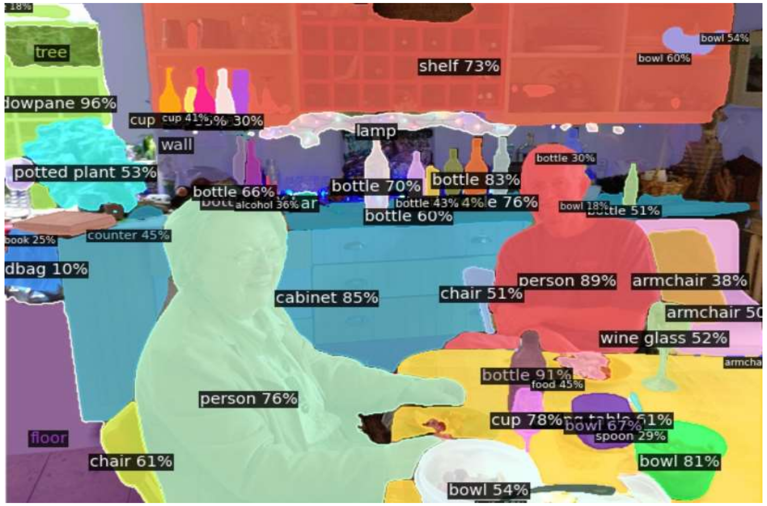
北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

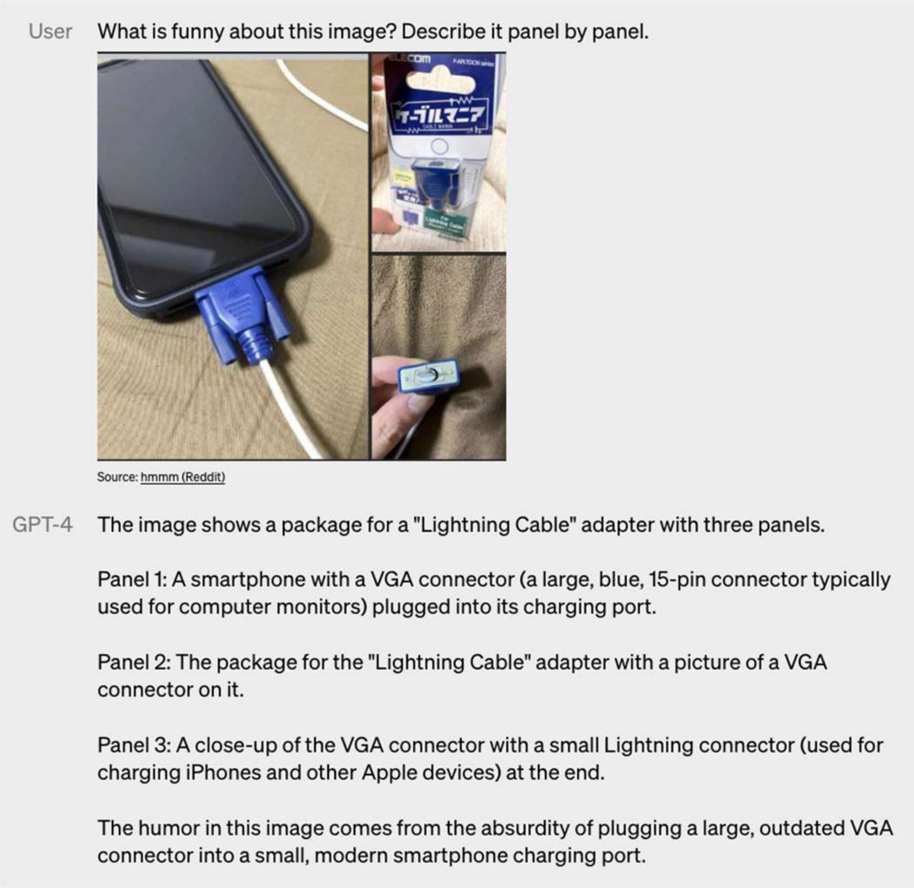
GVL
General Vision Lab

3D Scene Understanding, Generation, and Interaction for Embodied AI

Baoxiong Jia
General Vision Lab, BIGAI

Perception





User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

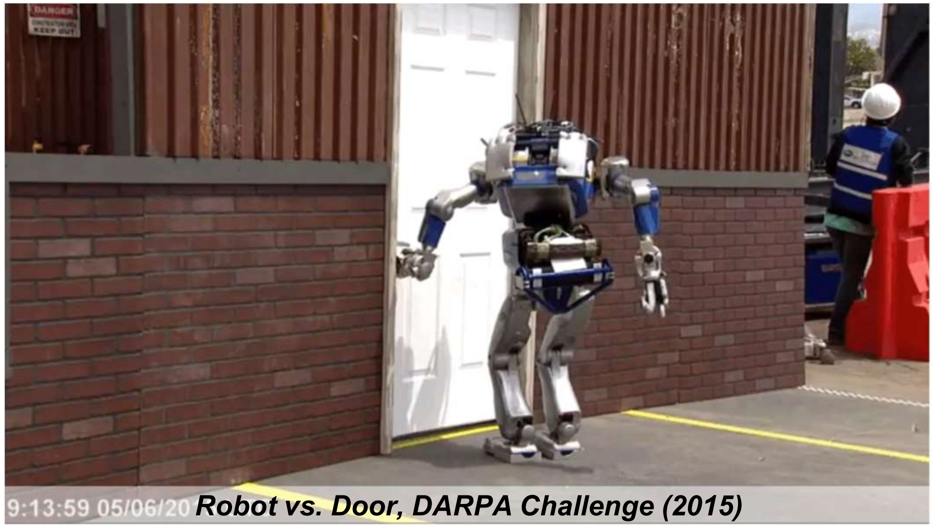
Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Embodied? ☹



Robot vs. Door, DARPA Challenge (2015)

Fu et al., Mobile-ALOHA (2024)

