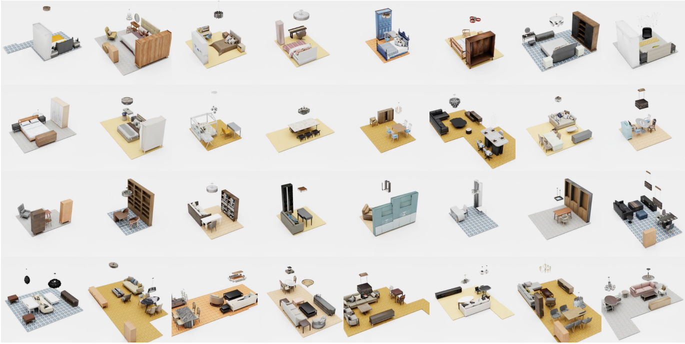
ATISS: Autoregressive Transformers for Indoor Scene Synthesis

Despoina Paschalidou[1,3,4] Amlan Kar[4,5,6] Maria Shugrina[4] Karsten Kreis[4]
Andreas Geiger[1,2,3] Sanja Fidler[4,5,6]

[1]Max Planck Institute for Intelligent Systems Tübingen [2]University of Tübingen
[3]Max Planck ETH Center for Learning Systems
[4]NVIDIA [5]University of Toronto [6]Vector Institute

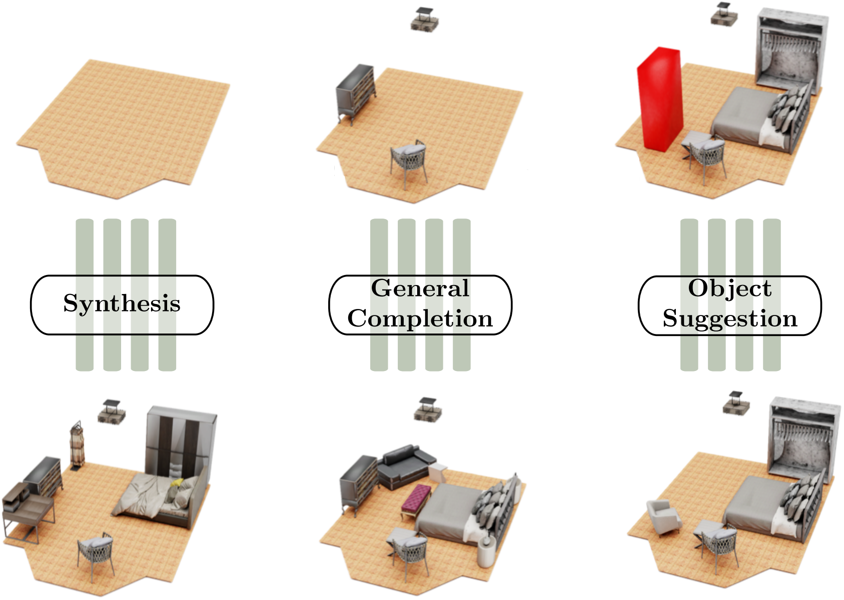
https://nv-tlabs.github.io/ATISS



Motivation

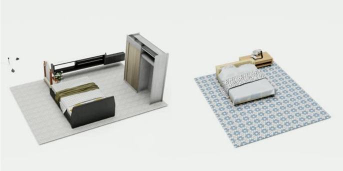
Can we learn a **generative model of object arrangements** trained for **scene synthesis** that can also perform a number of **interactive scenarios** with versatile user input?

Motivation



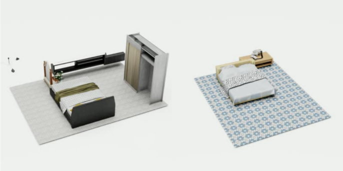
Existing scene synthesis methods
impose unnatural constraints on the scene generation process
because they represent **scenes as ordered sequences of objects**.

FastSynth, Ritchie et al. CVPR 2019



SceneFormer, Wang et al. ARXIV 2020

Existing scene synthesis methods
impose unnatural constraints on the scene generation process
because they represent **scenes as ordered sequences of objects**.



FastSynth, Ritchie et al. CVPR 2019

SceneFormer, Wang et al. ARXIV 2020

We pose scene synthesis as an **unordered set generation problem**.

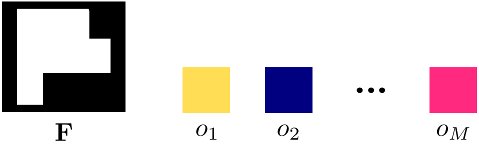
Scene Parametrization

A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.

$\mathbf{F}$ o_1 o_2 o_M

Scene Parametrization

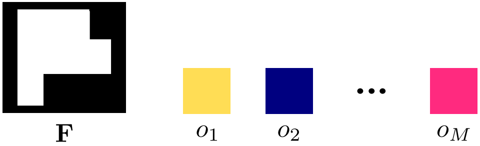
A scene comprises an **unordered set of M objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape $\mathbf{F}$**.



