HOLODECK: Language-Guided 3D Embodied AI Environment Generation

Yue Yang*[1], Fan-Yun Sun*[2], Luca Weihs*[4], Eli Vanderbilt[4], Alvaro Herrasti[4], Winson Han[4], Jiajun Wu[2], Nick Haber[2], Ranjay Krishna[3,4], Lingjie Liu[1], Chris Callison-Burch[1], Mark Yatskar[1], Aniruddha Kembhavi[3,4], Christopher Clark[4]

CVPR 2024 | Project Page

Presenter: **, University of Texas at Austin

1. University of Pennsylvania, 2. Stanford University, 3. University of Washington, 4. Allen Institute for Artificial Intelligence

📌 Existing Approaches in Embodied AI Environment Creation

TEXAS
The University of Texas at Austin

- 🎨 **Manual Design**: Handcrafted 3D scenes created by artists
  ⚠️ **Challenges**: Expensive, Labor-intensive and slow production

- 📸 **3D Scanning**: Captures real-world scenes through scanning
  ⚠️ **Challenges**: Limited interactivity and customization

- ⚙️ **Procedural Generation** (e.g., PROCTHOR): Uses rule-based algorithms to generate environments
  ⚠️ **Challenges**: Rigid layouts with limited fine-grained detail

- 🖼️ **2D-to-3D Model Adaptation**: Converts 2D images into 3D scenes
  ⚠️ **Challenges**: Prone to artifacts and unrealistic results

🚀 HOLODECK: Revolutionizing 3D Scene Creation

TEXAS
The University of Texas at Austin

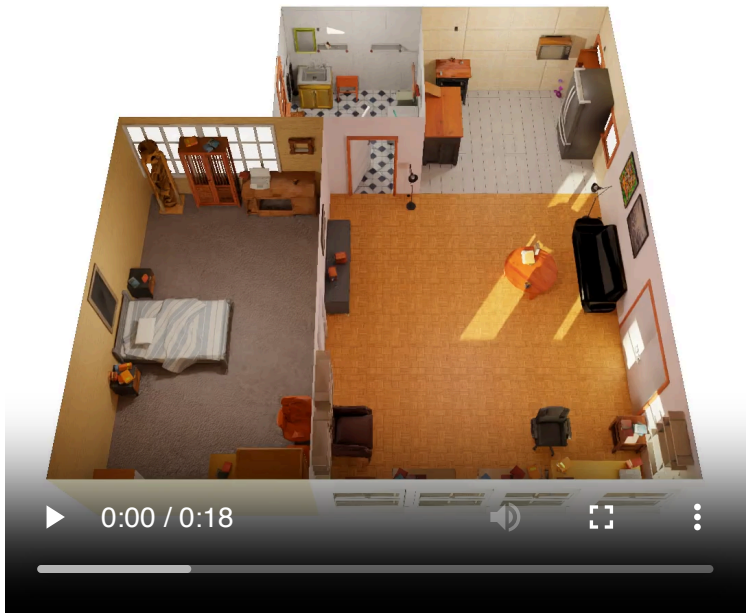
- 💡 **LLM-Powered 3D Environment Generation**
  Seamless integration of LLMs (e.g., GPT-4) with vast 3D asset libraries
  Transforms natural language prompts into **high-fidelity** 3D environments
  Fine-grained control over layout, style, and scene semantics

- 📝 **Intelligent Text-Based Scene Design**
  Rapid synthesis of **diverse, interactive, and customizable** scenes
  Constraint-based layout optimization ensures realistic object placement

- 🚀 **Breakthrough in AI Training & Simulation**
  Scales up 3D scene generation **without manual effort**
  Optimized object placement for better simulation accuracy
  Enables more efficient training of embodied AI agents

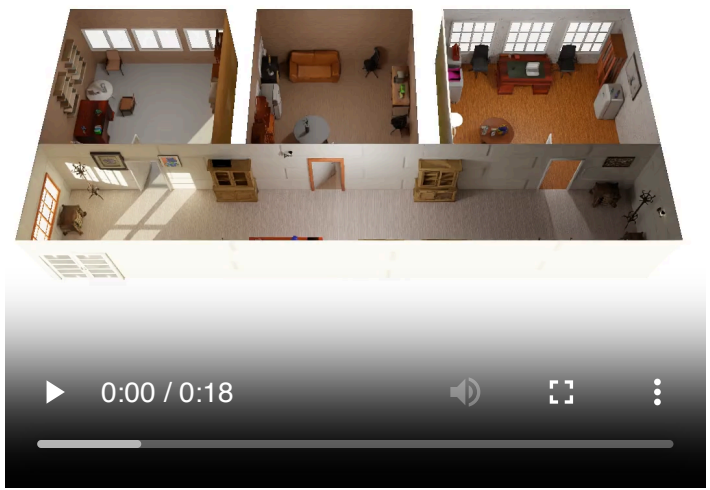
🏠 HOLODECK Demo

- A 1B1B apartment of a researcher who has a Cat 🐱



HOLODECK Demo

- Three professors' office connected to a long hallway, the professor in office 1 is a fan of Star Wars.



LLM's Role in Scene Generation 🚀🤖

- **Abstract-to-Concrete Mapping** 🎨
  Input: "a spa with large hot tubs and massage tables"
  Output: LLM translates this abstract concept into detailed scene elements (e.g., specific models, spatial arrangements)

- **Common Sense Knowledge** 💡
  Automatically includes expected items (e.g., musical instruments 🎹 in a music room)

- **Interactivity** 💬
  Allows dynamic adjustments of parameters (e.g., color, position) through dialogue

- **Spatial Reasoning** 🏠
  Understands physical relationships (e.g., cups on tables ☕, objects adjacent to each other) Uses spatial constraints and a solver to optimize layouts, ensuring realistic scene compositions.