Moravec’s Paradox

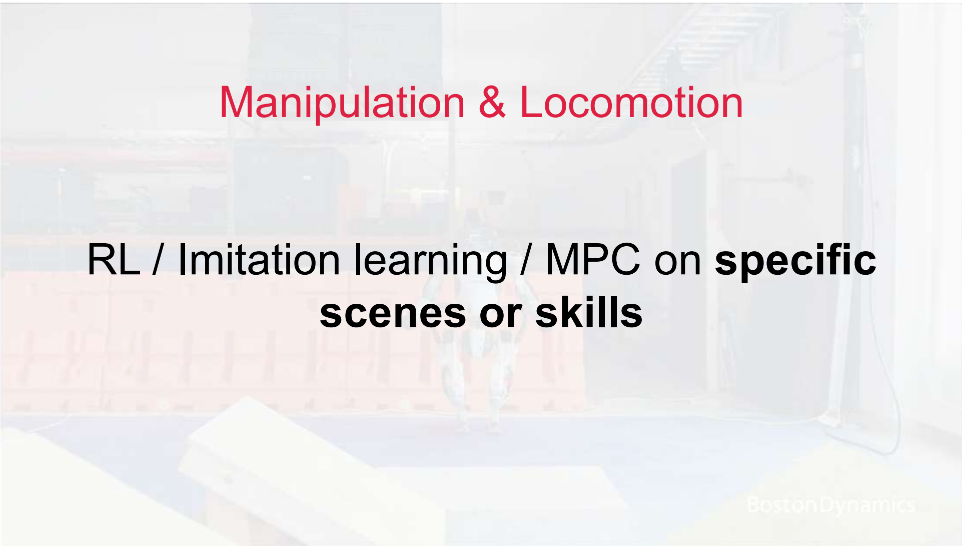
> *It’s comparatively **easy** to make computers exhibit adult level performance on **intelligence tests or playing checkers**, and **difficult or impossible** to give them skills of a one-year old when it comes to **perception and mobility**.*

Hans Moravec,Mind Children, 1988

Embodied AI

> *“The embodiment hypothesis is the idea that* ***intelligence emerges in the interaction of an agent with an environment*** *and as a result of sensorimotor activity”*

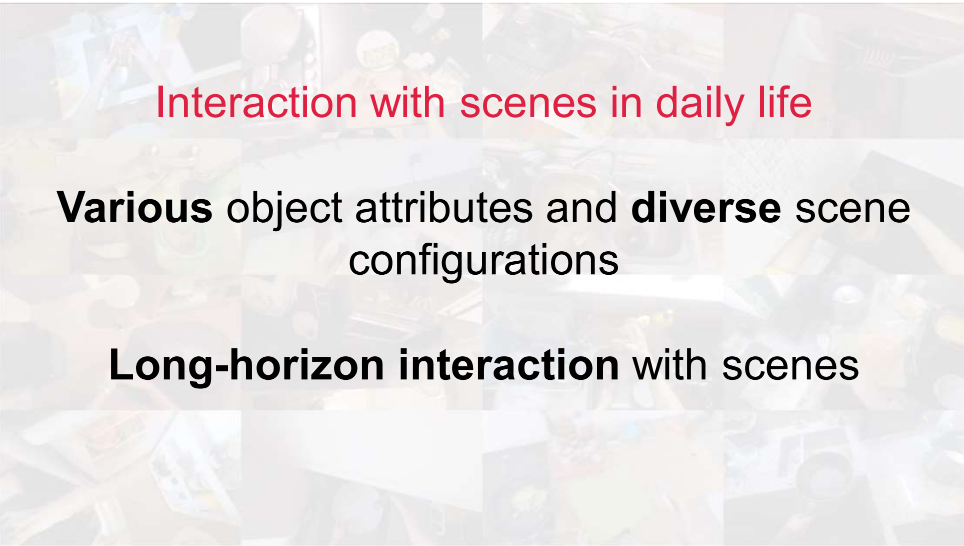
Smith & Gasser, The Development of Embodied Cognition: Six Lessons from Babies, 2005



Manipulation & Locomotion

RL / Imitation learning / MPC on **specific scenes or skills**

Boston Dynamics, Atlas | Partners in Parkour, 2022
https://www.youtube.com/watch?v=tF4DML7FIWk



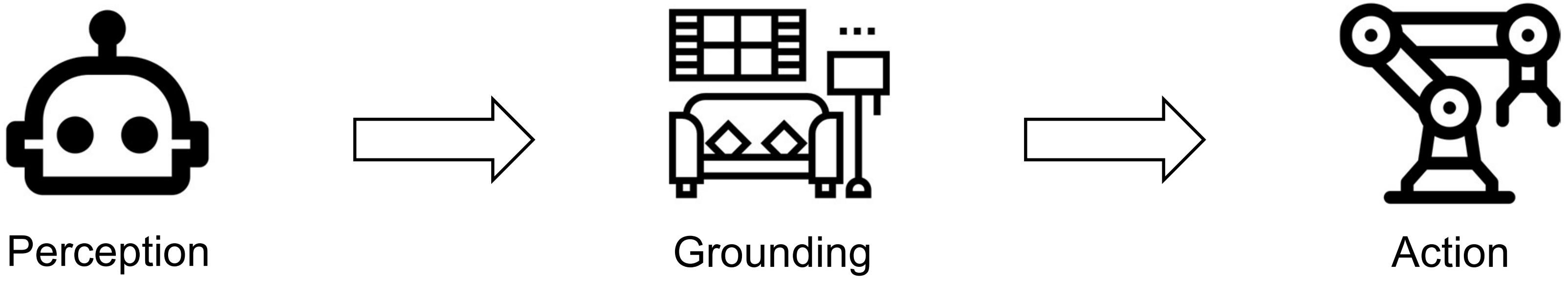
Interaction with scenes in daily life

Various object attributes and **diverse** scene configurations

Long-horizon interaction with scenes

Damen et al., Scaling Egocentric Vision: The Epic-Kitchens Dataset, 2018

From the scene perspective



Perception → Grounding → Action

Perception

- 3D or even 4D data capture
- Representation efficiency
- …

Grounding

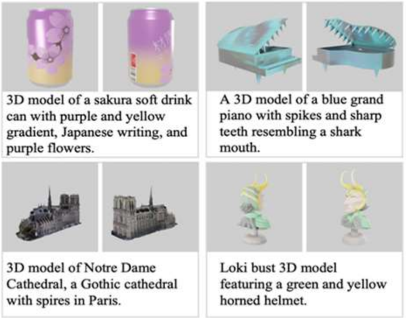
- Spatial relationships in situations
- Affordance & functionality
- …

Action

- Object Geometry / Physics
- Embodiment gap
- …

3D scene understanding for EAI

It is a dark blue couch in the center of this room.
This is a long bar table behind stools.
There is a brown wooden desk in the corner of this room.
This is a bed with blue sheets near the desk.

