use their respective object embeddings generated by the object encoder.

For tokens of fixed vocabulary length such as program structure, constraint type, and direction, an MLP head is enough to decode their tokens, but when choosing the reference object index the vocabulary is variable length as the number of objects in a scene is variable. To address this, we pass each reference object head through an MLP to form a pointer embedding  $v_j \in \mathbb{R}^d$  [30]. For a matrix of

object embeddings  $X \in \mathbb{R}^{txd}$  where  $t$  is the number of objects in the scene, we compute the reference object index  $r_j$  as

$$r_j = \arg \max(\text{Softmax}(Xv_j)) \quad (1)$$

The dot product of the pointer embedding with the object embeddings forms a probability distribution over the objects. The reference object is the object with the highest probability mass.  $S$

## 6 Program Self Training

In this section we describe our program self training algorithm which discovers DSL programs from 3D scene data. It also helps improve next-object distribution locations predicted by our system. An overview of our algorithm is shown in Figure 3.

Our algorithm falls under the PLAD [12] family, a conceptual framework for unsupervised program bootstrapping. These methods iteratively improve a dataset of programs by searching for new and better programs, retraining on them, and then repeating the process. Our algorithm also takes inspiration from Talton et al. [27], a work which uses probabilistic context free grammars (PCFGs) to learn a procedural model from a set of examples. The optimization begins with the “most-specific” grammars and converges to a grammar which is not too specific and not too general.

Our optimization begins with programs that represent a subset of all possible valid placements. For each object in a scene, we use geometric heuristics to apply every possible constraint to the object so that the extracted program will only place the object where it was originally found. Further details are described in the supplemental material. We add additional placement modes to these “most restrictive” programs through a search and filtering process. Similar to Ganeshan et al. [10] we use domain specific operations to edit a subset of programs per iteration of self training. This is both for computational feasibility and distributional stability between iterations.

### 6.1 Candidate Program Generation

For a given partial scene and object to add we search for new programs by sampling the generative model described in Section 5 and then relaxing both the inferred and original program. Program relaxation for our representation entails randomly removing constraints from the program tree to produce a new program. This can help generalize the overly restrictive programs produced by our initial naive approach. For example, an object found in the corner of a room might initially be constrained to rest against both walls. Removing one of these constraints would allow the program to predict more general placements along one of the walls.

We employ simple filtering techniques on candidate programs. Severely unconstrained programs that elect to place the object anywhere in the scene and programs which do not predict any placements are removed from consideration. We also ensure that each candidate program predicts only one valid orientation. Programs which predict placements for multiple orientations are split into subtrees that each predict only one. This constraint reduces the number of repeated subtrees present in the final programs and improves quantitative performance of this algorithm. It is worth noting that this constraint does not restrict the possible orientations the final

program can represent. Our later program combination step produces final programs which predict valid placements for multiple different object orientations.

## 6.2 Scene Classifier

While it might seem unintuitive to use a scene classifier to classify programs, we choose this route because both hard positives and negatives for programs are hard to generate. We train a real fake scene classifier to predict how in or out of distribution a particular object placement is. Positive examples are generated by randomly subsampling scenes. Negative examples come from randomly perturbing the rotation and location of a single object. Even though this process can generate negative examples that should be labeled positive, we find it sufficient for learning a useful decision boundary. We sample a program’s predicted mask multiple times, insert the object in question at those sampled locations, and then compute the scene classifier’s real probability for each insertion. The program’s final score is the average of these probabilities.

We perform experiments on two architectures of scene classifiers. We test a CNN based classifier that uses the input representation of Ritchie et al. [23] with an additional input channel denoting the query object in the scene. We also test a transformer based classifier that uses the object encoder described in section 5 with an additional learned vector embedding added to the query object vector encoding. While our CNN based classifier reports higher precision (accepts less invalid programs), and results in better downstream quantitative metrics we report results using the transformer classifier due to constraints on computation. The transformer model is computationally cheaper than the CNN because it does not require rasterization of the top down view of the scene. More details on the scene classifiers and their evaluation are included in the supplemental material.

## 6.3 Combining Programs

By this stage in the pipeline we have, for a partial scene and object to place, multiple programs whose placement distributions have been scored by a scene classifier. Each program is also guaranteed to predict a placement distribution with only one possible orientation. We combine programs whose score is above a preset threshold to produce a new final program on which our generative model will be retrained.

Recall that our programs takes on the structure of a CSG tree. If two candidate programs predict two different placement modes in a scene, we can combine them into a single program that predicts both modes concurrently. We do this by creating a new tree with an **or** node as its root. The two programs are its children. This process is repeatable for any number of candidate programs as the newly combined program is now considered a single entity which we can combine with another program. For every possible orientation our language can represent we choose the candidate program with the largest mask. These programs are combined together to produce the final program. Figure 6 shows an example of this process.

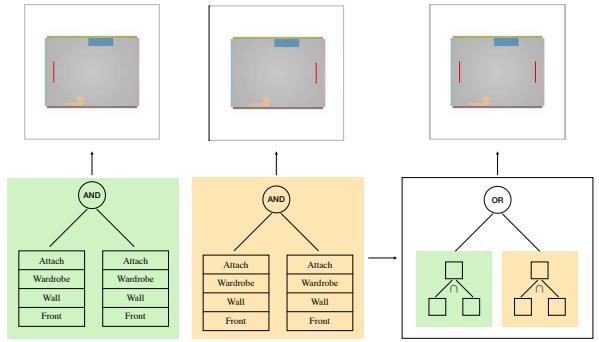


Fig. 6. Given two programs that produce different placement modes, we combine them into a new program with domain specific operations. We create a new tree with an “or” node as its root. The two programs are set as its children.

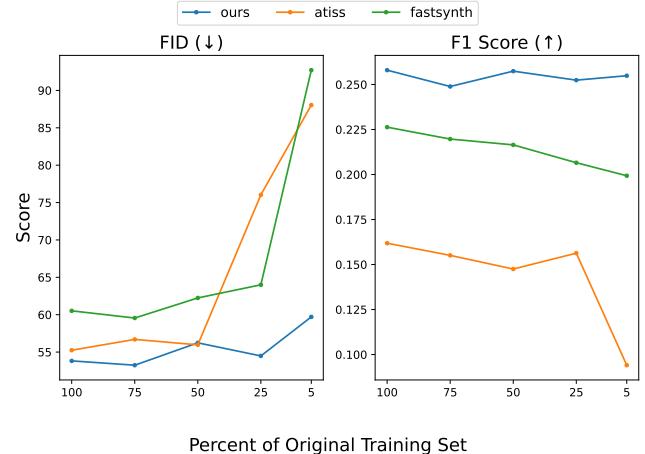


Fig. 7. Our system maintains performance in both scene synthesis and location distribution modeling with as little as 5% of the original training examples present. Other systems degrade in both final scene quality and the consistency of their per object location distribution.

## 7 Evaluation

We demonstrate how our system produces per object location distributions that are both more complete and more accurate than than previous data-driven autoregressive methods. Our system can model these distributions without sacrificing final scene quality. We also show our system’s superior performance in both modeling this location distribution and scene synthesis when data is sparse.

To demonstrate these claims we evaluate our system against FastSynth [23] and ATISS [20] – two data-driven autoregressive scene synthesis systems – on the 3DFRONT [7] dataset. We train individual models for four scene types: bedrooms, libraries, living rooms, and dining rooms. We evaluate two versions of the baseline methods. One version uses the recommended training time and the

Table 1. We report FID, Category KL Divergence, and Scene Classifier Accuracy of generated scenes. Methods denoted with “best F1” use the training epoch with the highest F1 score on the location distribution metrics. Methods denoted “recommended” use the recommended training time. Empty rows mean that the recommended training settings also produce the best F1 scores. Our method generates scenes of comparable quality to previous systems. Although our edge attention and self training algorithm improves our system’s ability to model per object location distributions, they do not significantly hurt or improve its scene synthesis capabilities. ATISS performance degrades when using the epochs with the highest F1 scores because they are often early epochs that have not yet memorized object placements. Fastsynth does not suffer from the same phenomenon.