Each object $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{r}_j, \mathbf{t}_j\}$ is modelled with four random variables that describe their **category, size, orientation and location**.

Scene Parametrization

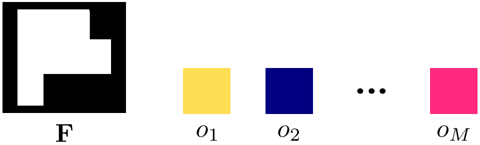
A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.



Each object $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{r}_j, \mathbf{t}_j\}$ is modelled with four random variables that describe their **category, size, orientation and location**.

Scene Parametrization

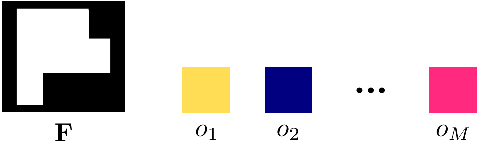
A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.



Each object $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{r}_j, \mathbf{t}_j\}$ is modelled with four random variables that describe their **category, size, orientation and location**.

Scene Parametrization

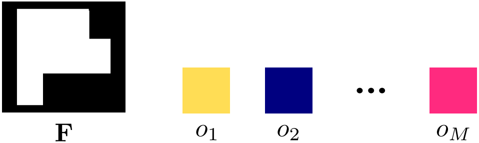
A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.



Each object $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{r}_j, \mathbf{t}_j\}$ is modelled with four random variables that describe their **category, size, orientation and location**.

Scene Parametrization

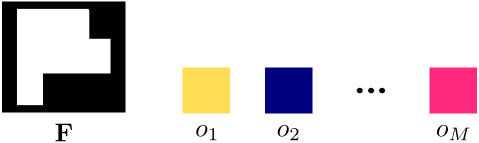
A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.



Each object $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{r}_j, \mathbf{t}_j\}$ is modelled with four random variables that describe their **category, size, orientation and location**.

Scene Parametrization

A scene comprises an **unordered set of M objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.

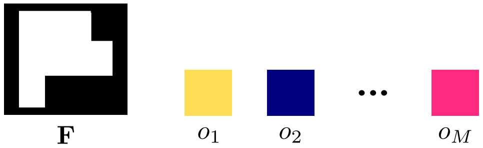


Each object $o_j = \{\mathbf{c}_j, \mathbf{s}_j, \mathbf{r}_j, \mathbf{t}_j\}$ is modelled with four random variables that describe their **category, size, orientation and location**.

$$\underbrace{p_\theta(o_j \mid o_{<j}, \mathbf{F})}_{\text{Probability of generating j-th object}} = p_\theta(\mathbf{c}_j | o_{<j}, \mathbf{F}) p_\theta(\mathbf{t}_j | \mathbf{c}_j, o_{<j}, \mathbf{F}) p_\theta(\mathbf{r}_j | \mathbf{c}_j, \mathbf{t}_j, o_{<j}, \mathbf{F}) p_\theta(\mathbf{s}_j | \mathbf{c}_j, \mathbf{t}_j, \mathbf{r}_j, o_{<j}, \mathbf{F})$$

Scene Parametrization

A scene comprises an **unordered set of M objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.



The **likelihood** of generating a scene **with any order** is:

$$\underbrace{p_\theta(\mathcal{O}|\mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with any order}}} = \sum_{\hat{\mathcal{O}} \in \pi(\mathcal{O})} \underbrace{\prod_{j \in \hat{\mathcal{O}}} p_\theta(o_j \mid o_{<j}, \mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with order } \hat{\mathcal{O}}}}$$

where $\pi(\mathcal{O})$ is a a permutation function that computes the set of permutations of all objects $\mathcal{O}$ in the scene.

Scene Parametrization

A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.

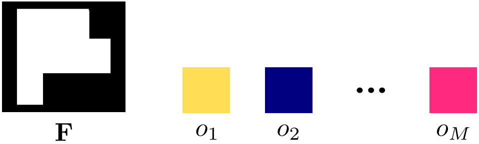
The **likelihood** of generating a scene **with any order** is:

$$\underbrace{p_\theta(\mathcal{O}|\mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with any order}}} = \sum_{\hat{\mathcal{O}} \in \pi(\mathcal{O})} \underbrace{\prod_{j \in \hat{\mathcal{O}}} p_\theta(o_j \mid o_{<j}, \mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with order } \hat{\mathcal{O}}}}$$

where $\pi(\mathcal{O})$ is a a permutation function that computes the set of permutations of all objects $\mathcal{O}$ in the scene.

Scene Parametrization

A scene comprises an **unordered set of M objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.