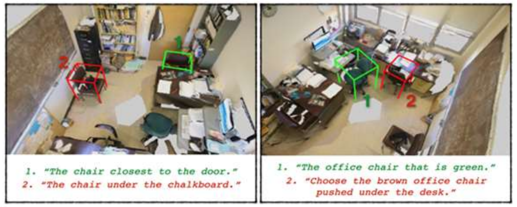


3D model of a sakura soft drink can with purple and yellow gradient, Japanese writing, and purple flowers.

A 3D model of a blue grand piano with spikes and sharp teeth resembling a shark mouth.

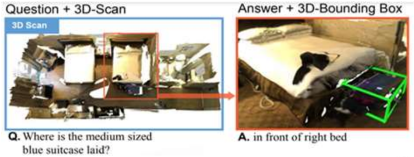
3D model of Notre Dame Cathedral, a Gothic cathedral with spires in Paris.

Loki bust 3D model featuring a green and yellow horned helmet.



1. "The chair closest to the door."
2. "The chair under the chalkboard."

1. "The office chair that is green."
2. "Choose the brown office chair pushed under the desk."

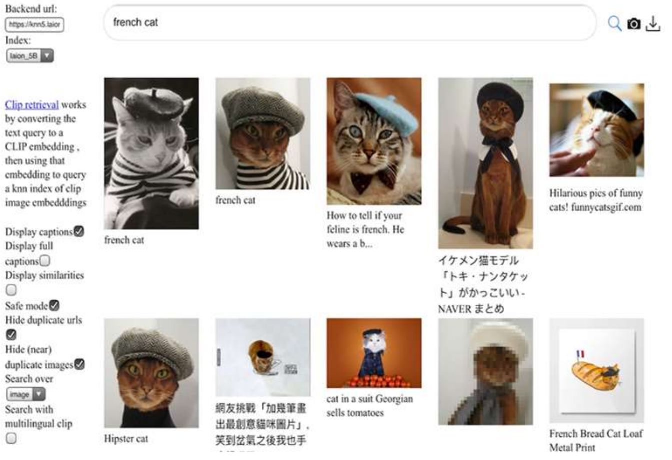


Q. Where is the medium sized blue suitcase laid?

A. in front of right bed

<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)

Hand holding a fresh mangosteen



Backend url: https://knn5.laior

Index: laion_5B

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑
Display full captions ☐
Display similarities ☐
Safe mode ☑
Hide duplicate urls ☑
Hide (near) duplicate images ☑
Search over image
Search with multilingual clip ☐

french cat

french cat

french cat

How to tell if your feline is french. He wears a b...

イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ

Hilarious pics of funny cats! funnycatsgif.com

Hipster cat

網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手

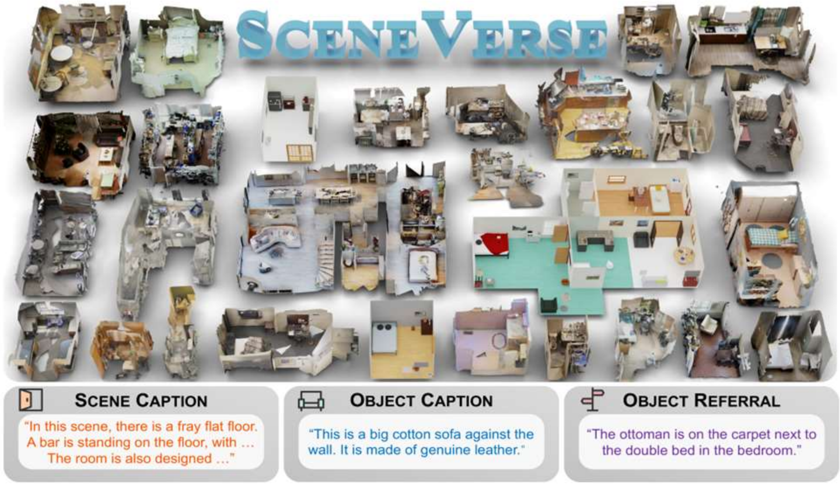
cat in a suit Georgian sells tomatoes

French Bread Cat Loaf Metal Print

Dataset	3D Data		Language		Total
	Scene	Object	Anno.	Syn.	
ScanRefer			52K	-	52K
ReferIt3D	|	|	42K	200K	242K
ScanQA	1.5K	33K	27K	-	27K
SQA3D			-	33K	33K
Multi3DRefer	|	|	52K	10K	62K
Cap3D	-	666K	58K	666K	724K
ScanScribe	3K	56K	94K	184K	278K

Dataset	2D Image-text pairs
MS-COCO	330K
Visual Genome	5.4M
WIT	5.5M
Conceptual Captions-12M	12M
YFCC100M	100M
LAION-5B	2.3B

Scaling 3D-VL with SceneVerse



SCENE CAPTION

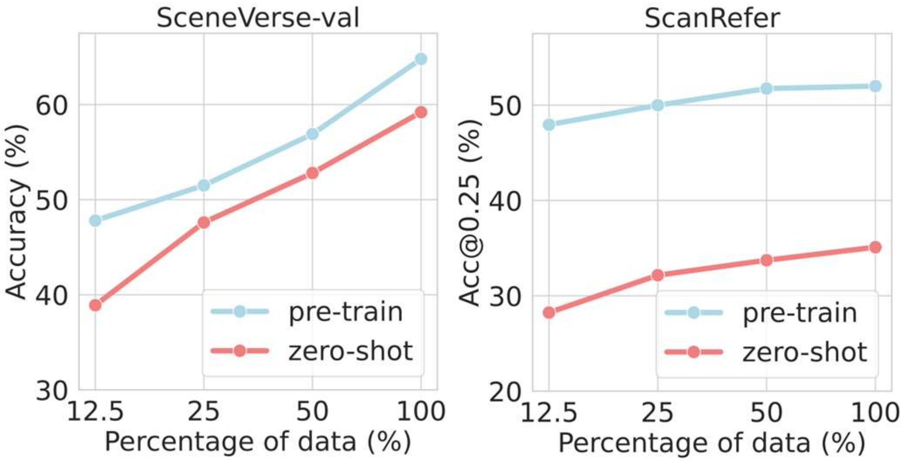
"In this scene, there is a fray flat floor. A bar is standing on the floor, with … The room is also designed …"

OBJECT CAPTION

"This is a big cotton sofa against the wall. It is made of genuine leather."