	All			Bedroom			Dining room			Living room			Library		
	FID ↓	CKL ↓	SCA %	FID ↓	CKL ↓	SCA %	FID ↓	CKL ↓	SCA %	FID ↓	CKL ↓	SCA %	FID ↓	CKL ↓	SCA %
ATISS (recommended)	55.245	0.0265	<b>76.62</b>	<b>20.148</b>	<b>0.0067</b>	<b>60.89</b>	96.063	0.0279	83.65	68.269	0.0356	80.60	36.499	0.0359	<b>81.34</b>
ATISS (best F1)	67.534	0.0214	82.87	20.748	0.0068	65.44	103.062	0.0216	89.95	75.369	0.0408	86.89	104.178	0.0252	92.71
Fastsynth (recommended)	60.517	0.20434	89.80	29.617	0.220	86.38	100.976	0.1923	90.798	68.845	0.1870	91.865	42.631	0.2180	90.163
Fastsynth (best F1)	61.02	0.2043	90.671	-	-	-	102.311	0.199	93.00	-	-	-	43.308	0.2112	91.44
Ours (no edge attention)	<b>52.847</b>	<b>0.0223</b>	78.23	22.712	0.0109	73.01	<b>90.375</b>	0.0325	81.30	<b>62.733</b>	<b>0.0231</b>	77.16	<b>35.567</b>	0.0227	81.46
Ours (no self training)	54.456	0.02482	77.69	24.170	0.0217	71.32	91.579	<b>0.0251</b>	<b>80.38</b>	64.623	0.0321	77.67	37.453	<b>0.0204</b>	81.38
Ours	53.823	0.0243	78.31	22.670	0.01133	72.52	90.665	0.0251	83.36	64.040	0.0229	<b>75.76</b>	37.916	0.0378	81.59

other uses the training epoch which maximizes performance on our proposed location distribution metric.

We justify our self training algorithm by evaluating our system without it. We also designed a variant of our method’s generative model which uses an augmented version of the regular self attention mechanism. This was motivated by empirical evidence that the model could not correctly attend to spatial relationships of objects in a scene. See the supplemental material for more details.

### 7.1 Location Distribution

We evaluate a scene synthesis system’s ability to model per object location distributions by measuring the Precision, Recall, and F1 Score between a location mask of possible centroid locations extracted from the method in question and a ground truth mask that comes from human annotators. We gave human annotators a partial scene and object and asked them to mark all the possible places that object can go. More details on this user study are shown in the supplemental material.

We generated 100 partial scenes and objects for each scene type. These partial scenes and objects follow the placement order used during scene generation. Our system outputs 4 masks that each represent a different orientation of the object, so before comparison it is collapsed into a single mask. For methods that output continuous values such as ATISS and FastSynth we generate masks by binarizing the output values with the threshold value that maximizes the F1 score on the eval set. Masks are dilated before comparison to eliminate sensitivity to small discrepancies.

We show performance of our system over the course of self training as well as performance of the baseline methods in figure 9. Our method adds placements modes increasing recall over the course of self training without sacrificing precision. Baseline methods give the highest F1 scores with low thresholds and early training epochs resulting in fuzzy masks that can cover many modes, but are imprecise. Recall drops when using the recommended settings because longer training time results in overfitting to particular object placements.

Qualitative examples of masks extracted from our methods are shown in Figure 8.

### 7.2 Scene Synthesis

We evaluate our system’s ability to perform scene generation from a given floor plan. Our programs do not automatically determine the category and size of the next object to place. We use the category prediction module from Ritchie et al. [23] to predict the object category of the next object to place and randomly sample category dimensions from the dataset of objects.

In accordance with previous work we report the FID scores of top down orthographic renderings of final scenes, the object category KL divergence between a set of generated scenes and an eval set from 3DFRONT, alongside the real fake scene classifier accuracy on the two scenes. Object models closest to the bounding box dimensions of each object are chosen from 3D-FRONT [7]. Images for FID come from orthographic renderings of generated scenes where each object model receives per-class coloring.

We finetune a pre-trained AlexNet [14] to classify these orthographic renderings and report its classification accuracy on a held out test set. Closer to 50% is better since the classifier struggles to differentiate between real generated scenes. Our quantitative results are shown in Table 1 and qualitative shown in Figure 10.

Our method generates scenes of comparable quality to previous systems. Although our edge attention and self training algorithm improves our system’s ability to model per object location distributions, they do not significantly hurt or improve its scene synthesis capabilities. ATISS performance degrades when using the epochs with the highest F1 scores because they are often early epochs that have not yet memorized object placements. Fastsynth does not suffer from the same phenomenon.

### 7.3 Data sparsity

One advantage our system has is that its performance does not degrade to the same degree when there are less data samples to train on. Previous systems require many example placements to learn complete distributions. They also rely on overfitting to object arrangements to produce high quality final scenes. Our system’s program bootstrapping algorithm can generalize sparse samples to more general placement rules. These rules allow our system to

maintain consistent performance even when training data is sparse. Even if our generative model overfits to particular programs, our program executor guarantees continuity of the location distribution and validity of placement.

We show in Figure 7 the effect that less training examples have on the predicted location distribution and show how our method does not show the same degree of degradation as other methods. We also report the FID score to measure final scene quality with respect to number of samples trained on.

## 8 Conclusion

In this work, we study how autoregressive indoor scene synthesis systems overfit to object placements seen during training, producing incomplete location distributions. We introduce a new approach to indoor scene synthesis which uses programs as an intermediate representation to address this issue. Our evaluation demonstrates how final scene quality and completeness of the intermediate location distributions used to produce those scenes are often independent. Current scene quality metrics do not accurately reflect this loss in placement diversity. We hope this work brings greater attention to this phenomenon, as well as neurosymbolic methods as a whole.

## 9 Ethics Statement

Scenes used to train and evaluate our method come from a primarily western design canon and represent only a subsection of the indoor spaces people inhabit. We also design our language with the assumptions that the inhabitants are able bodied. Indoor scenes generated by our method are thus potentially biased against underrepresented groups.

## Acknowledgments

Thank you Kenny Jones for the early conversations on program induction. Thank you Sheridan Feucht for the discussions and support. Thank you James Tompkin for the helpful feedback in early drafts. Thank you Luca Fonstad and Cal Nightingale for their help on the project. Thank you to everyone who volunteered their time to help with annotation. Funding was provided by NSF award #1941808.