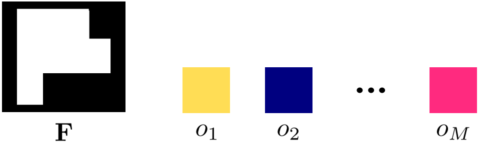


The **likelihood** of generating a scene **with any order** is:

$$\underbrace{p_\theta(\mathcal{O}|\mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with any order}}} = \boxed{\sum_{\hat{\mathcal{O}} \in \pi(\mathcal{O})}} \underbrace{\prod_{j \in \hat{\mathcal{O}}} p_\theta(o_j \mid o_{<j}, \mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with order } \hat{\mathcal{O}}}}$$

Scene Parametrization

A scene comprises an **unordered set of M objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.



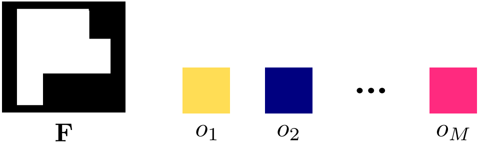
The **likelihood** of generating a scene **with all orders** is:

$$\underbrace{\hat{p}_\theta(\mathcal{O}|\mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with all orders}}} = \prod_{\hat{\mathcal{O}} \in \pi(\mathcal{O})} \underbrace{\prod_{j \in \hat{\mathcal{O}}} p_\theta(o_j \mid o_{<j}, \mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with order } \hat{\mathcal{O}}}}$$

ATISS is trained to **maximize the log-likelihood of all possible permutations of object arrangements** in a collection of scenes.

Scene Parametrization

A scene comprises an **unordered set of** M **objects** $\mathcal{O} = \{o_j\}_{j=1}^{M}$ and its **floor shape** $\mathbf{F}$.

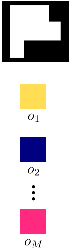


The **log-likelihood** of generating a scene **with all orders** is:

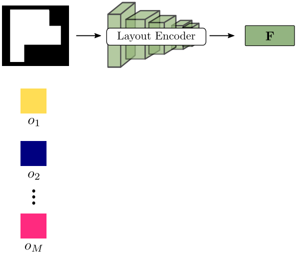
$$\underbrace{\log \hat{p}_\theta(\mathcal{O}|\mathbf{F})}_{\substack{\text{Log-likelihood of generating } \mathcal{O} \\ \text{with all orders}}} = \sum_{\hat{\mathcal{O}} \in \pi(\mathcal{O})} \underbrace{\sum_{j \in \hat{\mathcal{O}}} \log p_\theta(o_j \mid o_{<j}, \mathbf{F})}_{\substack{\text{Probability of generating } \mathcal{O} \\ \text{with order } \hat{\mathcal{O}}}}$$

ATISS is trained to **maximize the log-likelihood of all possible permutations of object arrangements** in a collection of scenes.

Scene Generation

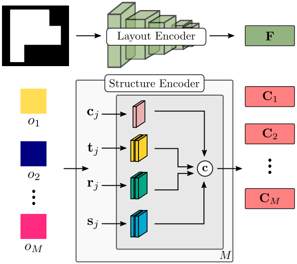


Scene Generation



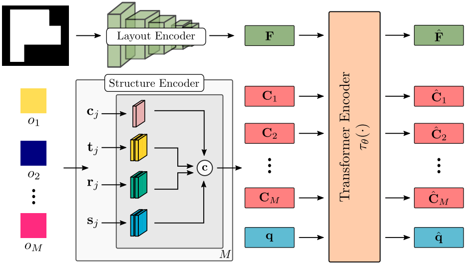
- **Layout encoder**: Computes a global feature representation for the floor.

Scene Generation



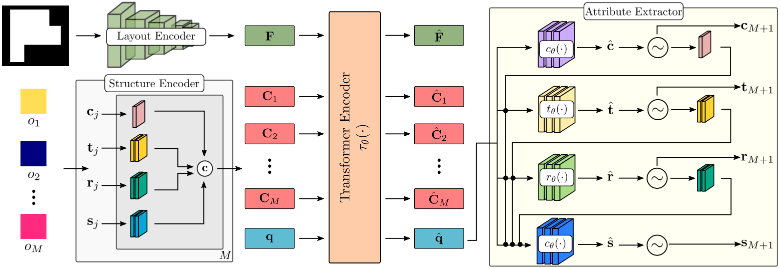
- **Layout encoder**: Computes a global feature representation for the floor.
- **Structure encoder**: Maps the j-th object to a per-object context embedding $\mathbf{C}_j$.

Scene Generation



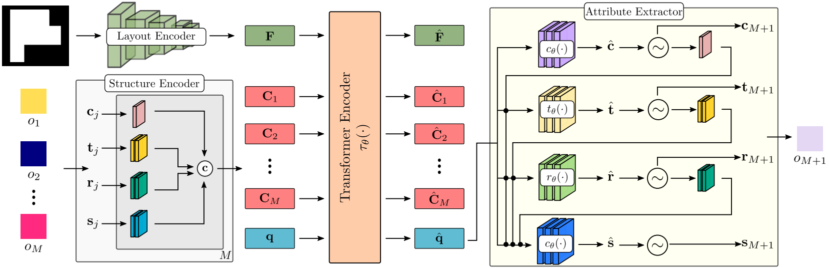
- **Layout encoder**: Computes a global feature representation for the floor.
- **Structure encoder**: Maps the j-th object to a per-object context embedding $\mathbf{C}_j$.
- **Transformer encoder**: Takes $\mathbf{F}, \{\mathbf{C}_j\}_{j=1}^{M}, \mathbf{q}$ and predicts the features $\hat{\mathbf{q}}$ of the next object to be added in the scene.