OBJECT REFERRAL

"The ottoman is on the carpet next to the double bed in the bedroom."



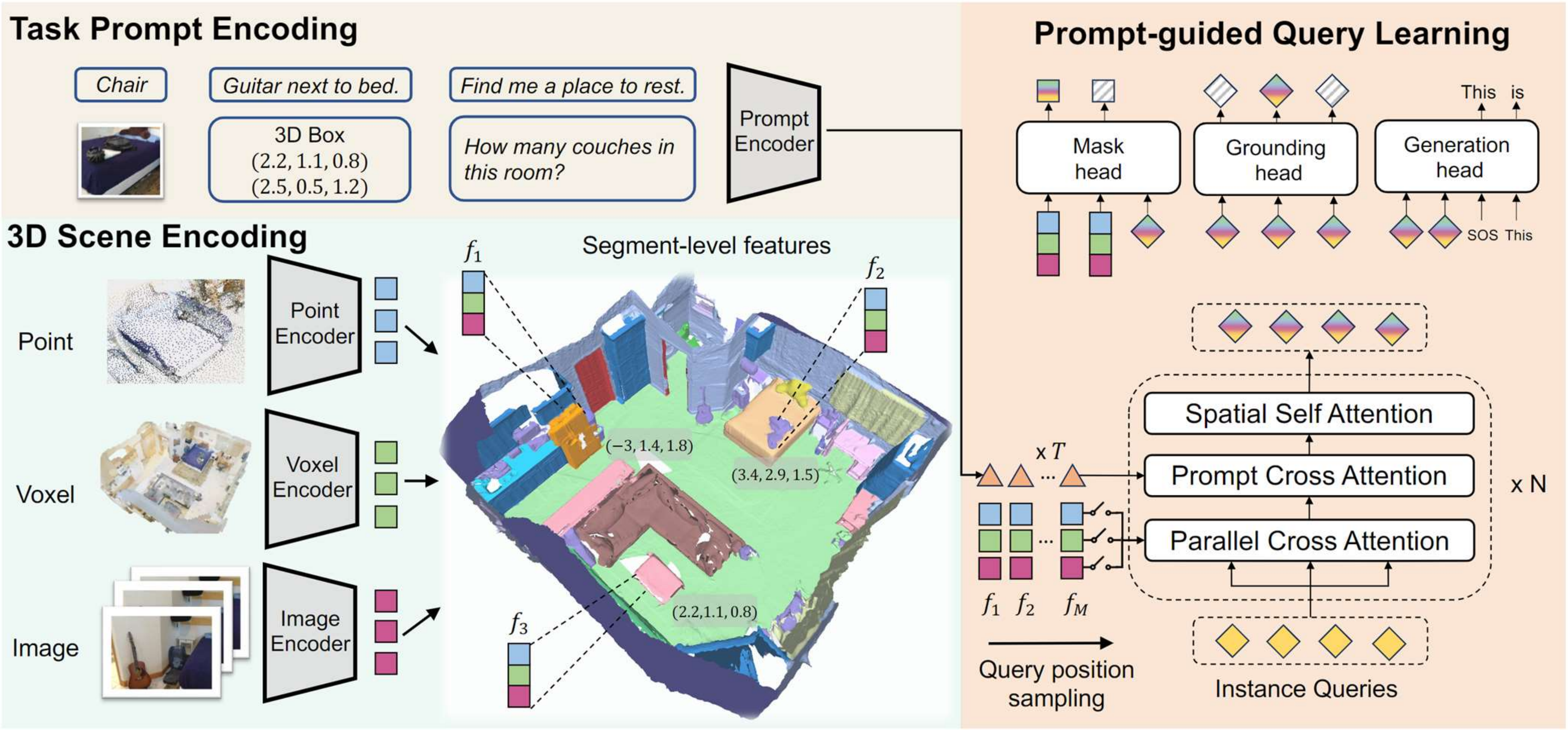
Scene Captioning

Scene Captioning

Scene Captioning

Jia et al., SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding, ECCV 2024

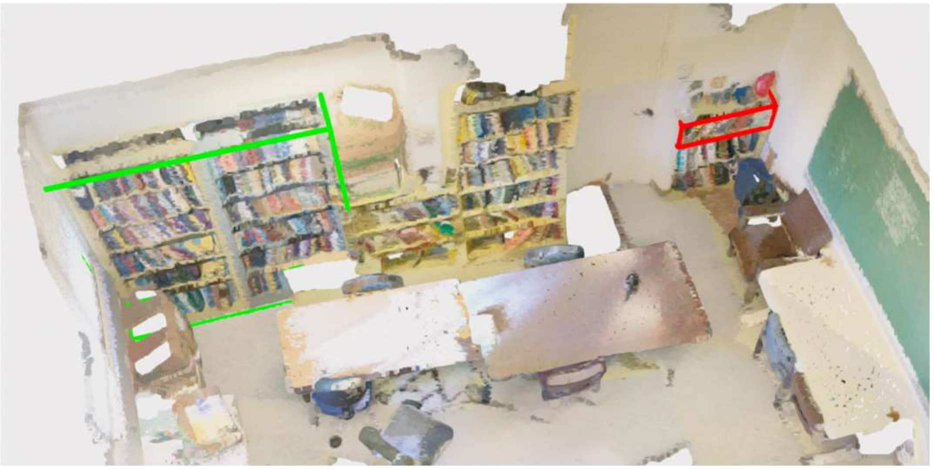
Uniting scene representations



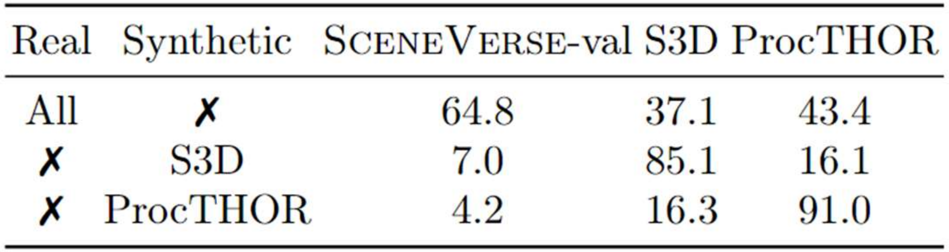
Zhu et al., Unifying 3D Vision-Language Learning via Promptable Queries, ECCV 2024

Findings & Takeaways

- Language
  - Relatively easy to scale
  - Quality of language matters
- Scene
  - Imbalanced classes
  - Domain gap between synthetic and real data
  - Domain gap between real-world datasets

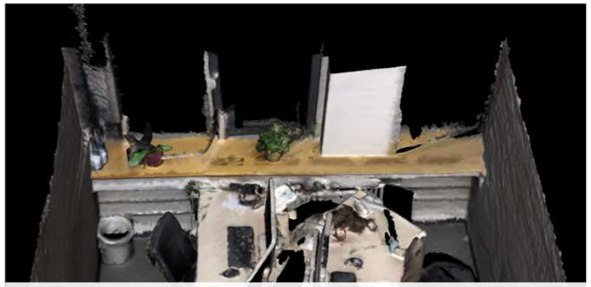


“In the corner of the room are boxes. the first two book shelves in the corner to the right of the boxes are the bookshelves we are looking for.”



