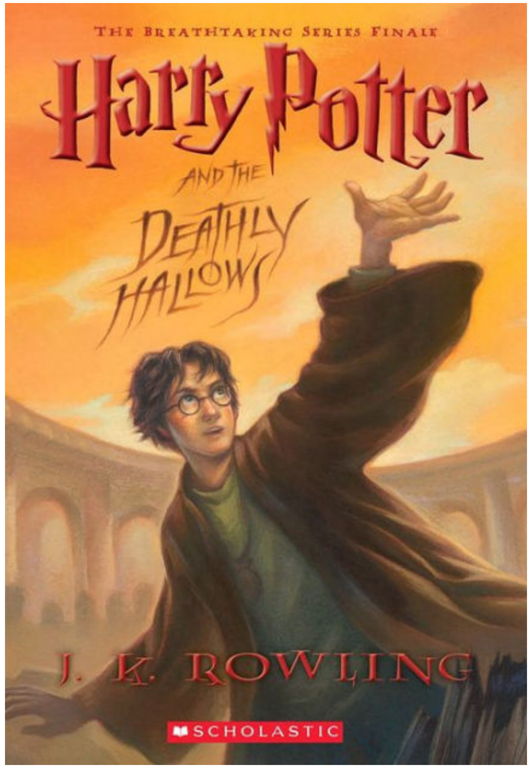
Grounding natural language to 3D

Angel Xuan Chang

2020-07-09

ALVR Workshop at ACL 2020

Wouldn't it be great?



WordsEye (Coyne and Sproat SIGGRAPH 2001)

An Automatic Text-to-Scene Conversion System

….the desk is against the back wall. the chair is in front of the desk. it is facing north. the computer is on the desk. a lamp is one foot to the left of the desk. a small pink trashcan is two feet to the right of the desk. a red stapler is one foot to the right of the computer.

wordseye™

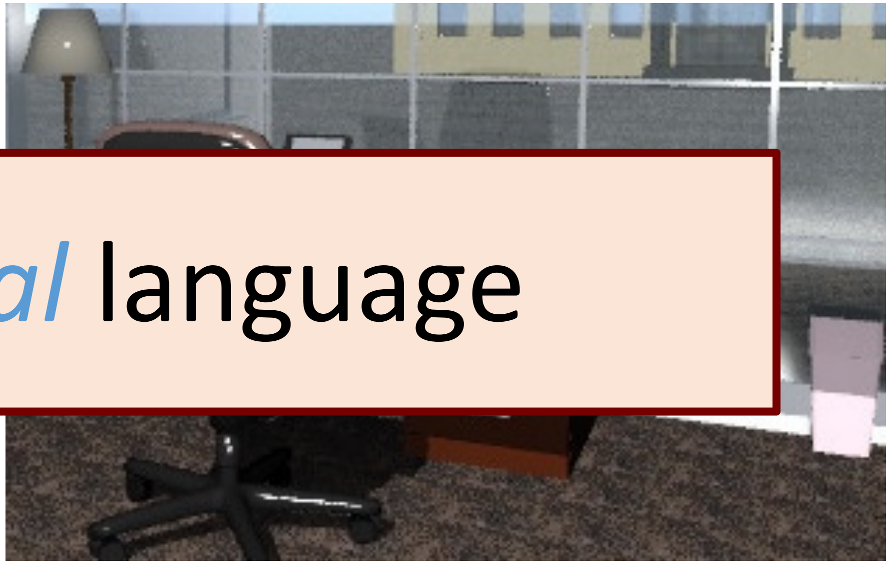
type a picture

https://www.wordseye.com/

WordsEye (Coyne and Sproat SIGGRAPH 2001)

An Automatic Text-to-Scene Conversion System

….the desk is against the back wall. the chair is in front of the desk. it is facing north.
the com
lamp is c
the desk
is two feet to the right of the desk. a red stapler is one foot to the right of the computer.