## References

- [1] Planner 5D. 2011. *Planner 5D: House Design Software / Home Design in 3D*. <https://planner5d.com>
- [2] Rio Aguinaga-Kang, Maxim Gumin, Do Heon Han, Stewart Morris, Seung Jean Yoo, Aditya Ganeshan, R. Kenny Jones, QiuHong Anna Wei, Kailiang Fu, and Daniel Ritchie. 2024. Open-Universe Indoor Scene Generation using LLM Program Synthesis and Uncurated Object Databases. *arXiv preprint arXiv:2403.09675* (2024).
- [3] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. 2017. Mode Regularized Generative Adversarial Networks. *ICLR* (2017). <https://arxiv.org/pdf/1612.02136.pdf>
- [4] Kevin Ellis, Maxwell Nye, Yewen Pu, Felix Sosa, Joshua B. Tenenbaum, and Armando Solar-Lezama. 2019. *Write, execute, assess: program synthesis with a REPL*. Curran Associates Inc., Red Hook, NY, USA.
- [5] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2021. DreamCoder: bootstrapping inductive program synthesis with wake-sleep library learning (*PLDI 2021*). Association for Computing Machinery, New York, NY, USA, 835–850. doi:10.1145/3453483.3454080
- [6] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. 2015. Activity-centric scene synthesis for functional 3D scene modeling. *ACM Transactions on Graphics (TOG)* 34 (2015), 1 – 13. [https://graphics.stanford.edu/~niessner/papers/2015/9synth/fisher2015activity\\_orig.pdf](https://graphics.stanford.edu/~niessner/papers/2015/9synth/fisher2015activity_orig.pdf)
- [7] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10933–10942.
- [8] Qiang Fu, Xiaowu Adaptive synthesis of indoor scenes via activity-associated object relation graphs Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. 2017. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics (TOG)* 36 (2017), 1 – 13. <https://dl.acm.org/doi/10.1145/3130800.3130805>
- [9] Samir Gadre, Kiana Ehsani, Shuran Song, and Roozbeh Mottaghi. 2022. Continuous Scene Representations for Embodied AI. *CVPR* (2022).
- [10] Aditya Ganeshan, R. Kenny Jones, and Daniel Ritchie. 2023. Improving Unsupervised Visual Program Inference with Code Rewriting Families. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- [11] Lin Gao, Jia-Mu Sun, Kaichun Mo, Yu-Kun Lai, Leonidas J Guibas, and Jie Yang. 2023. SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation with Fine-Grained Geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [12] R. Kenny Jones, Homer Walke, and Daniel Ritchie. 2022. PLAD: Learning to Infer Shape Programs with Pseudo-Labels and Approximate Distributions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
- [13] R. Kenny Jones, Renhao Zhang, Aditya Ganeshan, and Daniel Ritchie. 2024. Learning to Edit Visual Programs with Self-Supervision. In *Advances in Neural Information Processing Systems*.
- [14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) (NIPS’12). Curran Associates Inc., Red Hook, NY, USA, 1097–1105.
- [15] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baquan Chen, Daniel Cohen-Or, and Hao Zhang. 2018. GRAINS. *ACM Transactions on Graphics (TOG)* 38 (2018), 1 – 16. <https://arxiv.org/pdf/1807.09193.pdf>
- [16] Chen Liang, Jonathan Berant, Quoc V. Le, Kenneth D. Forbus, and N. Lao. 2016. Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision. *ArXiv abs/1612.01197* (2016). <https://aclanthology.org/P17-1003.pdf>
- [17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2999–3007. [https://openaccess.thecvf.com/content\\_iccv\\_2017/papers/Lin\\_Focal\\_Loss\\_for\\_ICCV\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_iccv_2017/papers/Lin_Focal_Loss_for_ICCV_2017_paper.pdf)
- [18] Paul C. Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. 2011. Interactive furniture layout using interior design guidelines. *ACM SIGGRAPH 2011 papers* (2011). [https://cs.stanford.edu/people/eschkuft/docs/siggraph\\_11.pdf](https://cs.stanford.edu/people/eschkuft/docs/siggraph_11.pdf)
- [19] Wamiq Reyaz Para, Paul Guerrero, Tom Kelly, Leonidas J. Guibas, and Peter Wonka. 2020. Generative Layout Modeling using Constraint Graphs. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2020), 6670–6680. <https://arxiv.org/abs/2011.13417>
- [20] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. 2021. ATISS: Autoregressive Transformers for Indoor Scene Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [21] Xavi Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Ruslan Partsey, Jimmy Yang, Ruta Desai, Alexander William Clegg, Michal Hlavac, Tiffany Min, Theo Gervet, Vladimir Vondrus, Vincent-Pierre Berges, John Turner, Oleksandr MakSYMets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. 2023. Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots.
- [22] Daniel Ritchie, Paul Guerrero, R. Kenny Jones, Niloy Jyoti Mitra, Adriana Schulz, Karl D. Willlis, and Jiajun Wu. 2023. Neurosymbolic Models for Computer Graphics. *Computer Graphics Forum* 42 (2023). <https://api.semanticscholar.org/CorpusID:258236273>
- [23] Daniel Ritchie, Kai Wang, and Yu-An Lin. 2018. Fast and Flexible Indoor Scene Synthesis via Deep Convolutional Generative Models. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 6175–6183. <https://arxiv.org/pdf/1811.12463.pdf>
- [24] RoomSketcher. 2024. *RoomSketcher*. <https://www.roomskeetcher.com>
- [25] Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, and Subhransu Maji. 2018. CSGNet: Neural Shape Parser for Constructive Solid Geometry. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [26] Armando Solar-Lezama. 2008. *Program synthesis by sketching*. Ph.D. Dissertation. USA. Advisor(s) Bodil Rastislav. AAI3353225.
- [27] Jerry O. Talton, Lingfeng Yang, Ranjitha Kumar, Maxine Lim, Noah D. Goodman, and Radomir Méch. 2012. Learning design patterns with bayesian grammar induction. *Proceedings of the 25th annual ACM symposium on User interface software and technology* (2012). <https://api.semanticscholar.org/CorpusID:17007327>
- [28] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. 2024. Diffuscene: Denoising diffusion models for generative indoor

- scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [29] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:13756489>
  - [30] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) (*NIPS’15*). MIT Press, Cambridge, MA, USA, 2692–2700.
  - [31] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X. Chang, and Daniel Ritchie. 2019. PlanIT. *ACM Transactions on Graphics (TOG)* 38 (2019), 1 – 15. <https://kwang-ether.github.io/pdf/planit.pdf>
  - [32] Kai Wang, Manolis Savva, Angel X. Chang, and Daniel Ritchie. 2018. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)* 37 (2018), 1 – 14. <https://dritchie.github.io/pdf/deepsynth.pdf>
  - [33] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. 2020. SceneFormer: Indoor Scene Generation with Transformers. *arXiv preprint arXiv:2012.09793* (2020).
  - [34] Karl D. D. Willis, Yewen Pu, Jieliang Luo, Hang Chu, Tao Du, Joseph G. Lambourne, Armando Solar-Lezama, and Wojciech Matusik. 2021. Fusion 360 gallery: a dataset and environment for programmatic CAD construction from human design sequences. 40, 4, Article 54 (July 2021), 24 pages. doi:10.1145/3450626.3459818
  - [35] Rundi Wu, Chang Xiao, and Changxi Zheng. 2021. DeepCAD: A Deep Generative Network for Computer-Aided Design Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6772–6782.
  - [36] Xiang Xu, Karl DD Willis, Joseph G Lambourne, Chin-Yi Cheng, Pradeep Kumar Jayaraman, and Yasutaka Furukawa. 2022. SkexGen: Autoregressive Generation of CAD Construction Sequences with Disentangled Codebooks. In *International Conference on Machine Learning*. PMLR, 24698–24724.
  - [37] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. 2023. Holodeck: Language Guided Generation of 3D Embodied AI Environments. *arXiv preprint arXiv:2312.09067* (2023).
  - [38] Yi-Ting Yeh, Lingfeng Yang, Matthew Watson, Noah D. Goodman, and Pat Hanrahan. 2012. Synthesizing open worlds with constraints using locally annealed reversible jump MCMC. *ACM Transactions on Graphics (TOG)* 31 (2012), 1 – 11. <https://api.semanticscholar.org/CorpusID:2270108>
  - [39] Lap-Fai Craig Yu, Sai-Kit Yeung, Chi-Keung Tang, Demetri Terzopoulos, Tony F. Chan, and S. Osher. 2011. Make it home: automatic optimization of furniture arrangement. *ACM SIGGRAPH 2011 papers* (2011). <https://web.cs.ucla.edu/~dt/papers/siggraph11/siggraph11.pdf>