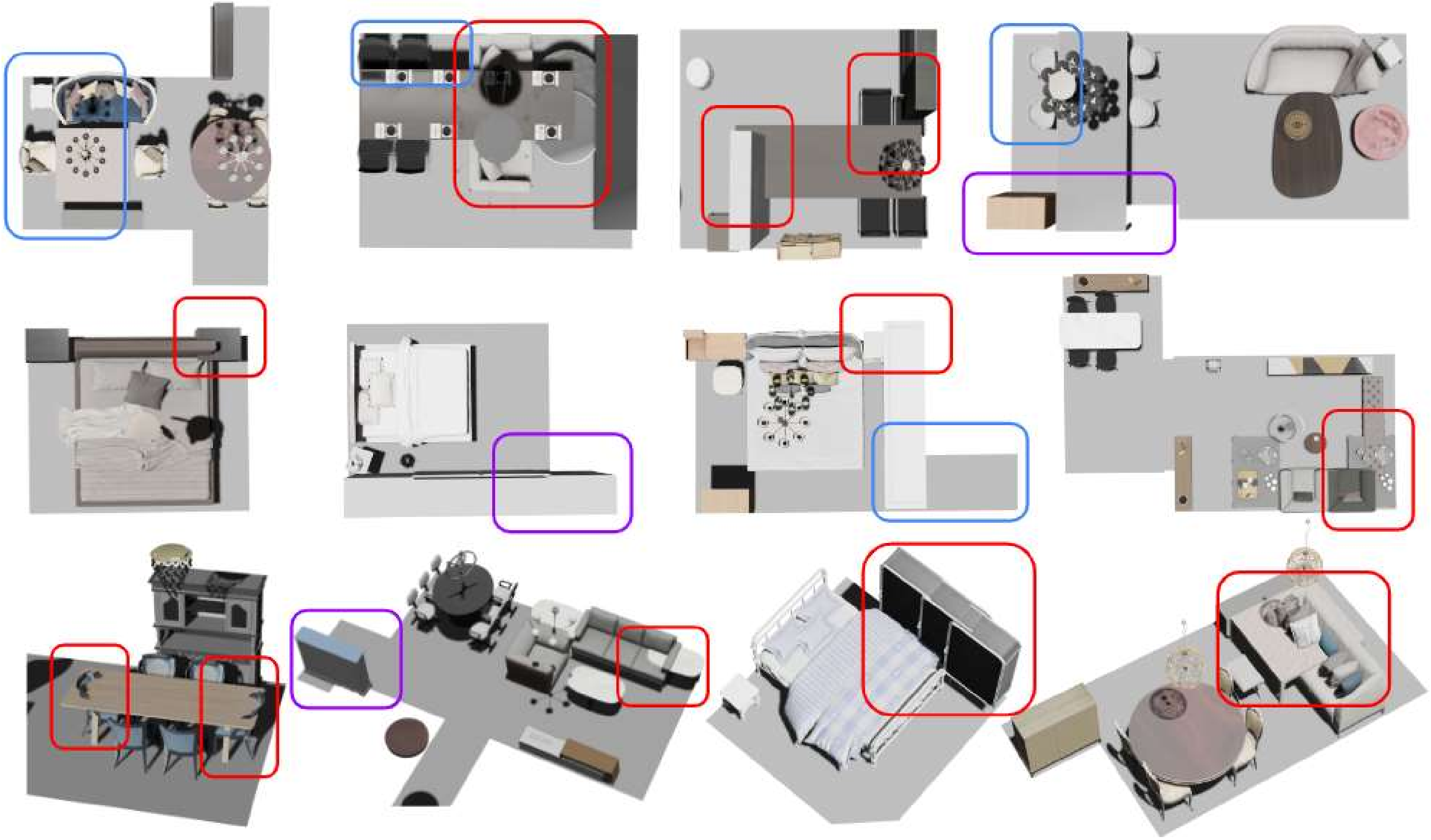
Real	Synthetic	SceneVerse-val	S3D	ProcTHOR
All	✗	64.8	37.1	43.4
✗	S3D	7.0	85.1	16.1
✗	ProcTHOR	4.2	16.3	91.0

Lamp



Plant

Can scene generation or reconstruction help?