NOT *natural* language

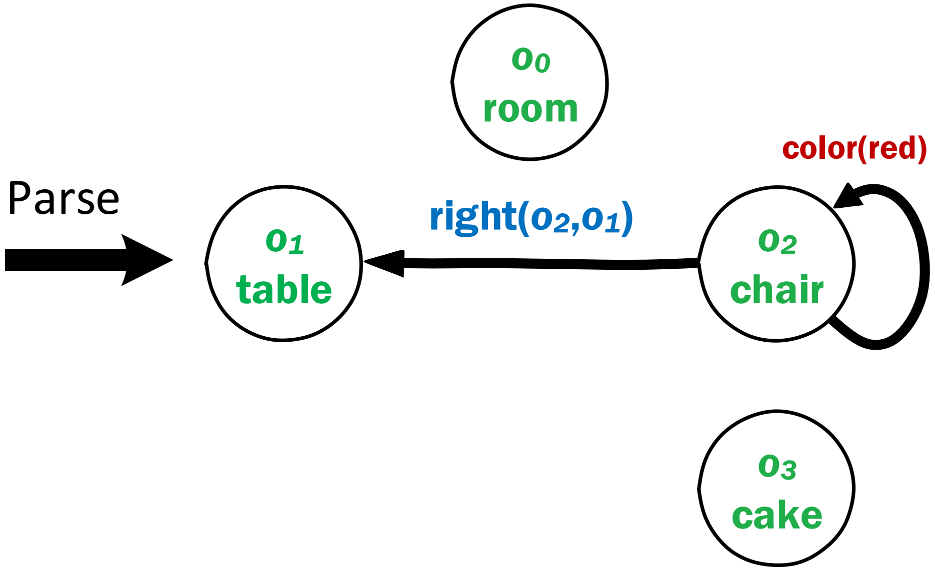
How do we handle natural, underspecified language?

"Living room with a red couch"

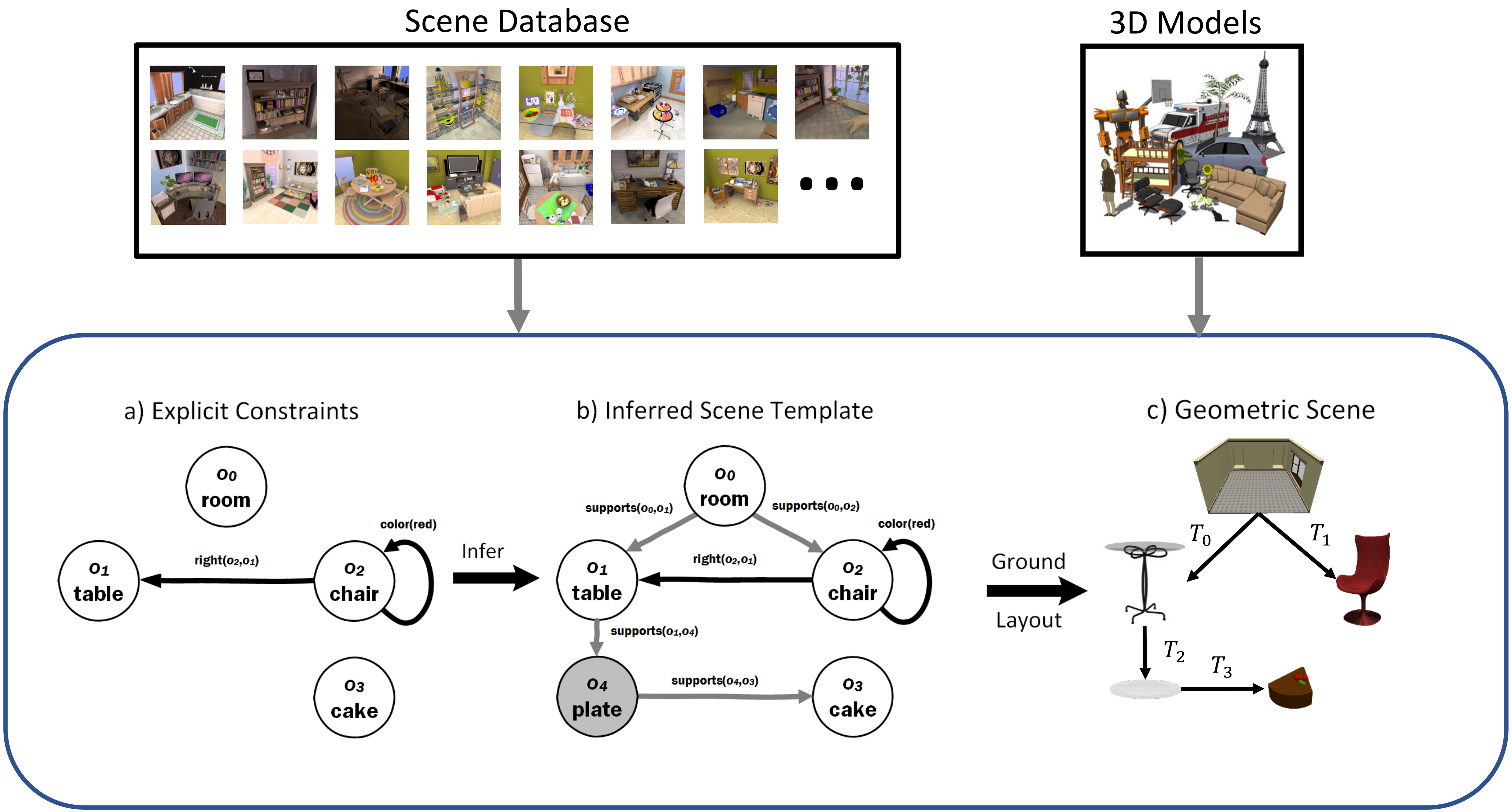
- learn common sense priors on how objects are arranged in the real world
- view scene description as constraints on the scene