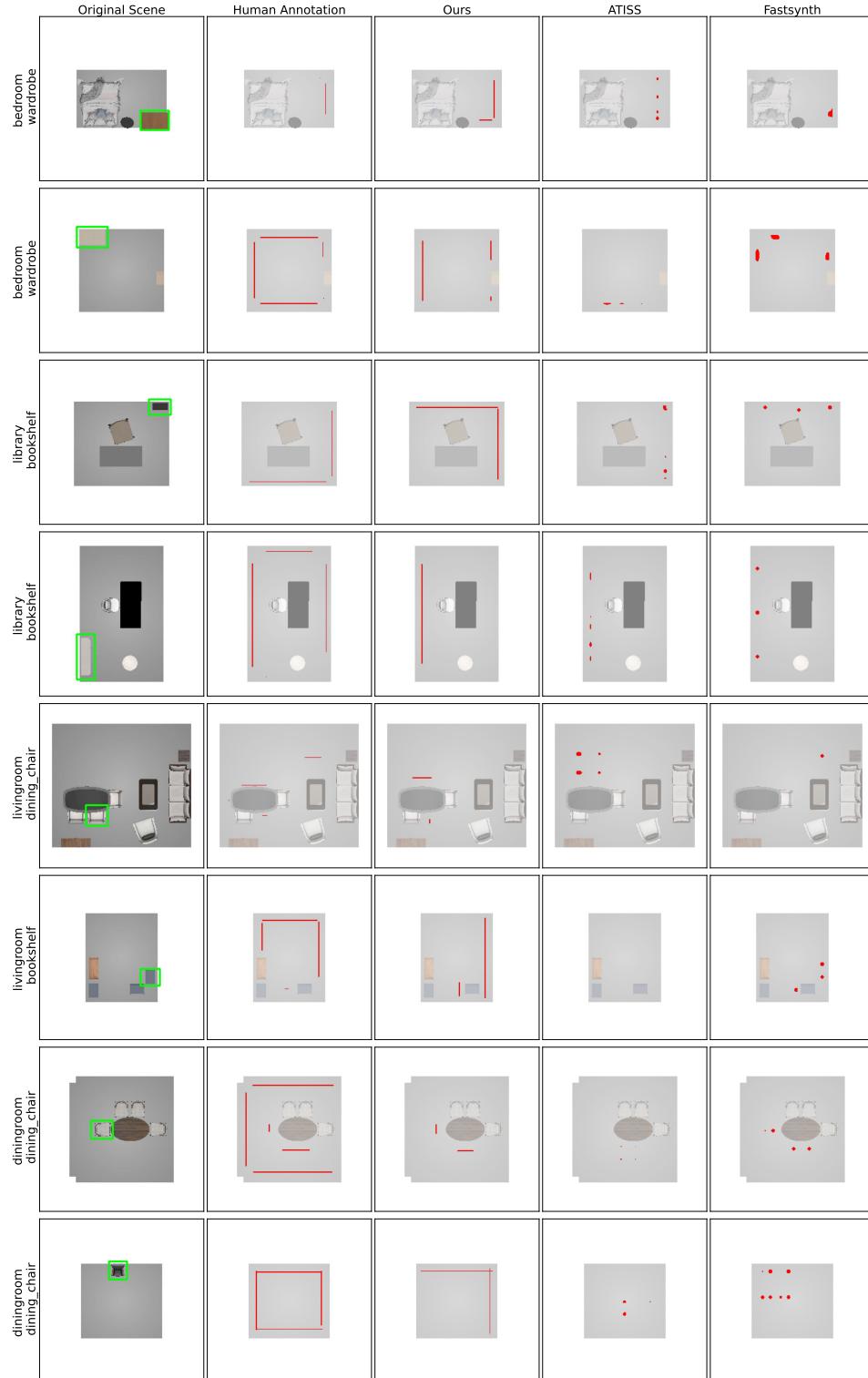


Fig. 8. Visualization of annotated masks alongside the location distributions predicted by our method, ATISS, and Fastsynth. Masks from our method come from the last iteration of self training. Masks for baseline methods come from the training epochs with the recommended training time. Our method predicts location distributions that are more consistent with human annotators. Previous methods overfit to particular placements seen during training.

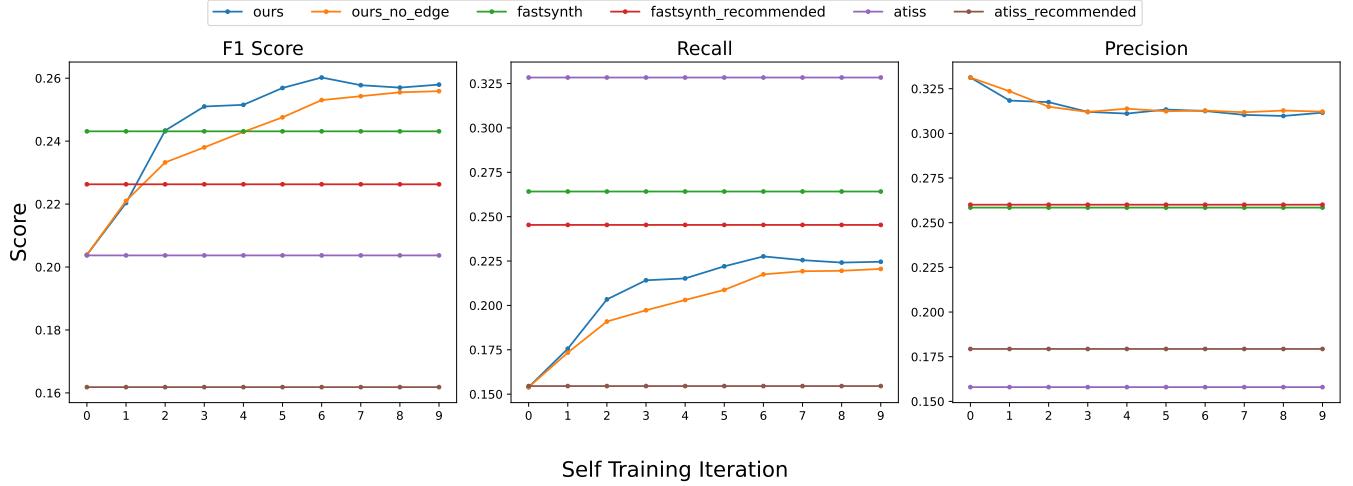


Fig. 9. We show the F1 score, precision, and recall of per object location distributions compared against human annotated masks. The x axis is each iteration of our program bootstrapping algorithm. Our method adds placements modes without sacrificing precision. Baseline methods that produce continuous values are binarized with the threshold value that maximizes F1 score. We choose the best performing training epoch and report it. Methods denoted with "recommended" show performance using the originally proposed training settings. Recall drops when using the recommended settings because the recommended training time results in overfitting to particular object placements.