HOLODECK Pipeline Overview

- **Floor & Wall Modules:**
  Develop floor plans, construct wall structures, and select appropriate materials for the floors and walls.

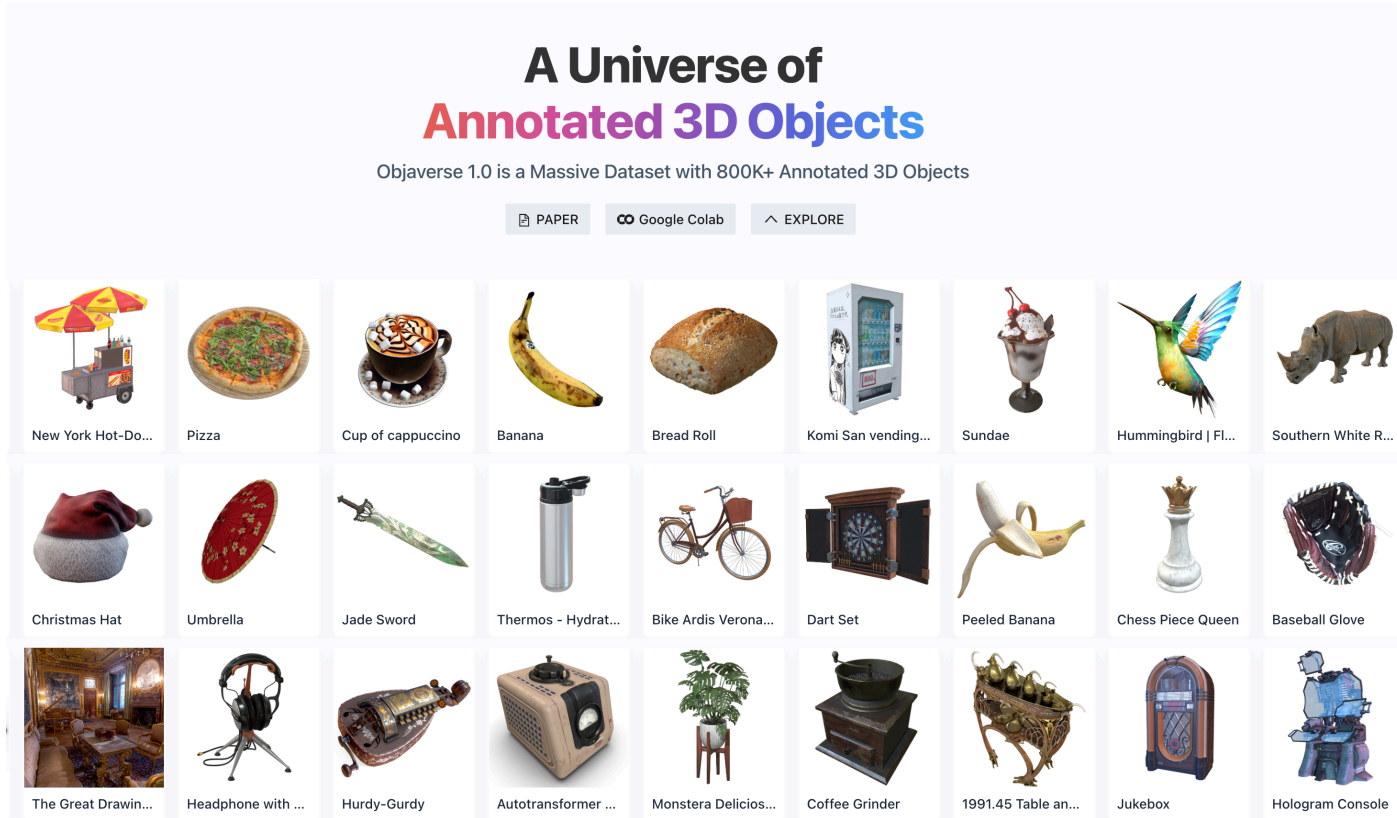
- **Door & Window Module:**
  Integrate doorways and windows into the environment.

- **3D Asset Selection:**
  Retrieve appropriate 3D assets froms from Objaverse.[1].

- **Constraint-Based Layout Design:**
  Arrange the assets within the scene by utilizing spatial relational constraints to ensure that the layout of objects is realistic.

[1]: Objaverse is a comprehensive dataset comprising over 800,000 annotated 3D objects, facilitating various AI and machine learning applications

Objaverse 1.0



Leveraging Objaverse Assets for HOLODECK

- **Asset Curation**:
  Curating a subset of indoor design assets from Objaverse 1.0.