Language as constraint for 3D scene graphs

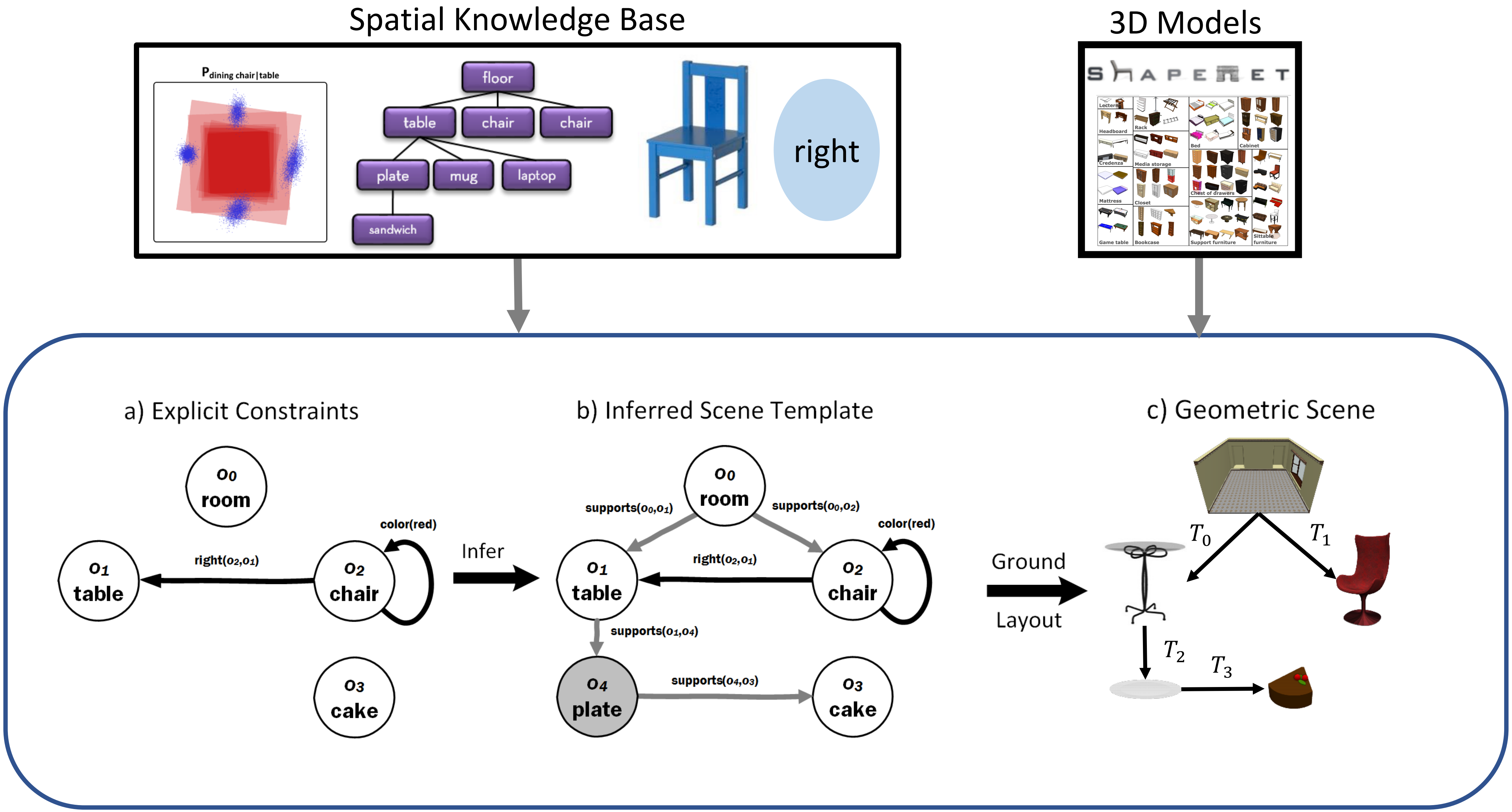
"There is a room with a table and a cake. There is a red chair to the right of the table."



objects, attributes and relations



Scene Database
3D Models
a) Explicit Constraints
b) Inferred Scene Template
c) Geometric Scene
o_0 room
o_1 table
o_2 chair
o_3 cake
color(red)
right(o_2,o_1)
Infer
o_4 plate
supports(o_0,o_1)
supports(o_0,o_2)
supports(o_1,o_4)
supports(o_4,o_3)
Ground
Layout
T_0
T_1
T_2
T_3



Spatial Knowledge Base
3D Models
$P_{\text{dining chair}|\text{table}}$
floor
table
chair
chair
plate
mug
laptop
sandwich
right
a) Explicit Constraints
b) Inferred Scene Template
c) Geometric Scene
o0 room
o1 table
o2 chair
o3 cake
right(o2,o1)
color(red)
Infer
supports(o0,o1)
supports(o0,o2)
supports(o1,o4)
supports(o4,o3)
o4 plate
Ground
Layout
T_0
T_1
T_2
T_3

What is some spatial “common sense” that we have captured?