Scene Generation



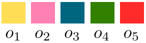
- **Layout encoder**: Computes a global feature representation for the floor.
- **Structure encoder**: Maps the j-th object to a per-object context embedding $\mathbf{C}_j$.
- **Transformer encoder**: Takes $\mathbf{F}, \{\mathbf{C}_j\}_{j=1}^{M}, \mathbf{q}$ and predicts the features $\hat{\mathbf{q}}$ of the next object to be added in the scene.
- **Attribute extractor**: Predicts the object attributes of the next object.

Scene Generation

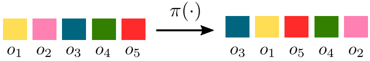


- **Layout encoder**: Computes a global feature representation for the floor.
- **Structure encoder**: Maps the j-th object to a per-object context embedding $\mathbf{C}_j$.
- **Transformer encoder**: Takes $\mathbf{F}, \{\mathbf{C}_j\}_{j=1}^M, \mathbf{q}$ and predicts the features $\hat{\mathbf{q}}$ of the next object to be added in the scene.
- **Attribute extractor**: Predicts the object attributes of the next object.

Training Overview

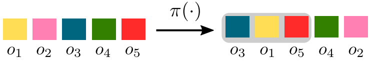


Training Overview



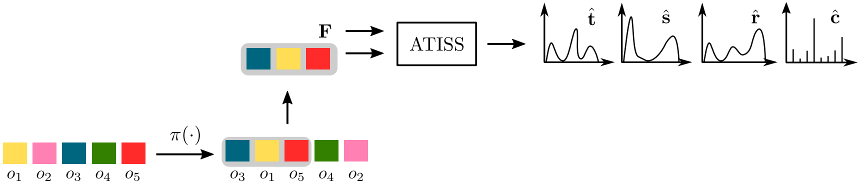
- Randomly permute the M objects of a scene.

Training Overview



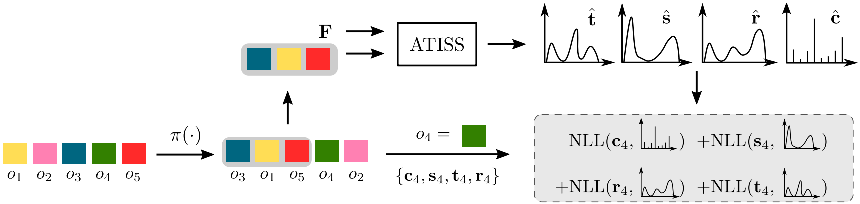
- Randomly permute the M objects of a scene.
- Randomly select the first T objects to compute the context embedding $\mathbf{C}$.

Training Overview



- Randomly permute the M objects of a scene.
- Randomly select the first T objects to compute the context embedding $\mathbf{C}$.
- Conditioned on the $\mathbf{C}$ and $\mathbf{F}$, ATISS **predicts the attribute distributions of the next object**.