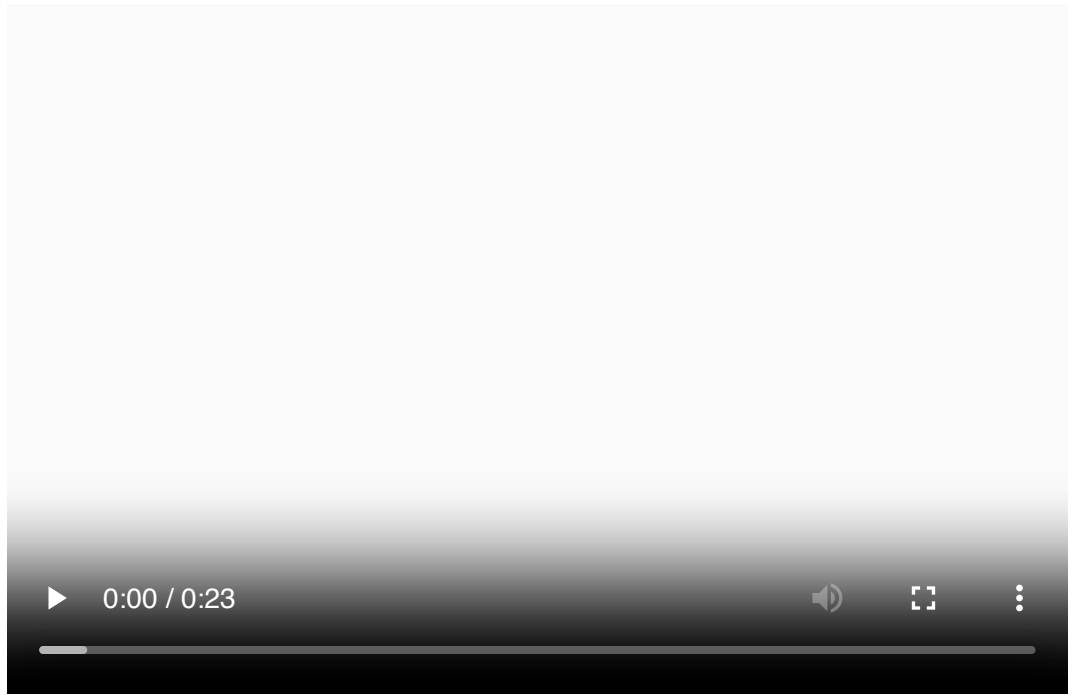
- **Automatic Annotation**:
  Annotating assets using GPT-4-Vision with details such as textual descriptions, scale, and canonical views.

- **Library Size**:
  The library includes 51,464 annotated assets, combining Objaverse and PROCTHOR assets.

- **Optimization for AI2-THOR**:
  Reducing mesh counts, generating visibility points, and adding colliders to minimize loading time for embodied AI applications.

HOLODECK Pipeline Overview

- Given any text input, Holodeck generates 3D interactive embodied environments by utilizing a series of specialized modules through multiple rounds of conversation with an LLM (GPT-4).



Overall Prompt Design

- **Task Description**:
  Outlines the context and goals of the task.

- **Output Format**:
  Specifies the expected structure and type of outputs and

- **One-shot Example**:
  A concrete example to assist the LLM's comprehension of the task.

Floor Plan Prompt

```
Floor plan Prompt: You are an experienced room designer.
Please assist me in crafting a floor plan. Each room is a rectangle.
You need to define the four coordinates and specify an appropriate design scheme, including each room's color, ma
Assume the wall thickness is zero.
Please ensure that all rooms are connected, not overlapped, and do not contain each other.
The output should be in the following format:
room name | floor material | wall material | vertices (coordinates).
Note: the units for the coordinates are meters.
For example:
living room | maple hardwood, matte | light grey drywall, smooth | [(0, 0), (0, 8), (5, 8), (5, 0)]
kitchen | white hex tile, glossy | light grey drywall, smooth | [(5, 0), (5, 5), (8, 5), (8, 0)]

Here are some guidelines for you:
1. A room's size range (length or width) is 3m to 8m. The maximum area of a room is 48 m^2.
   Please provide a floor plan within this range and ensure the room is not too small or too large.
2. It is okay to have one room in the floor plan if you think it is reasonable.
3. The room name should be unique.

Now, I need a design for {input}.
Additional requirements: {additional requirements}.
Your response should be direct and without additional text at the beginning or end.
```

Wall Height Prompt

```
I am now designing {input}.
Please help me decide the wall height in meters.
Answer with a number, for example, 3.0.
Do not add additional text at the beginning or in the end.
```

Floor & Wall Module

- **Room Definition**:
  Rooms are rectangles, defined by coordinates. **GPT-4** provides room placement, dimensions, and connectivity.

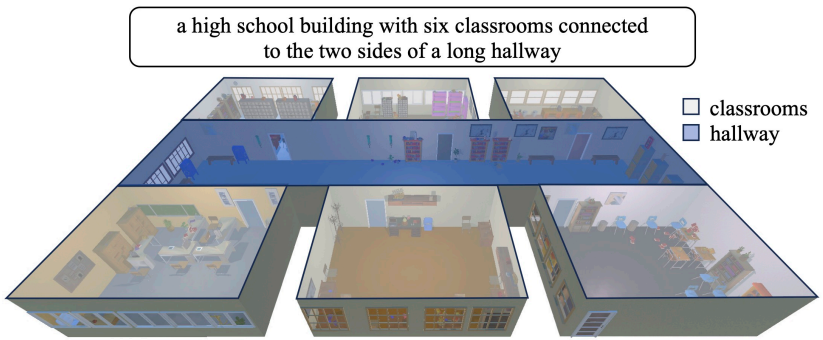
- **Layout Generation**:
  Produces multi-room layouts with constraints (e.g., minimum area of 9 m^2).