Scene database

Stanford Scene Database

133 scenes using
2455 models

3 objects

Average of
26 objects

103 objects

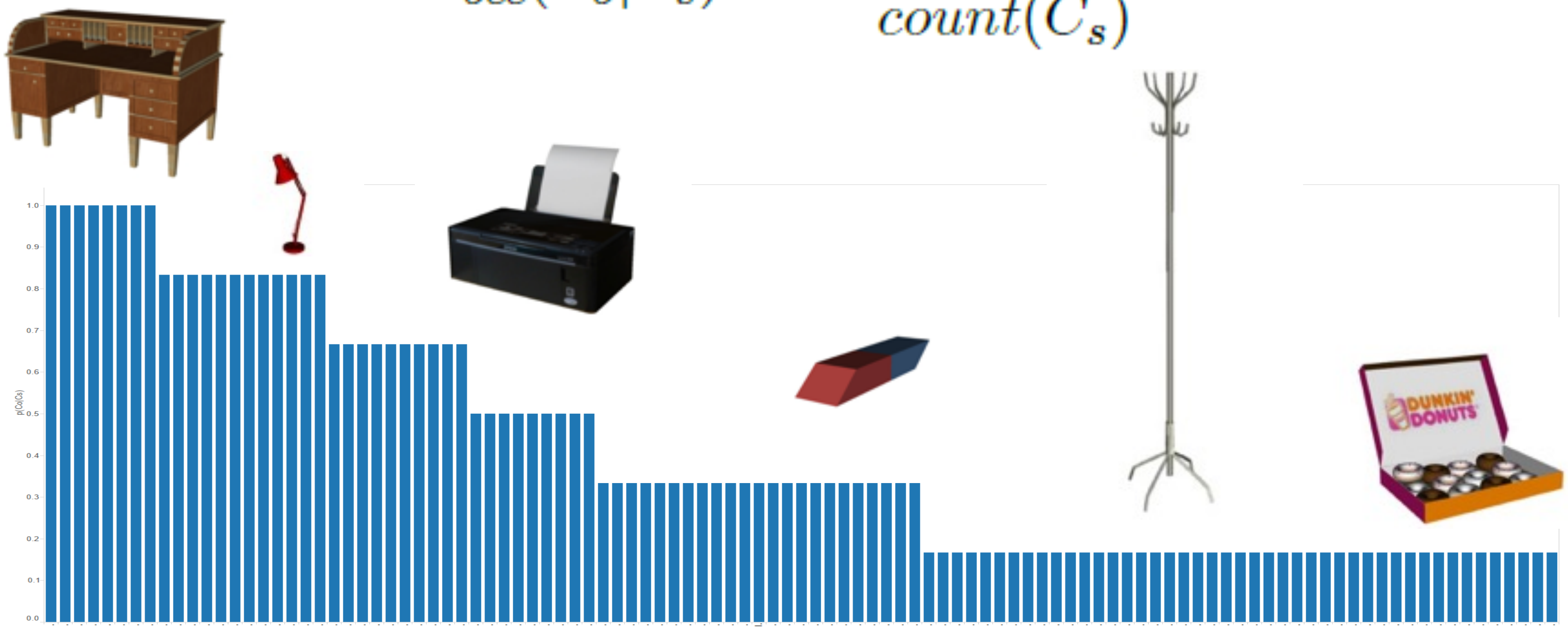
Example-based Synthesis of 3D Object Arrangements
Fisher et al, SIGGRAPH Asia 2012

Object occurrences

What goes in an office?