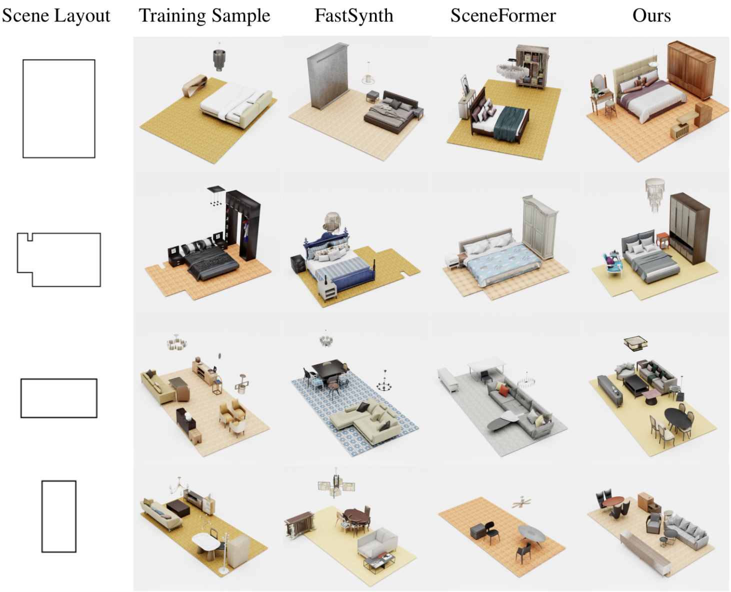
Training Overview



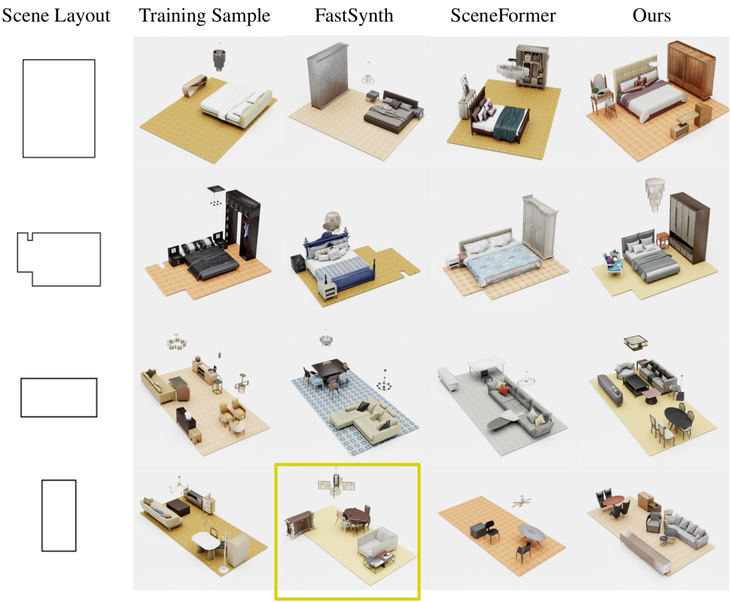
- Randomly permute the M objects of a scene.
- Randomly select the first T objects to compute the context embedding $\mathbf{C}$.
- Conditioned on the $\mathbf{C}$ and $\mathbf{F}$, ATISS **predicts the attribute distributions of the next object**.
- ATISS is trained to maximize the log likelihood of the $T + 1$ object from the permuted set of objects.

How well does it work?