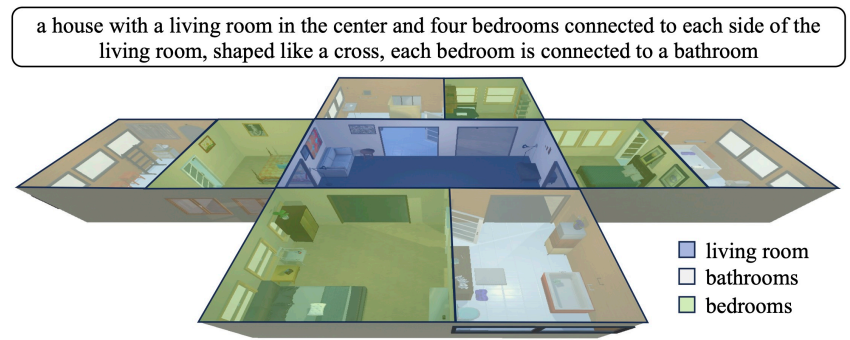
- **Material Selection**:
  Chooses from 236 materials and 148 colors for floor and wall customization.

- **Contextual Customization**:
  Adapts materials to scene type (e.g., **concrete** for cells, **bricks** for walls).

Floorplan Customizability

- HOLODECK can interpret complicated input and craft reasonable floor plans correspondingly.

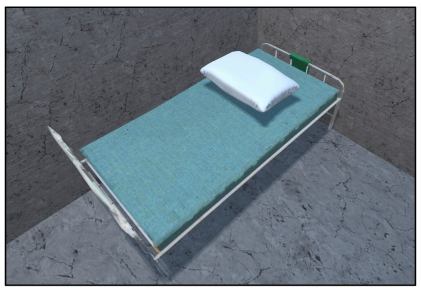




Material Customizability

- HOLODECK can select appropriate floor and wall materials to make the scenes more realistic.

a prison cell



a bedroom of a girl who loves the **pink color**

a wine cellar with **red wall bricks**

a 80s bar with **checkered** flooring

Doorway & Window Module

- **Room Connections**:
  Proposes doorways and windows based on LLM queries.

- **Door & Window Types**:
  40 door styles and 21 window types, each customizable by size, height, quantity, and more.

- **Tailored Designs**:
  Adapts to specific needs, e.g., wider doors for **wheelchair accessibility** or **floor-to-ceiling windows** in a **sunroom**.

Door & window Customizability

- HOLODECK can adjust the size, quantity, position, etc., of doors & windows based on the input.

an apartment for a disabled person who needs to use wheelchair

a sunroom with floor-to-ceiling windows covering all walls

Object Selection Module

- **Asset Retrieval**:
  Utilizes the extensive **Objaverse** asset collection to place objects in the scene.

- **LLM Integration**:
  Proposes objects based on descriptions and dimensions, e.g., "multi-level cat tower, 60×60×180cm."

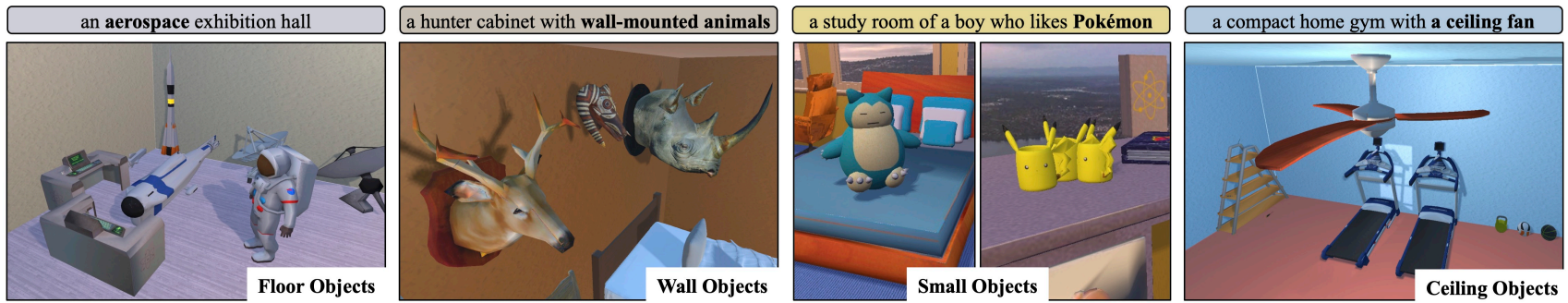
- **Retrieval Function**[2]:
  Considers visual and textual similarity and dimensions to ensure the assets match the design.

- **Custom Object Placement**:
  Places objects on floors, walls, other items, or even ceilings, enhancing scene customization.

[2] They use CLIP to measure the visual similarity, Sentence-BERT for the textual similarity, and 3D bounding box sizes for the dimension.

Objects Customizability

- HOLODECK can select and place appropriate floor/wall/small/ceiling objects conditioned on the input.



Constraint-based Layout Design Module

- **Objective:**
  Generates object positioning and orientation for layouts.

- **Challenge with Direct Bounding Boxes:**
  LLMs can provide absolute bounding box values.
  This often leads to out-of-bound errors and object collisions.