Probability that object of category C_o is found in scene type C_s

$$P_{occ}(C_o|C_s) = \frac{count(C_o \text{ in } C_s)}{count(C_s)}$$

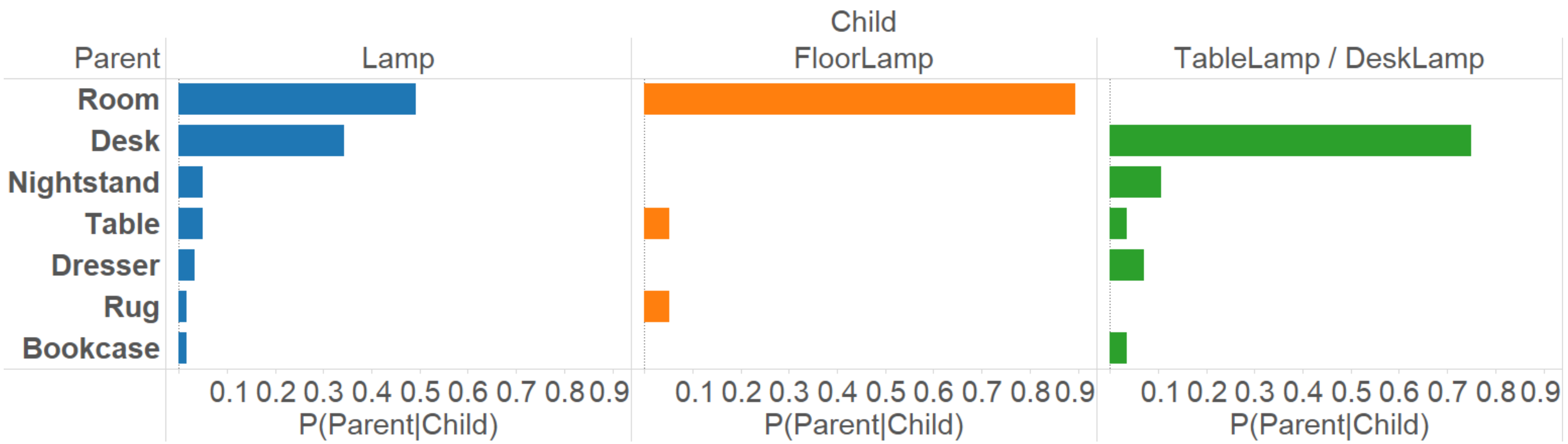


Support hierarchy

What goes on top of what?

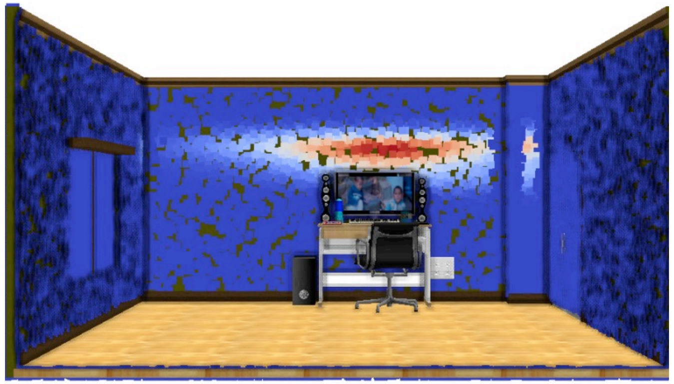
Probability that parent category C_p supports child category C_c

$$P_{support}(C_p|C_c) = \frac{count(C_c \text{ on } C_p)}{count(C_c)}$$



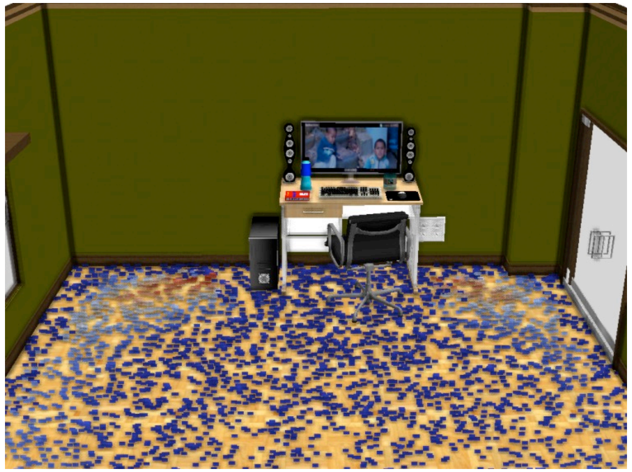
Semantic queries – Where can X go?