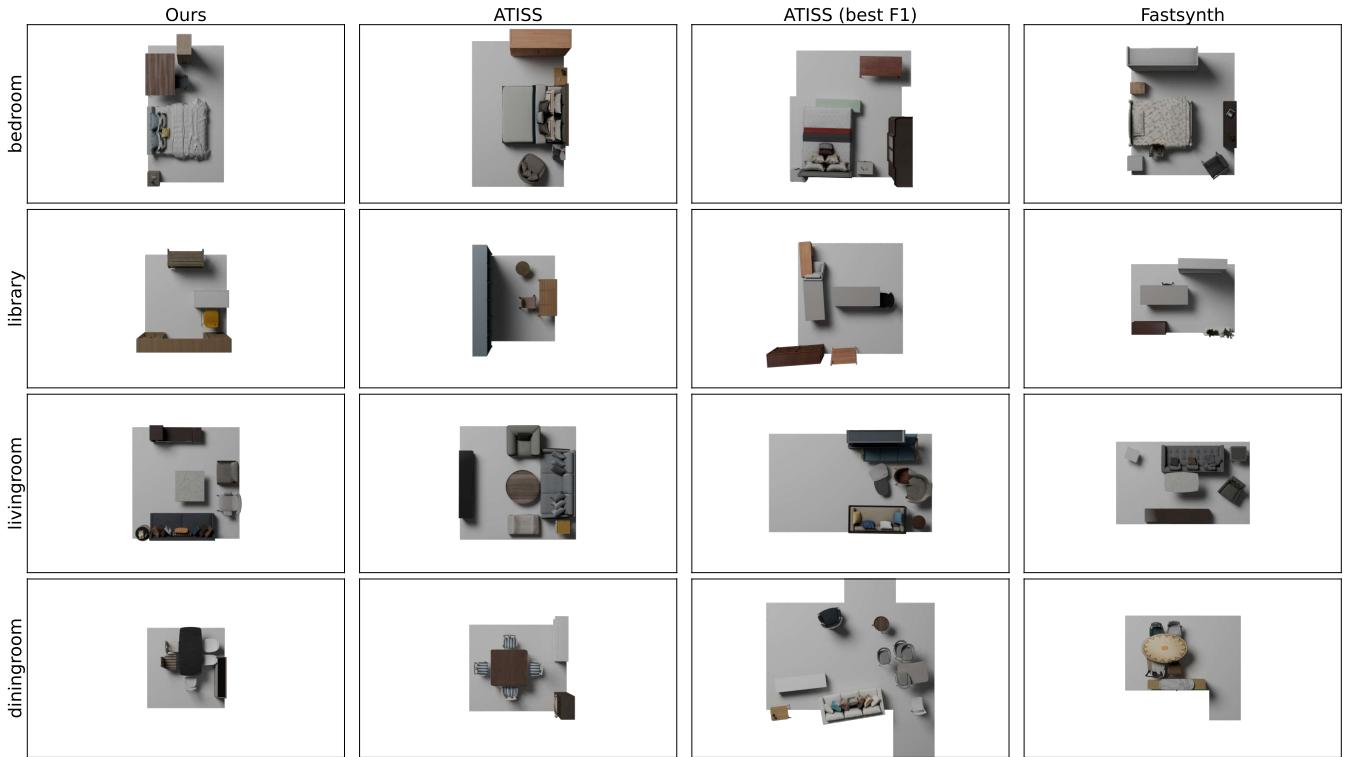


Fig. 10. Examples of scenes generated by ATISS, FastSynth, and our method. Our method is capable of generating scenes of comparable quality to previous methods. Note the degradation in final scene quality when ATISS uses the training settings which produce the highest F1 score.

## A Further Details on Constraints

We assume the input objects to each constraints have the following assumed properties. Objects in each category are aligned in a canonical coordinate frame. They must also come labeled with whether they are meant for holding humans (i.e. beds and chairs). Objects and their original scenes must also come with semantically meaningful sizes, scales, and distances since geometric heuristics described for each constraint are based on physically meaningful quantities such as the average reaching distance. Objects and scenes which satisfy this criteria come from preprocessing the 3DFRONT [7] dataset.

Shown in Figure 11 are examples of executed constraints.

## B Edge Attention mechanism

We describe the edge attention mechanism that augments our base transformer model.

Consider the setting in Figure 12. The partial scene contains a single bed and nightstand to the right of it. The query object is also a nightstand. Given this setting our generative model should predict a program that places the nightstand on the left of the bed 100 % of the time. Instead, the logits corresponding to which side of the bed the program will place the nightstand is split 50-50 left and right. This is likely due to nightstands appearing on either side of the bed with equal frequency. We find this is empirically true on both a model trained on real scene data and a toy setting containing just beds and nightstands. More details of this toy setting are shown in the appendix.

This experiment demonstrates that ordinary attention does not correctly account for spatial relationships between objects in our model. One interpretation of a transformer is that it is an edgeless graph neural network. Edge values denoting which side an object is on in relation to another object should provide the necessary information the network needs to correctly reason over the spatial relationships of objects in the room. As such, we augment the attention mechanism in our transformer model to introduce inter object relationships to the input signal.

In ordinary attention key, query, and value vectors are computed from linear projections of the input. The key and query vectors compute the attention weights used for a final weighted sum of the value vectors. The output of regular self-attention  $Z$  is defined as

$$Z = \text{Softmax}\left(\frac{W_q X (W_k X)^T}{\sqrt{d_k}}\right) W_v X \quad (2)$$

Our edge attention mechanism adds inter object information to this computation. We extract directional relationships between objects and encode them into a matrix of edge values  $E \in \mathbb{R}^{txd}$ . Encoding directional relationships means that each object must receive its own matrix of edge values. We allow the original embedding vector inform which edges should receive more weight. The category, size, and location information encoded in the original embedding vector of an object should inform which other objects it pays attention to. For example, a bed should pay more attention to its spatial relationship with a nightstand than with a chair.

For each object, denoted by its index  $i$ , we compute key and value vectors  $K'$  and  $V'$  with respect to it. The attention weights are

computed as  $QK^T + QK'^T$ , where the  $QK'^T$  term acts as a correction weight to the original  $QK^T$  attention weights. The weighted sum of  $V'$  using these attention weights is added to the normal output of attention for the object. Given edge values  $E_i \in \mathbb{R}^{txd}$ , our edge attention mechanism adds inter object information to each output of the original attention mechanism  $Z_i$ .

$$Z_i = Z_i + \text{Softmax}\left(\frac{W_q X (W_k X)^T + W_q X (W_{ek} E_i)^T}{\sqrt{d_k}}\right) W_{ev} E_i \quad (3)$$

## C Initial Program Extraction Implementation

Here, we describe in greater detail how programs are initially extracted from scene data. As a pre-processing step, scenes with major inter-object bounding box collisions are removed from the dataset. These scenes can produce errors in the scene extraction process.

For every query object we first consider all the objects within attachment distance (15 cm). If a reference object is within attachment distance and also faces the same direction as the query object, an alignment constraint is also applied. Otherwise, if the two objects face each other, a face constraint is applied. If the query object is meant to hold humans such as a bed or a chair, the same process is applied for all objects within reaching distance (15 - 60 cm).

This process is very sensitive to hyper parameters and can often fail to produce valid programs. In the case where a null program is extracted (a program that produces no object placements), we search through its children and if a subtree produces a program which contains the original placement, it is accepted. It is also often the case that too many constraints are applied and there are "extraneous" constraints, or constraints which when applied do not change the final output. These constraints are also removed.

## D Scene Classifier Evaluation

We wrote a simple procedural grammar for bedrooms that can generate a dataset of scenes that comes with the ground truth location distributions used to place objects. We evaluate our scene classifier in this toy setting. Every object in a scene comes with a mask representing its "ground truth" location distribution. We use the initial program extraction to generate a program which places this object. We add this "ground truth" mask as a positive example to the evaluation set and negative examples are generated by randomly removing constraints from the extracted program.

We found that the scene classifier can overfit to particular locations seen during training, so we employ various techniques to reduce overfitting. We employ focal loss [17] with different  $\lambda$  parameters, weight examples based on the level of perturbation of the object, and employ random translation during training.

Table 2 shows the performance of different experimental choices of the scene classifier with a acceptance threshold of 0.6. The metrics p-cons and n-cons measures for a positive and negative example, how consistently the classifier correctly labels a mask. We seek to minimize the false positive rate which is in this settings corresponds to incorrectly predicting an object as in distribution.

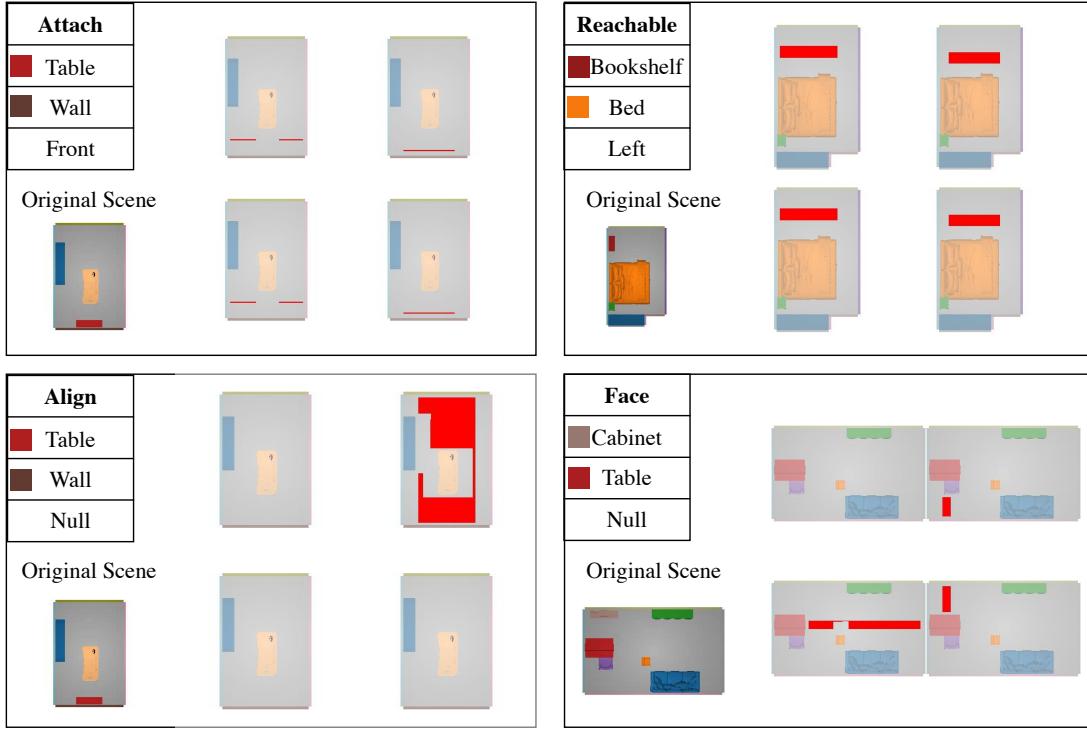


Fig. 11. Constraint Examples: Shown are examples of constraints, their input scene and object, and their executed masks. Objects are colored by how they appear in the scene visualization. The original scene shows where the query object was originally placed. There are four masks for each constraint. Each mask represents a possible orientation of the query object. For example the align constraint contains placement options for only one orientation (the orientation of the reference wall object). The other constraints contain placement options for all possible orientations, but those locations are constrained based on the input arguments.