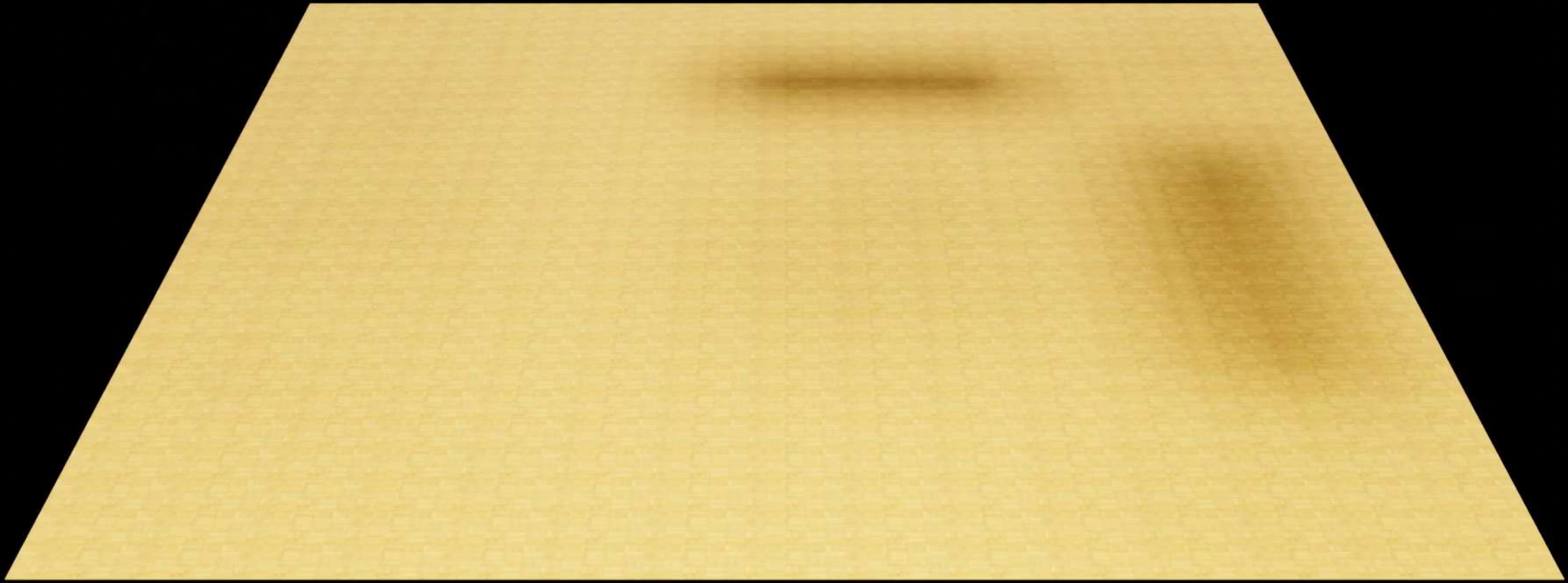


🟥 Collisions between objects

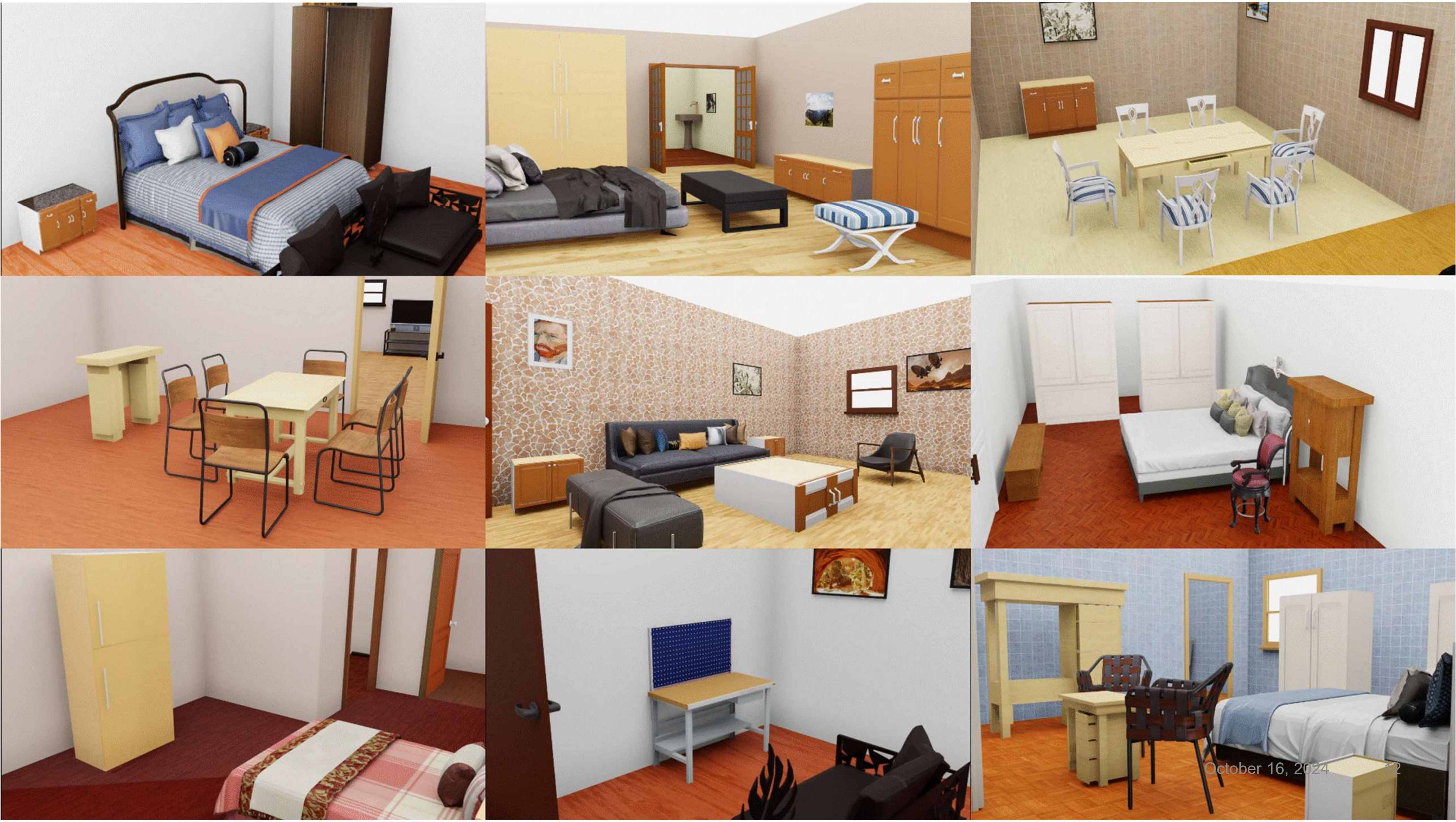
🟪 Objects outside of floor plan

🟦 Areas unreachable to agents

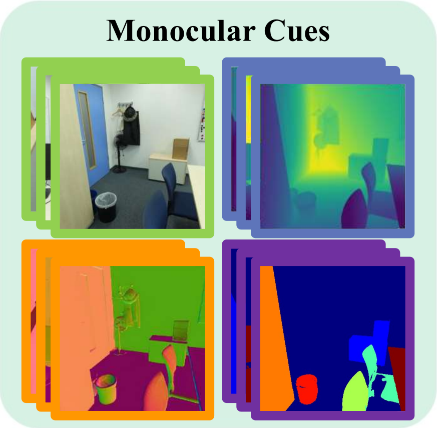
Fu et al., 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics, ICCV 2021