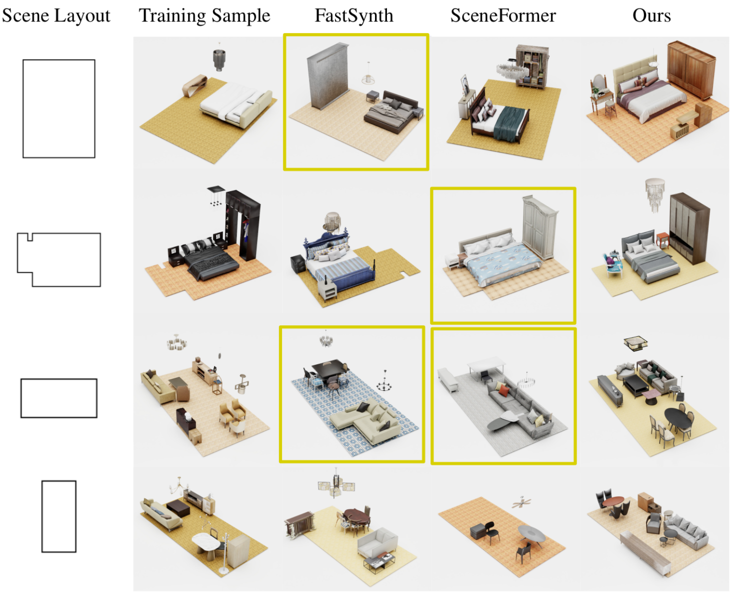
Scene Synthesis



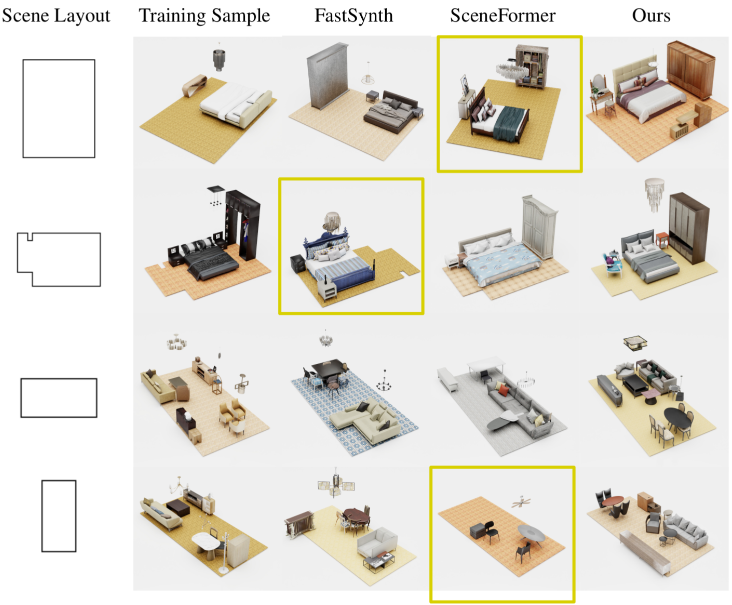
Scene Synthesis



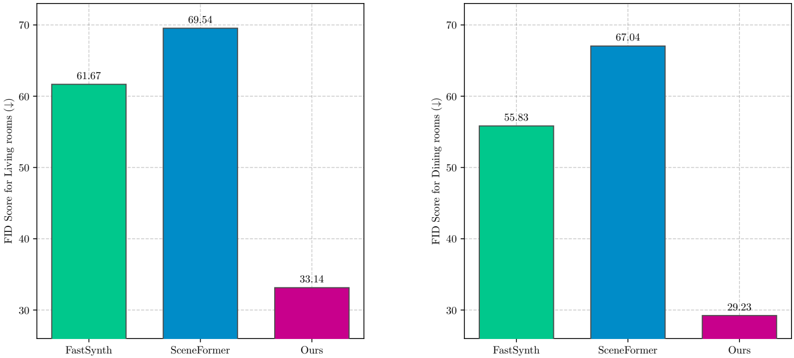
Scene Synthesis



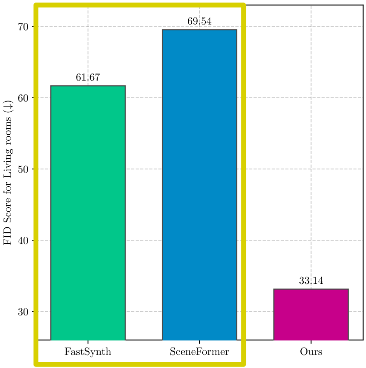
Scene Synthesis

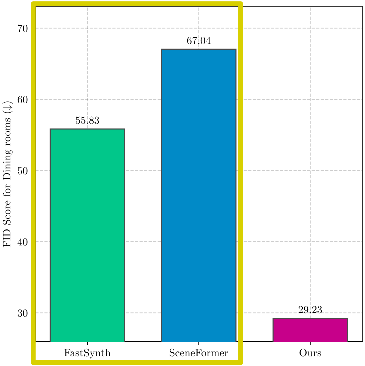


Scene Synthesis

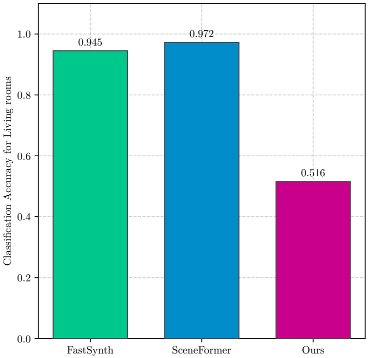


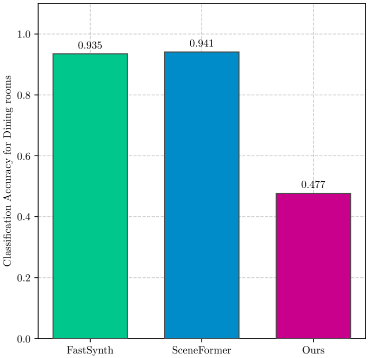
Scene Synthesis



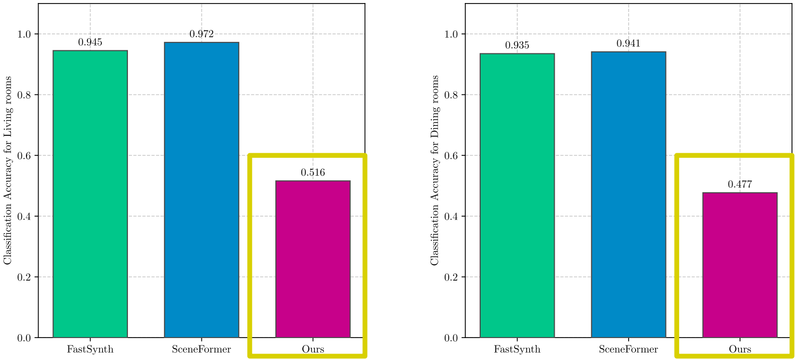


Scene Synthesis

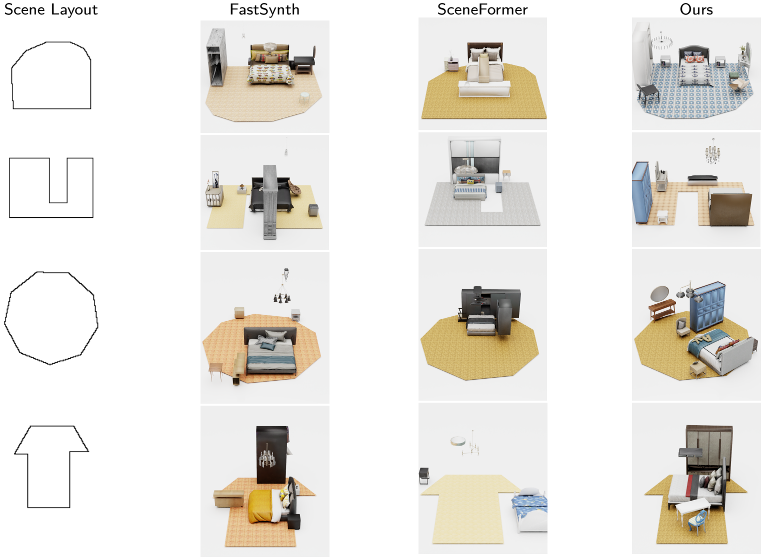




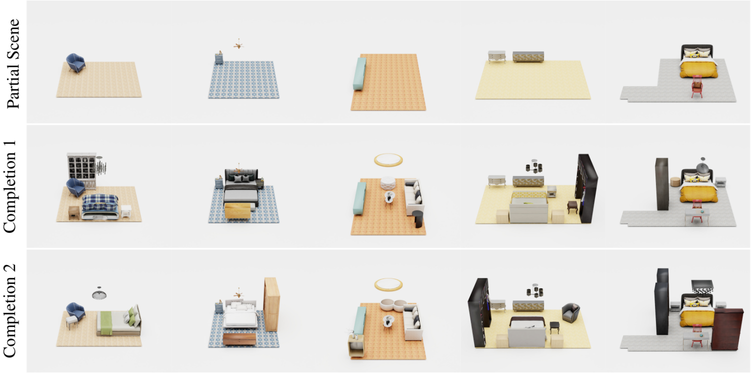
Scene Synthesis



Generalization Beyond Training Data

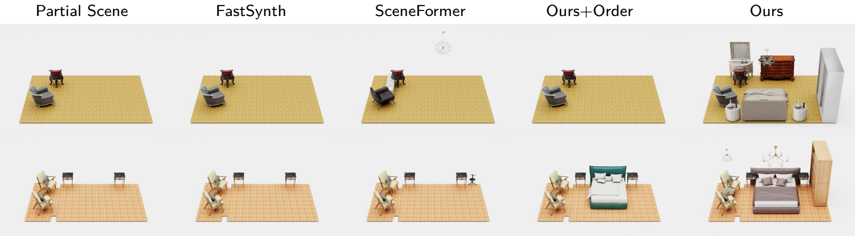


Scene Completion