Table 2. We measure the quantitative performance of different design choices for the scene classifier. We wrote a procedural grammar that can generate a dataset of scenes that comes with "ground truth" location distributions, and use those masks to create an evaluation set. We seek to minimize the false positive rate which corresponds to incorrectly predicting an object as in distribution.

method	accuracy	tn	fp	fn	tp	p-cons	n-cons
resnet34	0.9386	0.5400	0.0131	0.0483	0.3986	0.9440	0.9160
resnet34-focal-loss-2	0.9724	0.5359	0.0172	0.0103	0.4366	0.9920	0.8960
resnet34-focal-loss-5	0.9669	0.5303	0.0228	0.0103	0.4366	0.9960	0.8760
resnet34-weighted	0.9538	0.5124	0.0407	0.0055	0.4414	0.9960	0.7720
resnet34-weighted-focal-loss-2	0.8890	0.5372	0.0159	0.0952	0.3517	0.8720	0.9080
transformer	0.9324	0.5097	0.0434	0.0241	0.4228	0.9480	0.7520
transformer-focal-loss-2	0.9538	0.5131	0.0400	0.0062	0.4407	0.9840	0.7760
transformer-focal-loss-5	0.9441	0.5221	0.0310	0.0248	0.4221	0.9680	0.8240
transformer-no-pe	0.9393	0.5290	0.0241	0.0366	0.4103	0.9760	0.8760
transformer-shift-scene-around	0.9456	0.5120	0.0435	0.0109	0.4336	0.9840	0.7560
transformer-weighted	0.9469	0.5117	0.0414	0.0117	0.4352	0.9920	0.7640
transformer-weighted-focal-loss-2	0.9634	0.5186	0.0345	0.0021	0.4448	1.0000	0.8120

## E Scene annotation software

We built a browser based scene annotation software to facilitate the annotation of partial scenes, objects, and where they could go. This

was built as a web app in react. The software allows users to draw rectangles denoting possible centroid locations of the object for a given orientation. Users can visualize what these proposed placements look like in the scene before confirming these placements. We

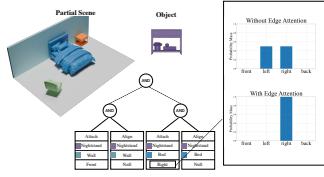


Fig. 12. In this example, we want to place a nightstand in a room that already contains a bed and a nightstand on the left side of it. The program which specifies this placement should always predict placement on the right side of the bed. Without edge attention the program will incorrectly place the nightstand on the left side 50% of the time. With edge attention our model correctly attends to the spatial relationships between objects in the scene and predicts placement on the right side every time.

same data splits of 3D-FRONT as our method, and also without object ordering for a fair comparison with ATISS.

## G More mask examples

We show more examples of masks produced by our and baseline methods. These are shown in Figures 14, 15, 16, 17, and 18.

## H More scene examples

We show more examples of scenes generated by our method and baseline methods. These are shown in Figures 19 and 20.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009



Fig. 13. A screen shot of our annotation software. On the left shows a partial scene and the object hovers on the mouse pointer. Users can draw rectangles to visualize possible placements, and then confirm them once confident in their drawing.

recruited 15 university students and young working professionals to participate in the annotation. No time limit was enforced, but users on average spent on 18 seconds on each partial scene and object, or 30 minutes in total for 100 scenes. A screenshot of this software are shown in Figure 13

## F Baselines Details

The retraining of both baseline methods [20, 23] on 3D-FRONT [7] differ slightly from their original training setting. Both baselines are trained on the same scenes our method trains on and both without object ordering.

ATISS's [20] data preprocessing parses living rooms and dining rooms with a maximum side length of 13.2 meters, and bedrooms and libraries with a maximum side length of 6.2. Our data processing only parses scenes with maximum side of 6.2 for all room types. During our data preprocessing, scenes with significant inter-object penetration are removed from the dataset as they introduce errors to the initial program extraction process. This also reduces the total number of scenes in our data split.

Fastsynth [23] was not originally trained on 3D-FRONT. It was also trained with object ordering. We retrained Fastsynth on the