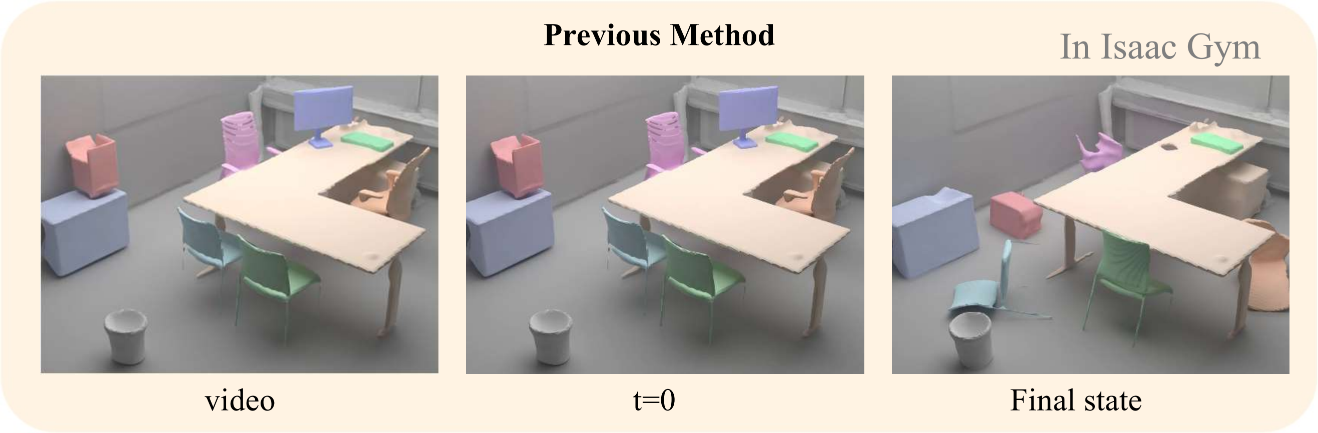


Yang et al., PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI, CVPR 2024 (Highlight)



Reconstruction?