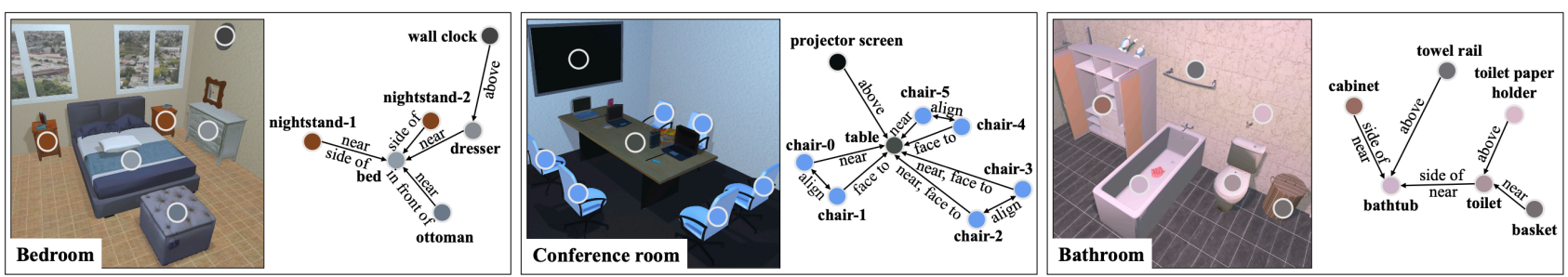
- **Novel Approach:**
  Uses LLM to generate **spatial relations** (e.g., "coffee table, in front of, sofa") instead of numerical values.
  Convert these relations into mathematical conditions.
  Use an optimization algorithm (with **DFS**) to place objects sequentially.

Constraint-based Layout Design Module

- **Spatial Constraints:**
  Types: Global (edge, middle), Distance (near, far), Position (in front of, side of, above, on top of), Alignment (center aligned), Rotation (face to).

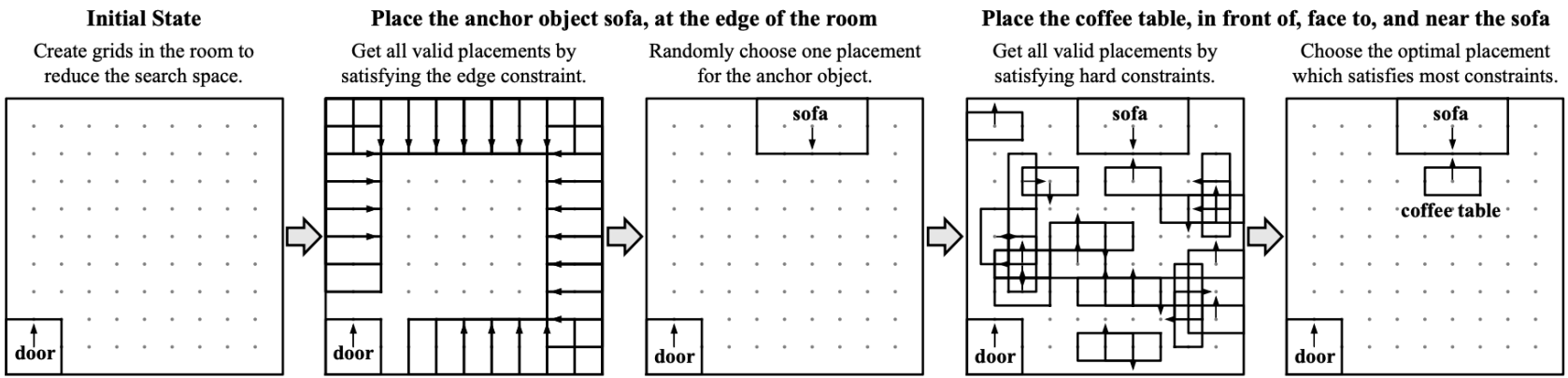
- **LLM** selects a subset of constraints for each object, forming a scene graph for the room. **Handling:**\ a. **Soft constraints:** Allow minor violations when perfect satisfaction isn't feasible.
  b. **Hard constraints:** Must be met (no collisions, all objects within boundaries).

- Examples of **Spatial Relational Constraints** generated by LLM and their solutions found by our constraint satisfaction algorithm.



DFS-based Constraint Satisfaction Example

a. DFS solver initiates grids to establish a finite search space.
b. First explore different placements for the **anchor** object selected by the **LLM**
c. Optimize the placement for the remaining objects, adhering to the hard constraints, and satisfying as many soft constraints as possible.
d. Generate multiple solutions, with the final selection meeting the most constraints.



Output Diversity.

- HOLODECK can generate **multiple variants** for the same input with different assets and layouts.

Human Evaluation Overview

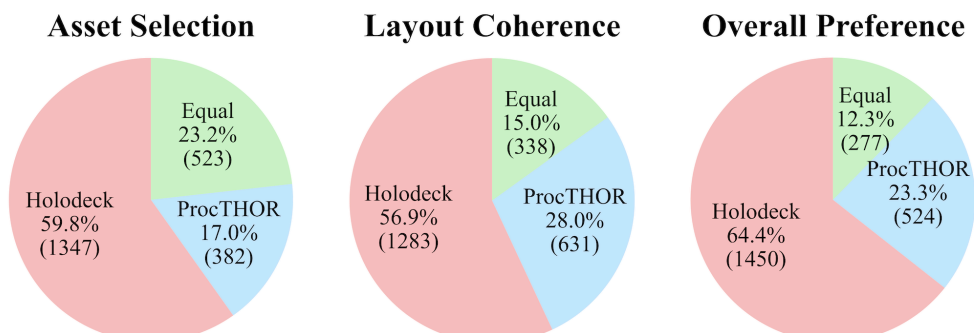
- **Participants**: 680 graduate students took part in three user studies.
- **Studies**:
  - A comparative study on residential scenes (vs. PROCTHOR), including human evaluations and CLIP Score experiments.

    a. 120 paired scenes (30 per type: bathroom, bedroom, kitchen, living room)

    b. The PROCTHOR baseline has access to the same set of Objaverse assets as HOLODECK.
  - An assessment of HOLODECK's performance on 52 diverse scene types.
  - An ablation study on layout design methods, comparing the Spatial Constraint approach with Absolute, Random, and Edge strategies.
- **Outcome**: HOLODECK produces scenes with higher quality and greater diversity.