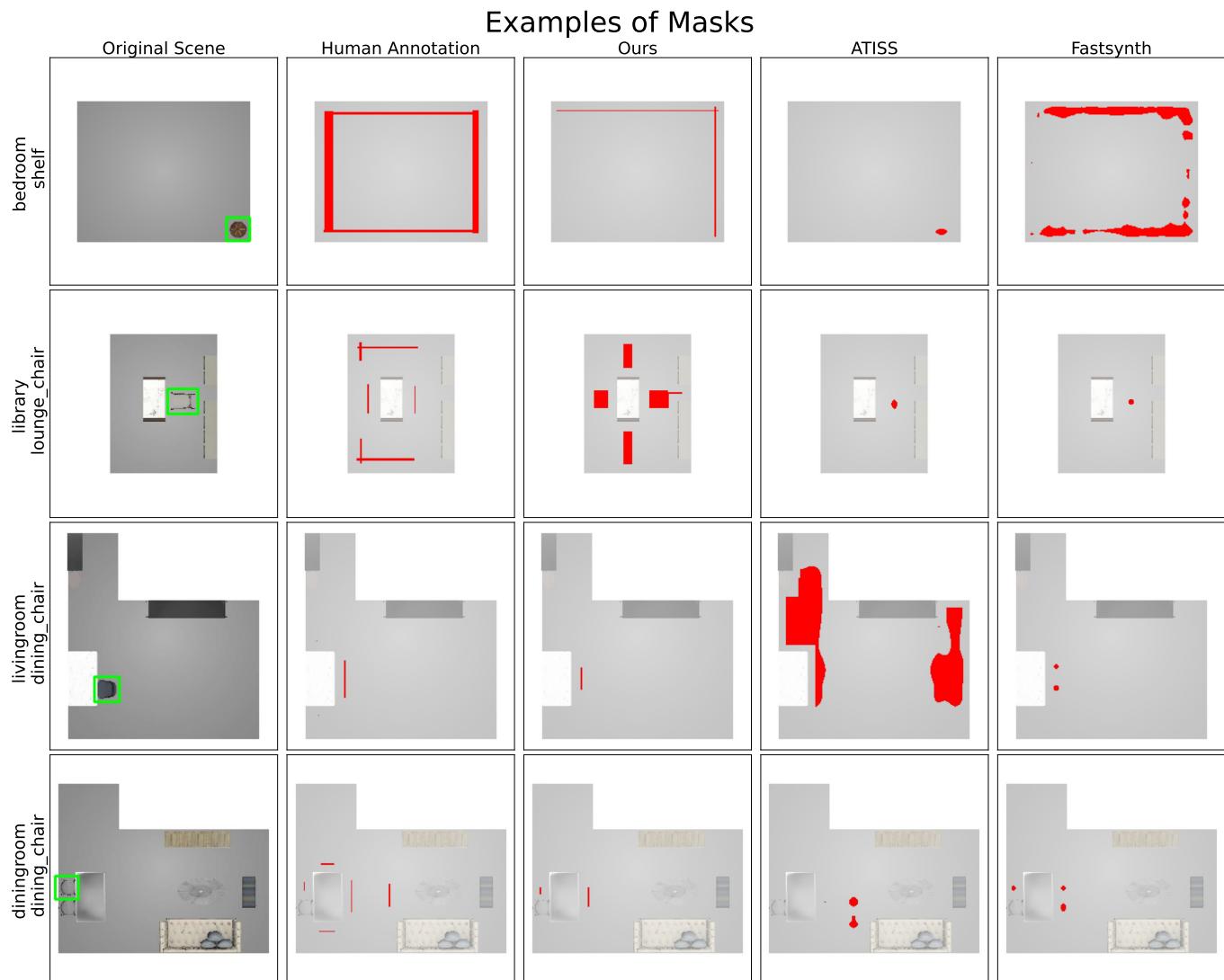


Fig. 14. Additional Mask Examples

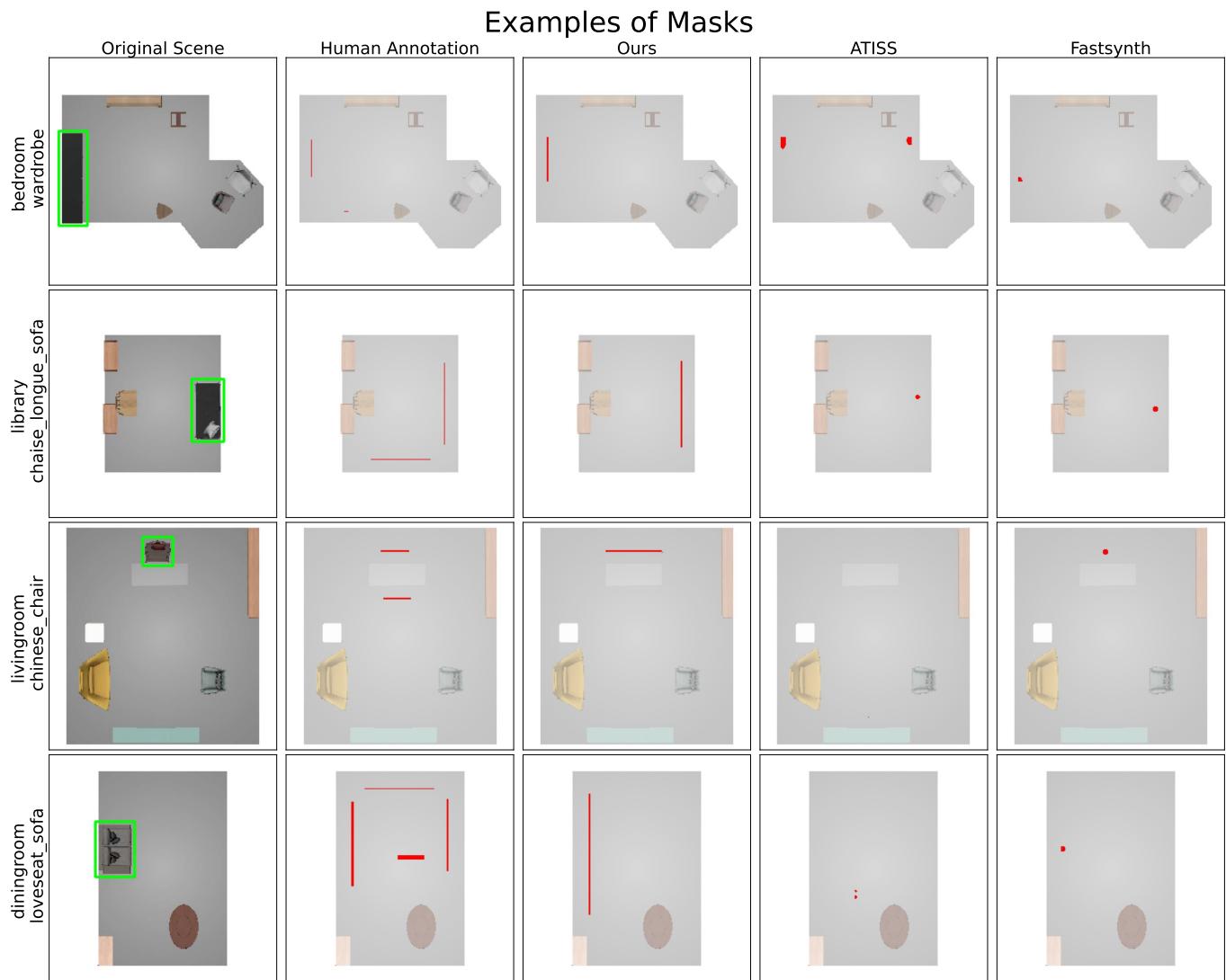


Fig. 15. Additional Mask Examples



Fig. 16. Additional Mask Examples

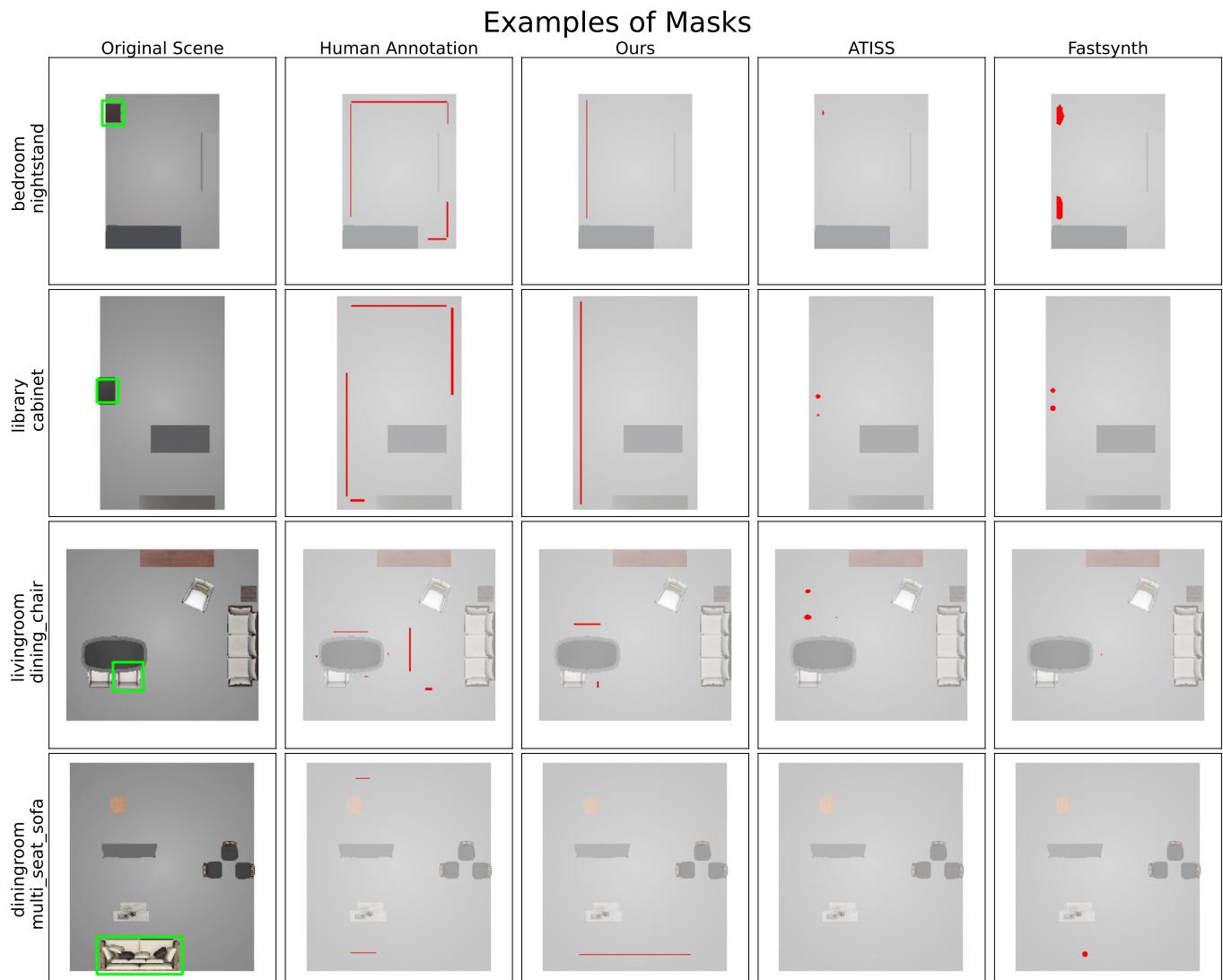


Fig. 17. Additional Mask Examples