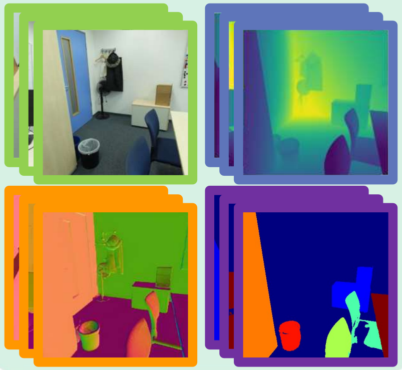
video

t=0

Final state

Reconstructing scenes with physical constraints

Monocular Cues

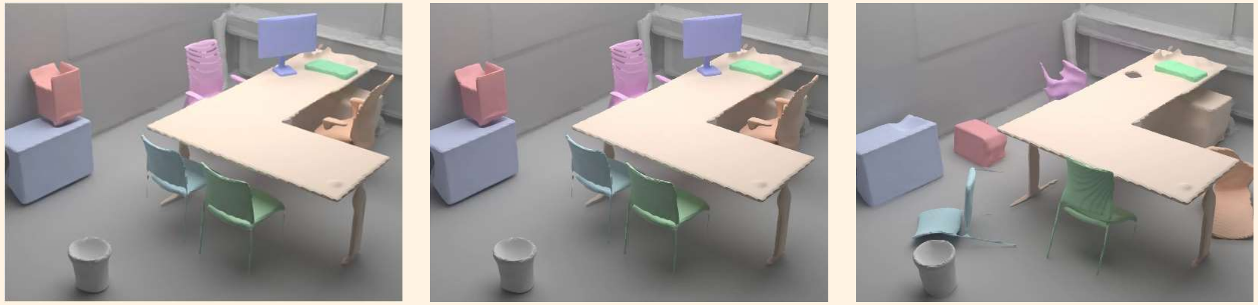


Physical Simulator

t=0 … Final state

Previous Method

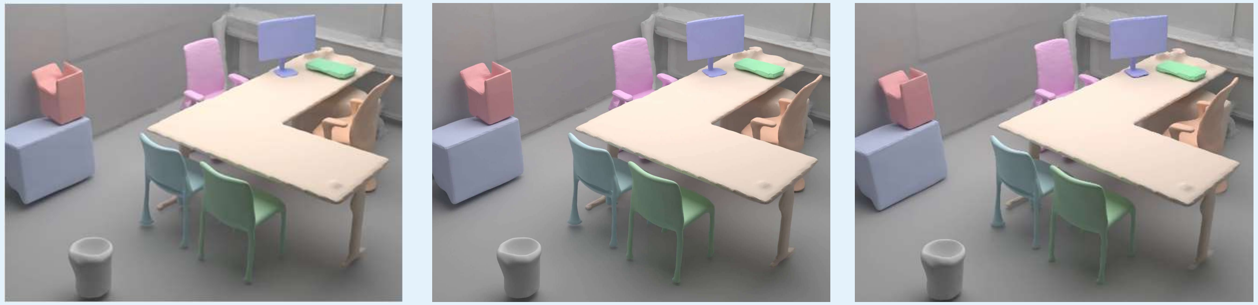
In Isaac Gym



video t=0 Final state

PHYRECON

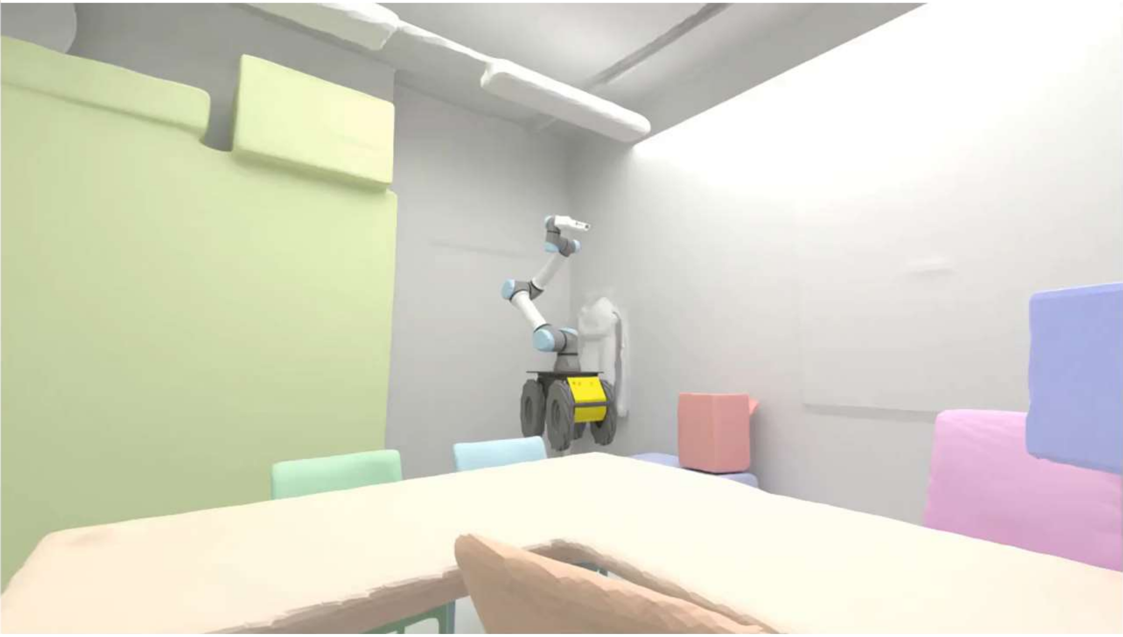
In Isaac Gym



video t=0 Final state

Ni et al., PhyRecon: Physically Plausible Neural Scene Reconstruction, NeurIPS 2024

- Generation
  - Insufficient training scale & diversity
  - Limited diversity both assets and layout
- Reconstruction
  - No object articulation, pick & place only
  - Limited efficiency and scaling potential



- Articulated asset reconstruction
- Retrieval augmented reconstruction

Afford-motion, CVPR 2024 Highlight

The man walks to the chair in a curve.

TRUMANS, CVPR 2024 Highlight

LingoMotions, SIGGRAPH ASIA 2024

Qualitative results of our synthesized motions

Text Instruction

Go to the sink to wash the hands

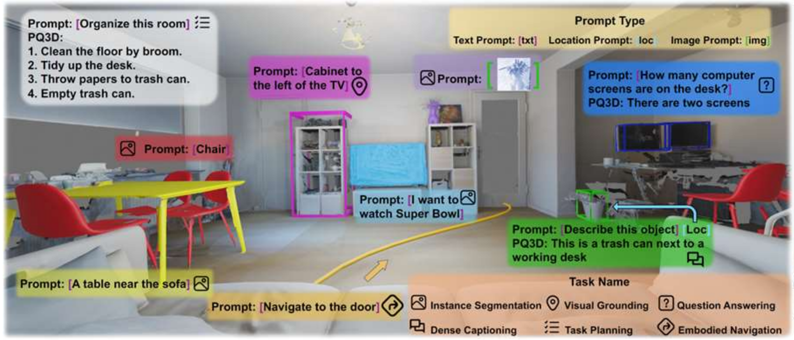
I am hungry. Could you give me some food? And pass me a cup of juice.

COME-Robot, ArXiv 2024

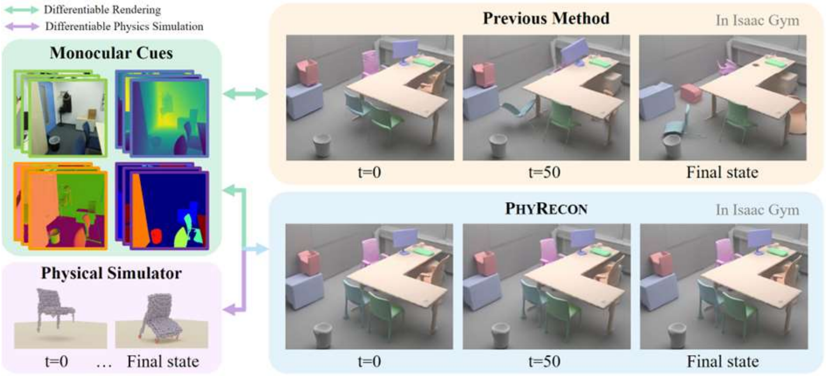
Above all

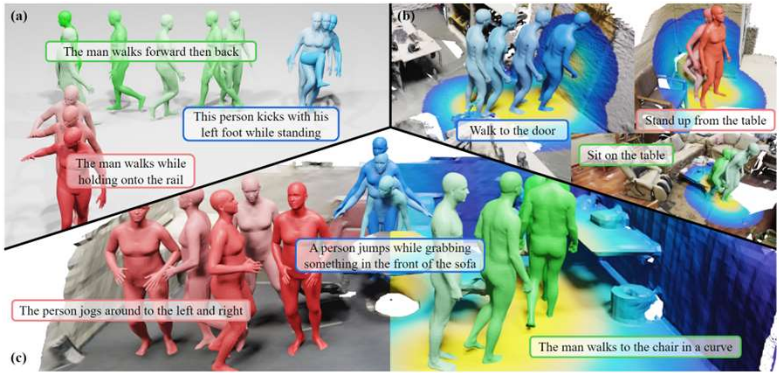
- Scaling works in 3D scene understanding
  - Unifying different domains and situation modeling
- Scene curation for embodied AI is still challenging
  - Ensuring naturalness and physics while maintaining diversity
- Interaction data from human-scene interaction, egocentric videos
  - Robust robot system for coordinating different modules for real-world applications
  - Transfer from motion to robots, agent agnostic policy learning

More to come at BIGAI


Prompt: [Organize this room]
PQ3D:
1. Clean the floor by broom.
2. Tidy up the desk.
3. Throw papers to trash can.
4. Empty trash can.
Prompt: [Cabinet to the left of the TV]
Prompt:
Prompt Type
Text Prompt: [txt] Location Prompt: loc Image Prompt: [img]
Prompt: [How many computer screens are on the desk?]
PQ3D: There are two screens
Prompt: [Chair]
Prompt: [I want to watch Super Bowl]
Prompt: [Describe this object] Loc
PQ3D: This is a trash can next to a working desk
Prompt: [A table near the sofa]
Prompt: [Navigate to the door]
Task Name
Instance Segmentation
Visual Grounding
Question Answering
Dense Captioning
Task Planning
Embodied Navigation

Physically Interactable 3D Scene
Walkable Map
Articulated Objects
Articulated Objects




(a)
The man walks forward then back
This person kicks with his left foot while standing
The man walks while holding onto the rail
(b)
Walk to the door
Stand up from the table
Sit on the table
A person jumps while grabbing something in the front of the sofa
The person jogs around to the left and right
The man walks to the chair in a curve
(c)

Thank you!