Comparative Analysis on Residential Scenes (Human Evaluation)

- **Evaluation Criteria**:
  - **Asset Selection**: Faithfulness to scene type
  - **Layout Coherence**: Realism in arrangement
  - **Overall Preference**
- **Results**: HOLODECK preferred in: Asset Selection (59.8%), Layout Coherence (56.9%), Overall Preference (64.4%)

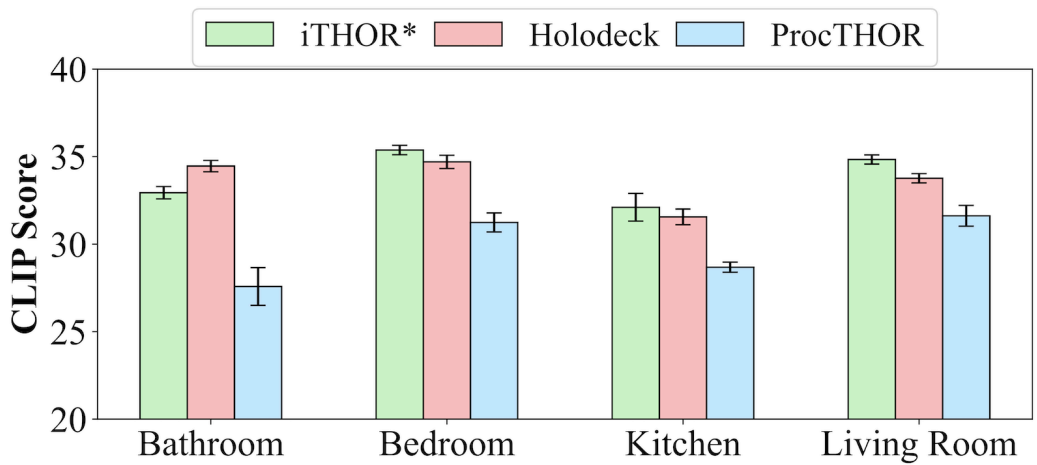


The pie charts show the distribution of annotator preferences, showing both the percentage and the actual number of annotations favoring each system.

Comparative Analysis on Residential Scenes (CLIP Score Details)

- **CLIP Score Experiment**:

  Quantifies the visual coherence between the scene's top-down view and its scene type prompt ("a top-down view of [scene type]").
- **Reference**: Human-designed scenes from iTHOR are used as the upper bound.
- HOLODECK Significantly outperforms PROCTHOR; closely approaches iTHOR performance.



HOLODECK's Performance on Diverse Scenes

- 52 scene types from the MIT Scenes Dataset:

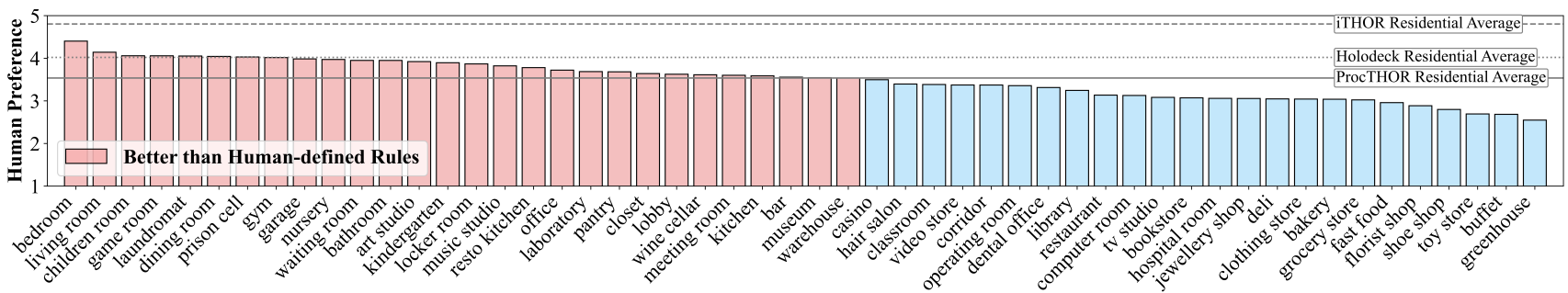
  Stores (e.g., deli, bakery), Home (e.g., bedroom, dining room), Public Spaces (e.g., museum, locker room), Leisure (e.g., gym, casino), Working Space (e.g., office, meeting room)

HOLODECK's Performance on Diverse Scenes (Human Evaluation)

- 52 scene types from the MIT Scenes Dataset:
- **Methodology**:
  - Generated five outputs per scene type, totaling 260 scenes.
  - Human annotators rated scenes based on asset selection, layout coherence, and match with scene type (1 to 5 scale).
  - Compared with residential scenes from PROCTHOR and iTHOR.

HOLODECK's Performance on Diverse Scenes (Human Evaluation)

- **Key Findings**:
  - HOLODECK achieved higher human preference scores in 28 out of 52 diverse scenes.
  - Holodeck can generate satisfactory (better than hard-coded rules) outputs for diverse scene types.
  - Demonstrated broad competence and flexibility, outperforming PROCTHOR in many cases.
  - Struggled with scenes requiring complex layouts or unique assets (e.g., dental x-ray machine).