poster

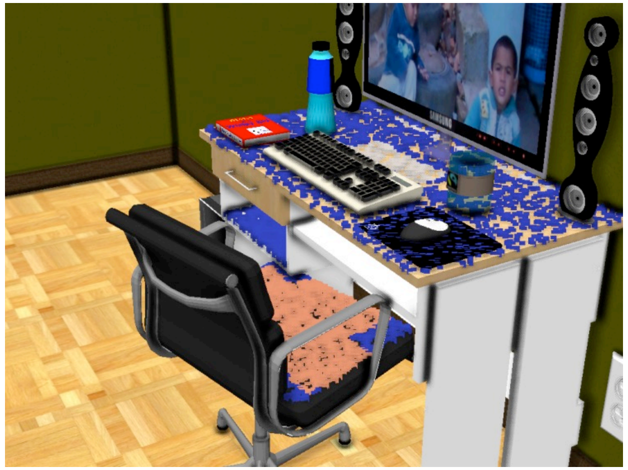


rug

floor lamp

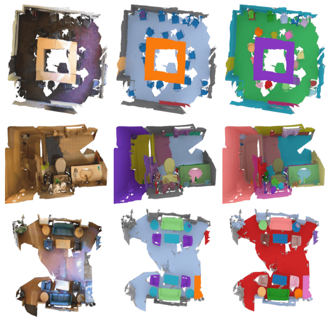


hat



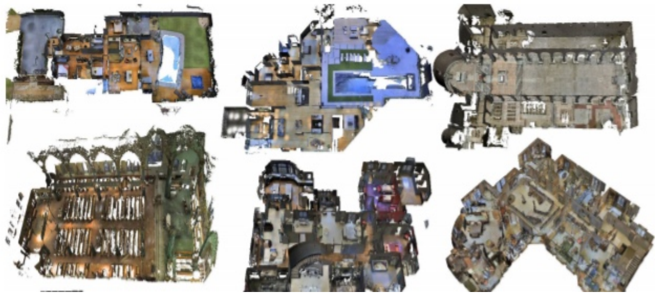
Datasets for semantic understanding in 3D

3D scenes



ScanNet

[Dai et al. 2017]



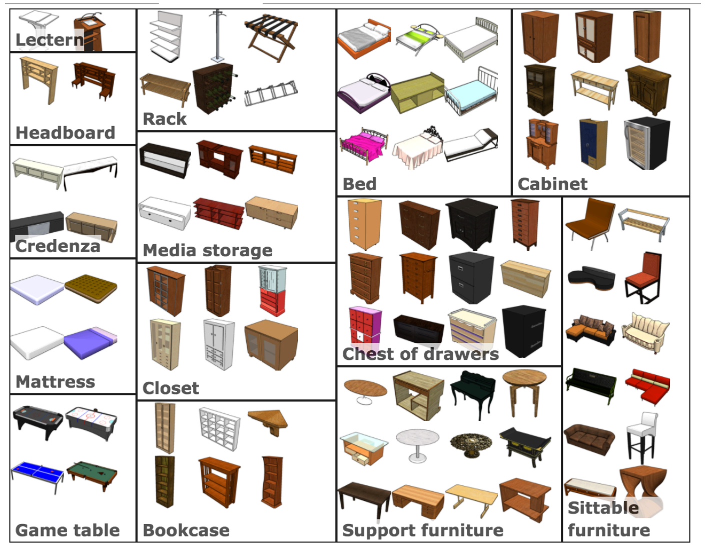
Matterport3D

[Chang et al. 2017]