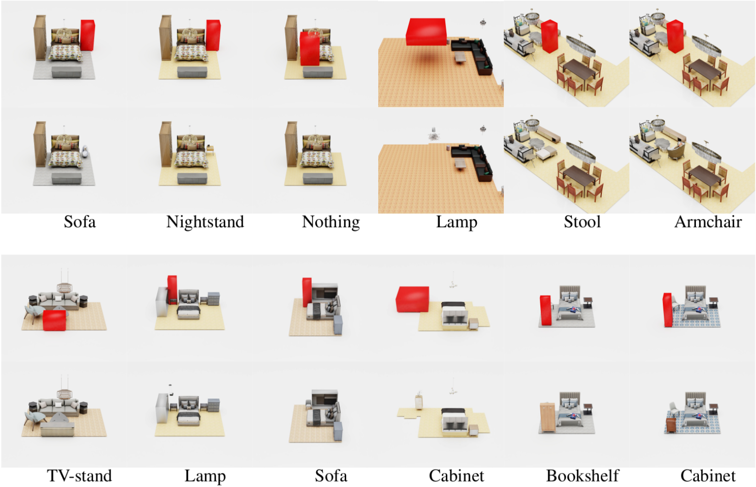
Scene Completion

FastSynth and SceneFormer can only generate objects in the order they were trained with. As a result, starting from partial scenes with less common objects, both models fail to generate plausible object arrangements.



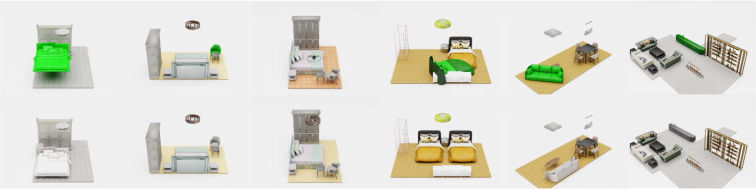
Objects Suggestion

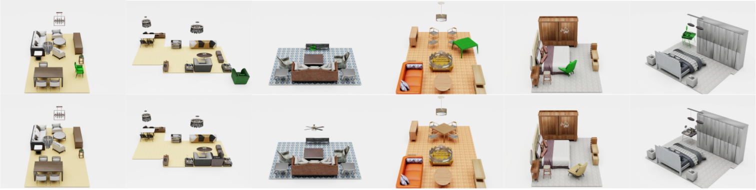
A user specifies **a region of acceptable positions to place an object**, marked as a red box and **our model suggests suitable objects to be placed at this location**. To perform this task, we compute the likelihood of an object conditioned on an arbitrary scene.