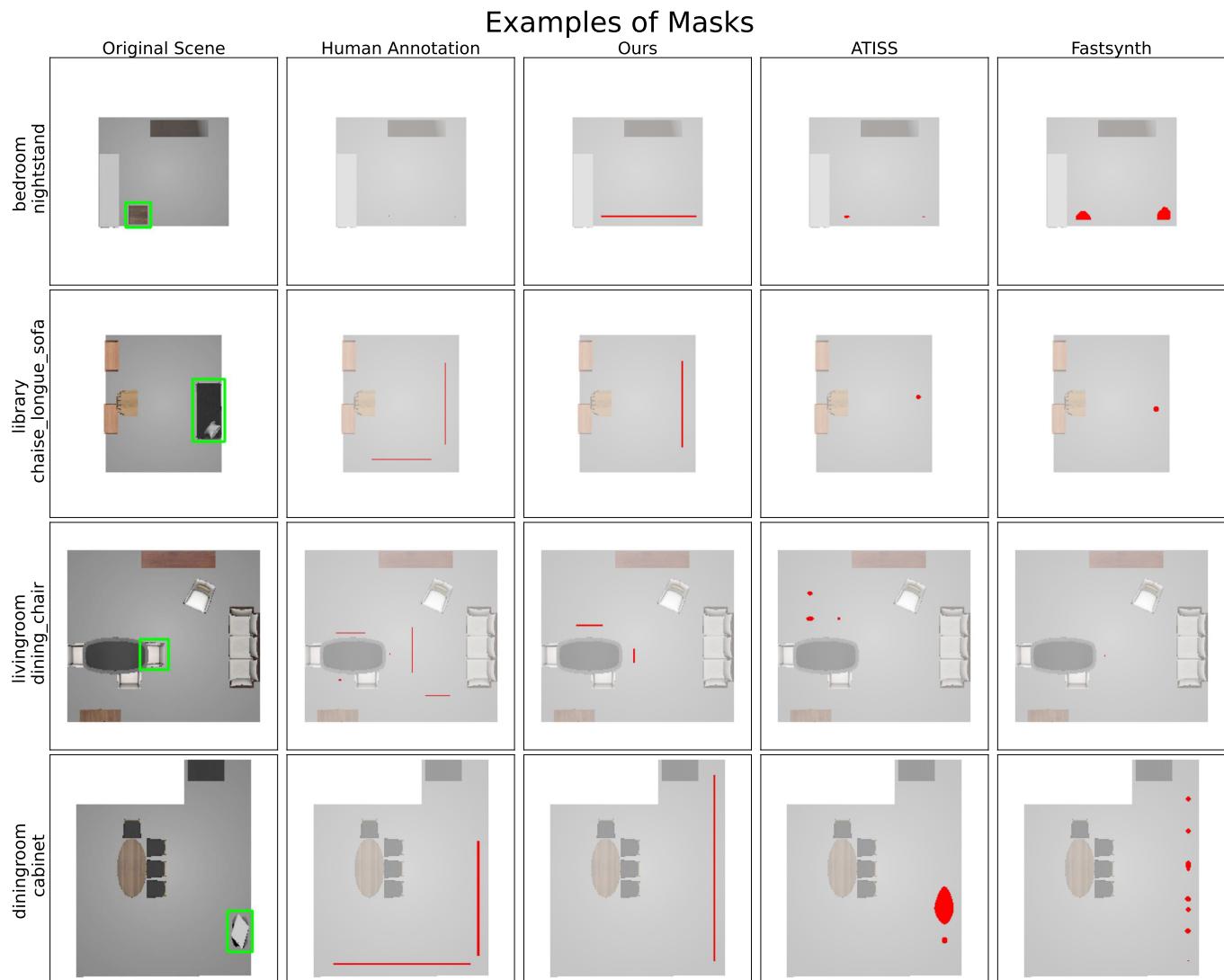


Fig. 18. Additional Mask Examples

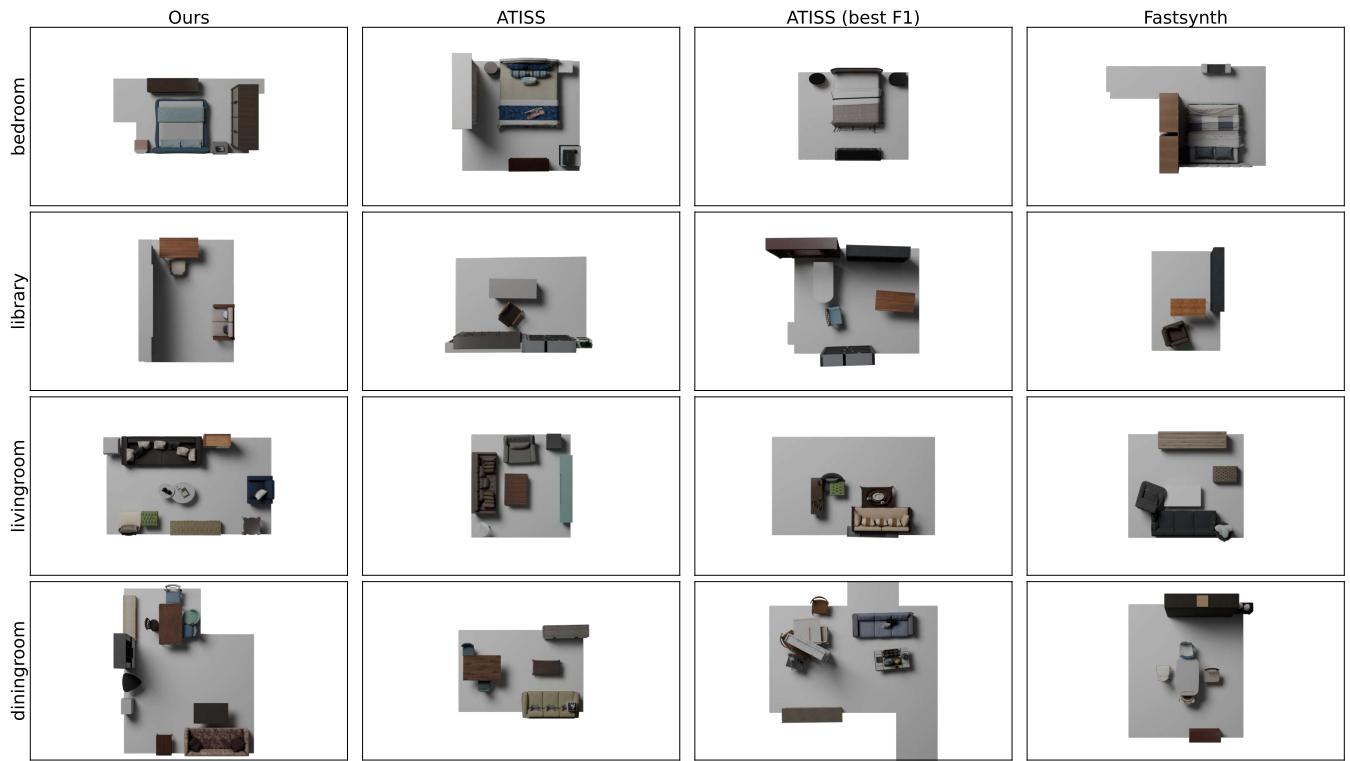


Fig. 19. More examples of scenes generated by our and baseline methods



Fig. 20. More examples of scenes generated by our and baseline methods