Human evaluation on 52 scene types from MIT Scenes with qualitative examples. The three horizontal lines represent the average score of iTHOR, HOLODECK, and PROCTHOR on four types of residential scenes (bedroom, living room, bathroom and kitchen.)

Ablation Study on Layout Design (Human Evaluation)

- **Layout Design Methods (Baselines)**

  a. **CONSTRAINT**: HOLODECK's spatial relational constraint-based layout.

  b. **ABSOLUTE**: Objects' absolute coordinates and orientations derived from LLM (similar to LayoutGPT).

  c. **RANDOM**: Objects are randomly placed within the room without collision.

  d. **EDGE**: Objects are placed along the room's walls.

Method	Bathroom	Bedroom	Kitchen	Living Room	Average
ABSOLUTE	0.369	0.343	0.407	0.336	0.364
RANDOM	0.422	0.339	0.367	0.348	0.369
EDGE	0.596	0.657	**0.655**	0.672	0.645
CONSTRAINT	**0.696**	**0.745**	0.654	**0.728**	**0.706**

HOLODECK's constraint-based approach outperforms the other methods significantly on bathroom, bedroom and living room. CONSTRAINT and EDGE perform similarly on kitchen, where it is common to align most objects against walls.

HOLODECK for ObjectNav: Synthesizing Training Environments

TEXAS
The University of Texas at Austin

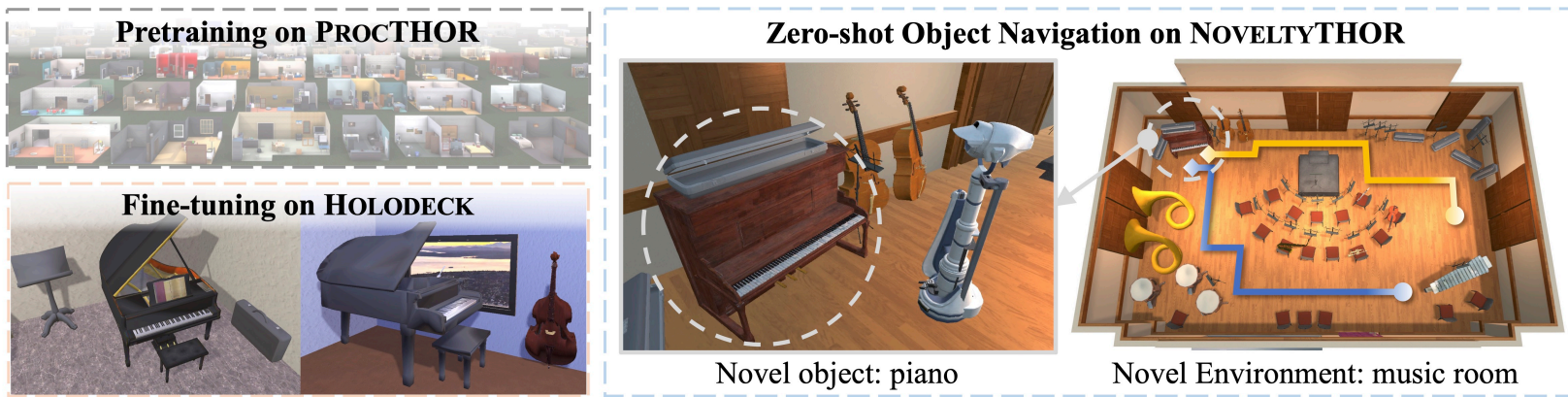
- Holodeck can aid embodied agents in adapting to new scene types and objects during object navigation tasks.

- **NOVELTYTHOR Benchmark:**
  92 object types across 5 categories: Office, Daycare, Music Room, Gym, Arcade.
  Scenes: 100 novel scenes per category generated by HOLODECK.
  Task: Train ObjectNav on HOLODECK scenes and evaluate on NOVELTYTHOR.

- **Model Setup**
  Pre-trained on PROCTHOR-10K, fine-tuned on 100 scenes (50M steps).
  PROCTHOR:
  +HOLODECK: Fine-tune with HOLODECK-generated scenes.
  +OBJAVERSE: Objaverse-enhanced object selection.

Object Navigation in Novel Environments

- Holodeck can aid embodied agents in adapting to new scene types and objects during object navigation tasks.



- **NoveltyTHOR**, an artist-designed benchmark to evaluate embodied agents in diverse scenes.