PRINCETON UNIVERSITY

TECHNISCHE UNIVERSITÄT MÜNCHEN

3D shapes

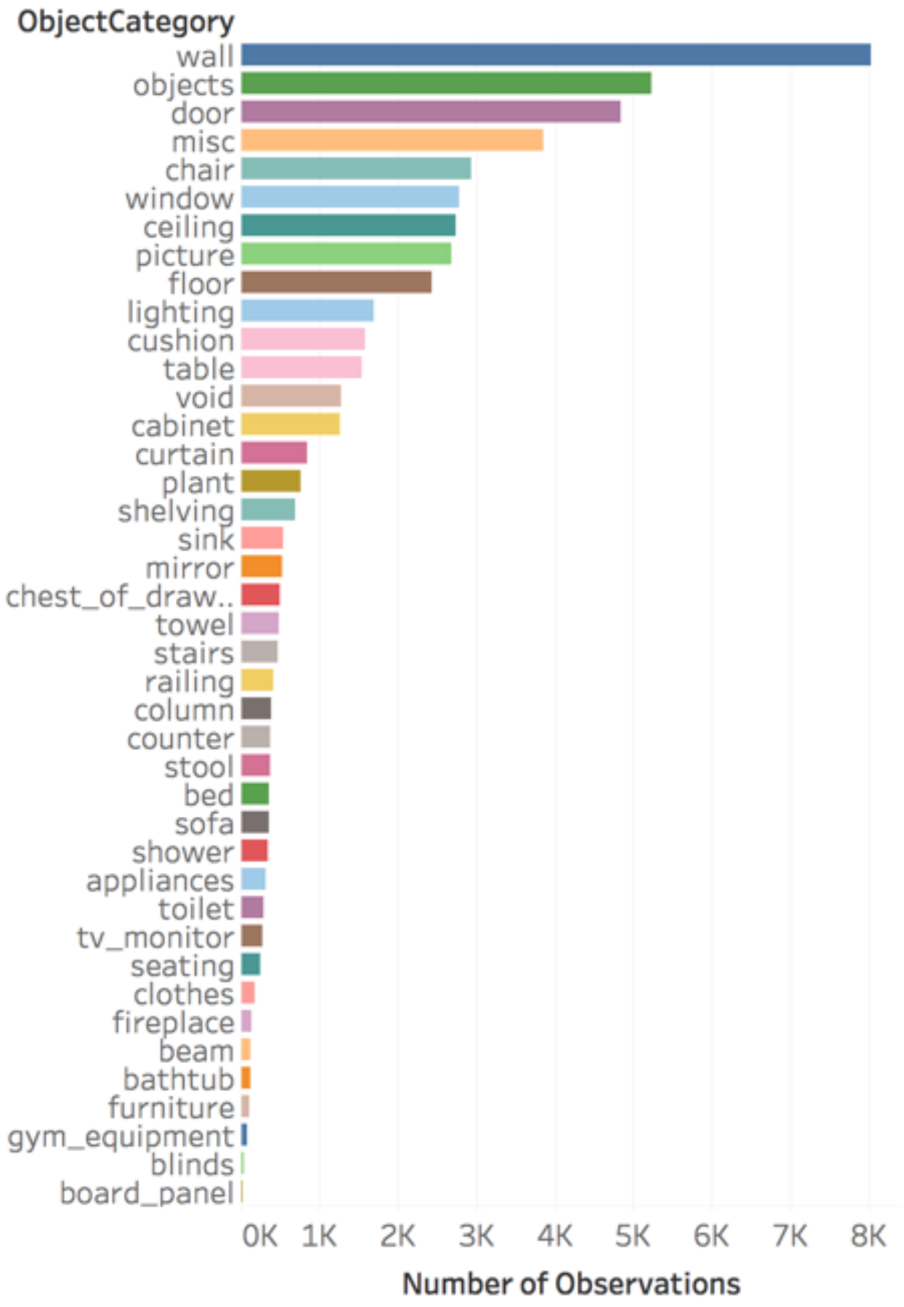