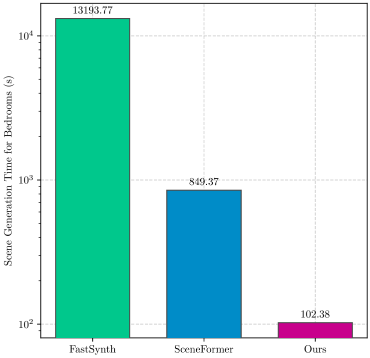
Failure Cases Correction

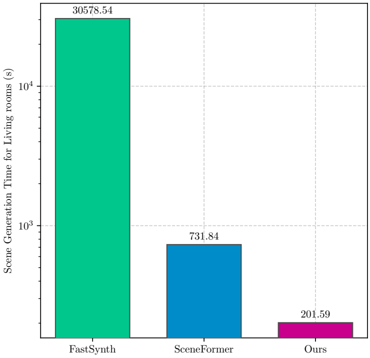
Our model **identifies and corrects unnatural object arrangements in a scene**. To identify such objects, our model **computes the likelihood of each object conditioned on the other objects** in the scene and objects with low likelihood are identified as problematic. For these objects a new location is sampled.





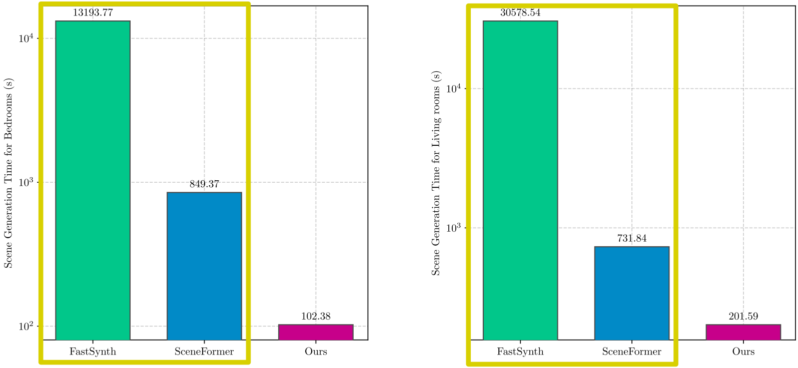
Generation Time





- At least $100\times$ faster than the CNN-based FastSynth for all room types.
- At least $4\times$ faster than the Transformer-based SceneFormer for all room types.

Generation Time



- At least $100\times$ faster than the CNN-based FastSynth for all room types.
- At least $4\times$ faster than the Transformer-based SceneFormer for all room types.

Summary

- We propose ATISS **a novel autoregressive model for unordered set generation**.

Summary

- We propose ATISS **a novel autoregressive model for unordered set generation**.
- We demonstrate that our unordered set formulation **opens up multiple interactive applications**.

Summary

- We propose ATISS **a novel autoregressive model for unordered set generation**.
- We demonstrate that our unordered set formulation **opens up multiple interactive applications**.
- ATISS has fewer parameters, **is simpler to implement and train and runs up to 8x faster** than existing methods.

Summary

- We propose ATISS **a novel autoregressive model for unordered set generation**.
- We demonstrate that our unordered set formulation **opens up multiple interactive applications**.
- ATISS has fewer parameters, **is simpler to implement and train and runs up to 8x faster** than existing methods.
- Limitations:

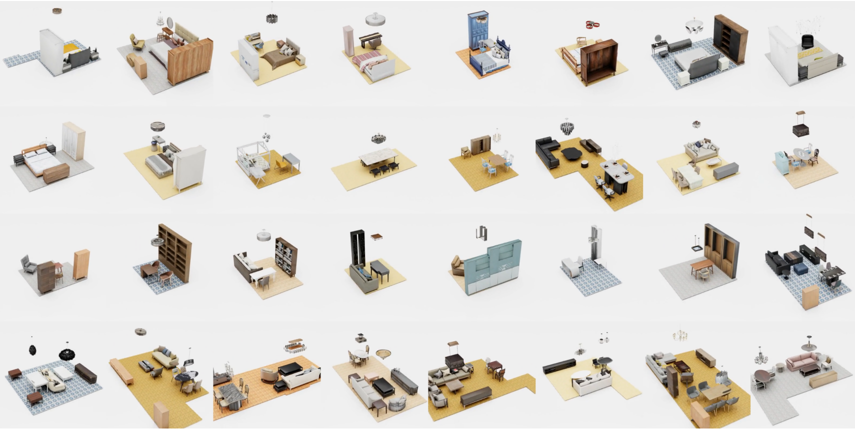
Summary

- We propose ATISS **a novel autoregressive model for unordered set generation**.
- We demonstrate that our unordered set formulation **opens up multiple interactive applications**.
- ATISS has fewer parameters, **is simpler to implement and train and runs up to 8x faster** than existing methods.
- Limitations:
  - The autoregressive generation of attributes need to follow a specific ordering.

Summary

- We propose ATISS **a novel autoregressive model for unordered set generation**.
- We demonstrate that our unordered set formulation **opens up multiple interactive applications**.
- ATISS has fewer parameters, **is simpler to implement and train and runs up to 8x faster** than existing methods.
- Limitations:
  - The autoregressive generation of attributes need to follow a specific ordering.
  - Separate object retrieval module.

Check out our project page for code and additional results!



https://nv-tlabs.github.io/ATISS