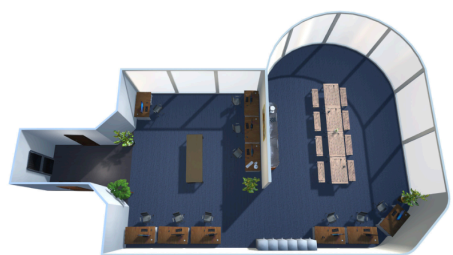
Music Room_01

Music Room_02

Daycare_01

Daycare_02

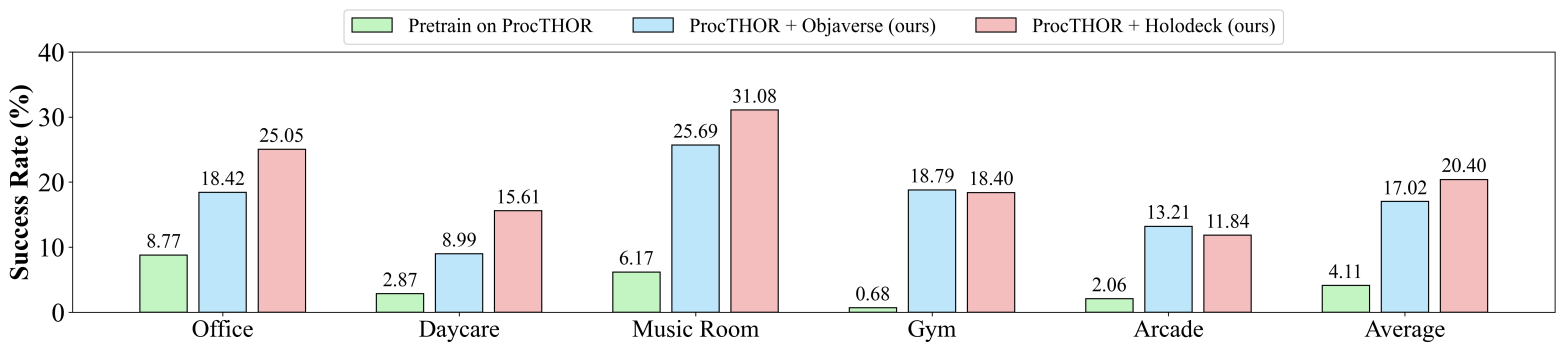
Office_01



Office_02

Results: HOLODECK vs. Baselines

- **HOLODECK** outperforms baselines on **Office**, **Daycare**, and **Music Room**. On Gym and Arcade, +HOLODECK and +OBJAVERSE perform similarly.
- **HOLODECK** creates **human-like object placements** (e.g., Music Room with piano, violin, cellos). **PROCTHOR** struggles due to poor object coverage, often performing close to **random**.
- HOLODECK excels in generating **realistic, commonsense layouts**, boosting agent performance on novel environments.



Conclusion & Future Work

- **HOLODECK: A New Paradigm for Embodied AI**
  - Leverages **large language models** to generate **diverse, interactive** environments.
  - Demonstrated effectiveness through **human evaluation** and **object navigation tasks**.
- **Future Directions**
  - Expanding **3D asset library** for richer scene diversity.
  - Exploring broader **Embodied AI applications**, beyond object navigation.

🚀 **HOLODECK sets the foundation for scalable, realistic AI training environments.**

Motivation & Importance

- Holodeck proposes a fully automated method for creating diverse 3D environments by leveraging large language models (LLMs) and a vast asset library (Objaverse).

- SciTechDaily's article, Not Science Fiction: Researchers Recreate Star Trek's Holodeck Using AI, highlights how AI-generated environments could revolutionize training for robotics, AR/VR, and more.

- The Verge's article on world modeling questions whether language-driven methods can fully capture the 3D physical fidelity and interactivity required for real-world tasks.

- **Critiques:**
  Is reducing manual labor alone sufficient justification for the approach?
  Can LLMs reliably capture the intricate details of 3D spatial structure and physics necessary for embodied AI?

Limitations of the Proposed Approach

- GPT-4 for common-sense spatial reasoning and semantic understanding.
  Constraint-based optimization helps in achieving coherent object placements.

- **Spatial Reasoning & Artifacts:**
  Heavy reliance on LLMs can sometimes result in hallucinations or imprecise object placements.

- **Limitations in Diversity:**
  Current experiments mainly address residential settings, leaving open how the system might handle more diverse or dynamic environments.

- **Critiques:**
  How well does the system perform with ambiguous, extreme, or atypical prompts?
  Would a hybrid approach combining LLM outputs with procedural or physics-based rules mitigate these issues?

Evaluation Methodology & Results

- Extensive human studies (680 participants) indicate a strong preference for Holodeck over procedural baselines like ProcTHOR in terms of asset selection and layout coherence.
  Zero-shot object navigation on the NoveltyTHOR benchmark shows improved agent performance.

- **Limitations & Biases:**
  CLIP scores and human ratings focus on aesthetics and layout but may not reflect real-world usability (Making 3D Scenes Interactive).
  Evaluations by graduate students may introduce cultural bias, lack diversity in perspectives, and fail to assess adaptability across different tasks and environments.

- **Critiques:**
  Can additional quantitative metrics (such as collision rate, physical plausibility scores, or simulation-based benchmarks) better validate performance?
  How might evaluations be expanded to assess diverse environments and dynamic interactions?

Future Research Directions

- **Dynamic and Physics-Aware Simulation:** Expanding from static to dynamic environments, incorporating real-time physics and interactive object behaviors.

- **Hybrid Pipelines:** Integrating LLM-based scene synthesis with procedural and physics-based models can potentially overcome current limitations.

- **Emerging Trends & Additional Resources:**
  Hierarchical Inpainting Approaches: Architect: Generating Vivid and Interactive 3D Scenes with Hierarchical 2D Inpainting proposes a method for refining scene details iteratively, which could enhance realism.
  World Modeling and Scalability: Recent initiatives by companies like DeepMind (The Verge on world modeling) signal the importance of scalable and physically accurate world models.

Future Research Directions

- **Challenges in 3D Content Generation:**
  Challenges and Opportunities in 3D Content Generation discusses the need for scalable, high-fidelity 3D synthesis.

- **Critiques:**
  How can future work improve the real-time interactivity and physical plausibility of AI-generated scenes? What additional domains (such as outdoor environments or dynamic human interactions) should be targeted to broaden the impact in embodied AI?

- **Further Reading and Resources:**
  PhyScene (arXiv)
  Challenges in 3D Content Generation (arXiv)
  Google's World Modeling Efforts (The Verge)

TEXAS
The University of Texas at Austin

Thanks

Q & A