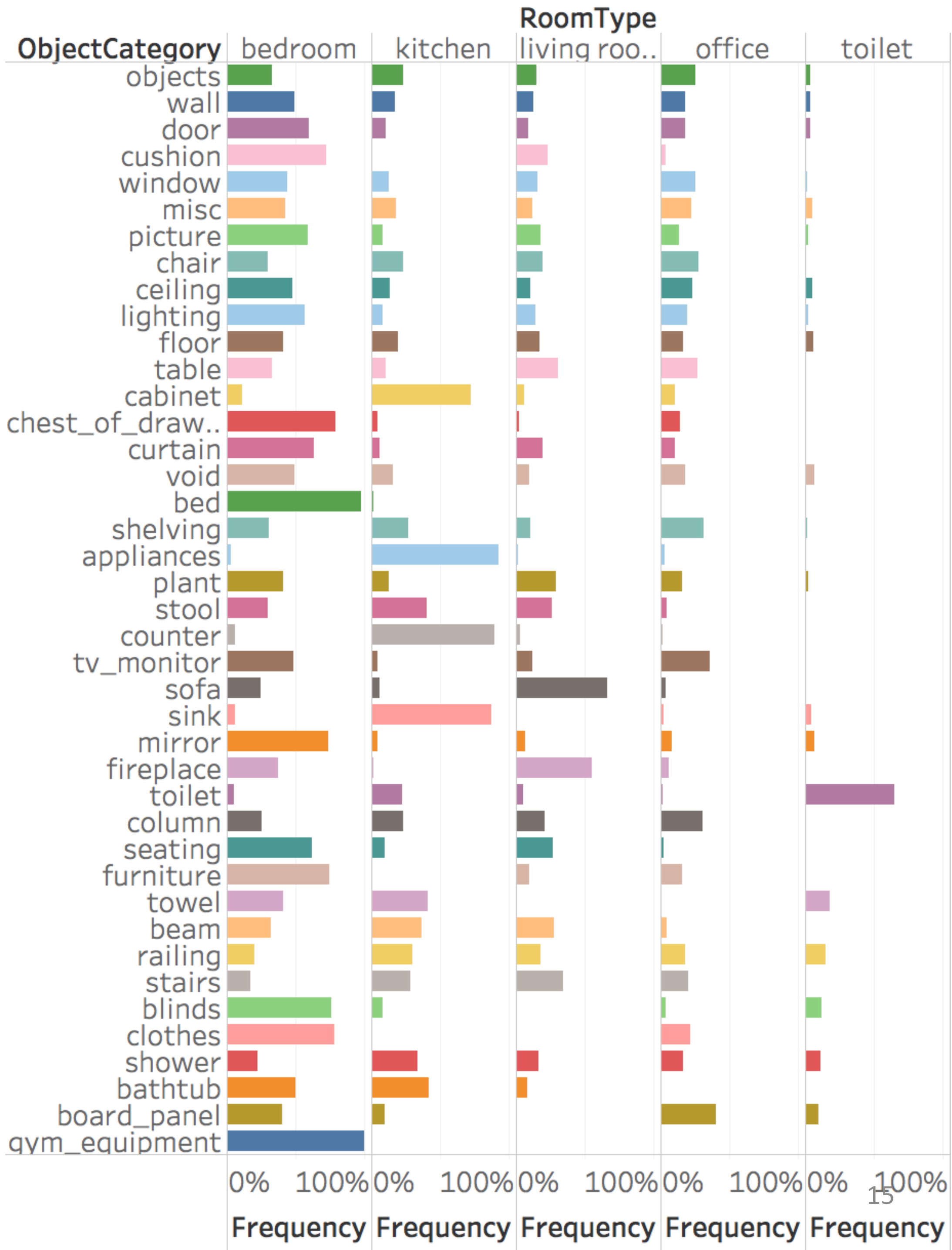
ShapeNet

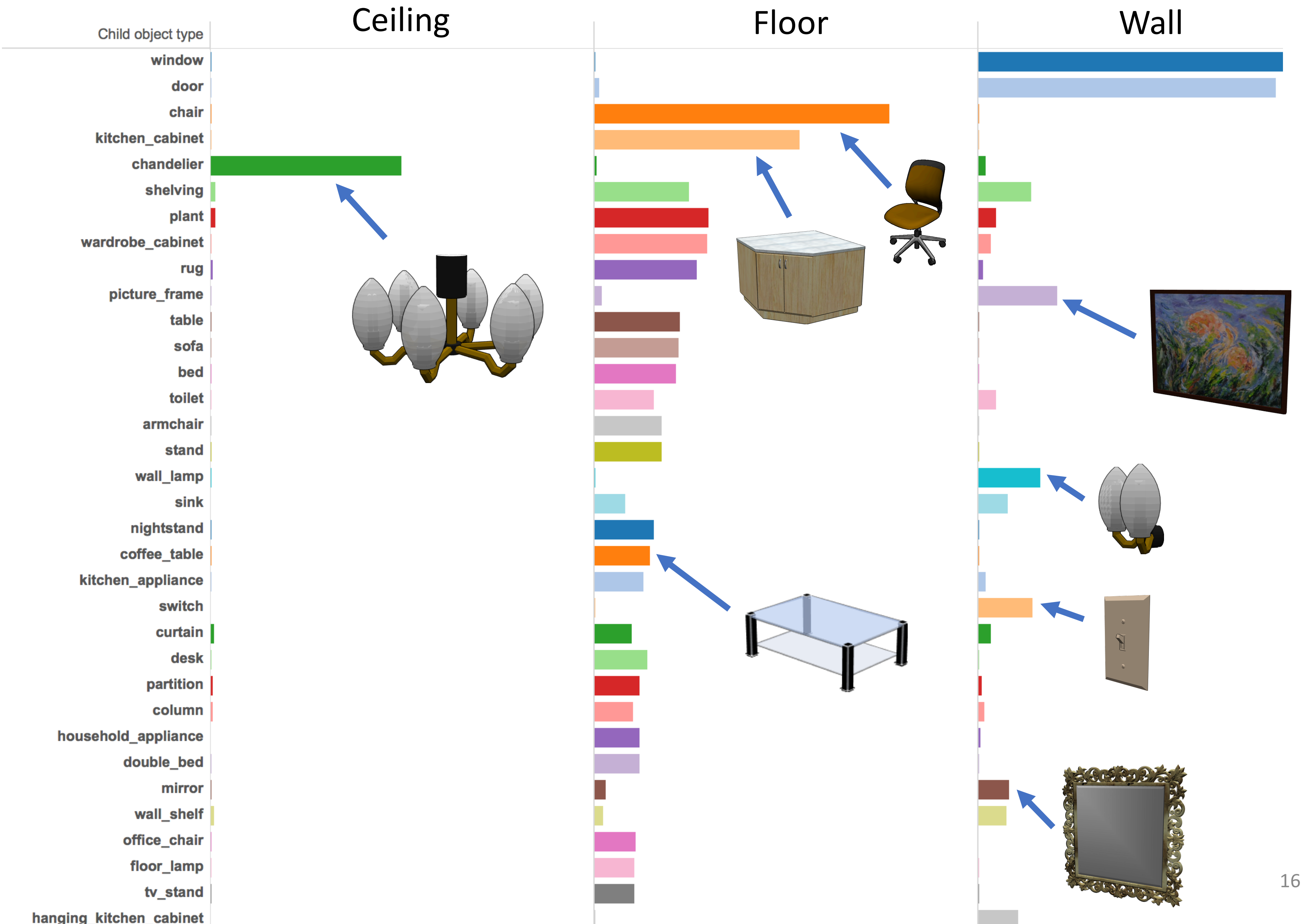
[Chang et al. 2015]

Stanford University

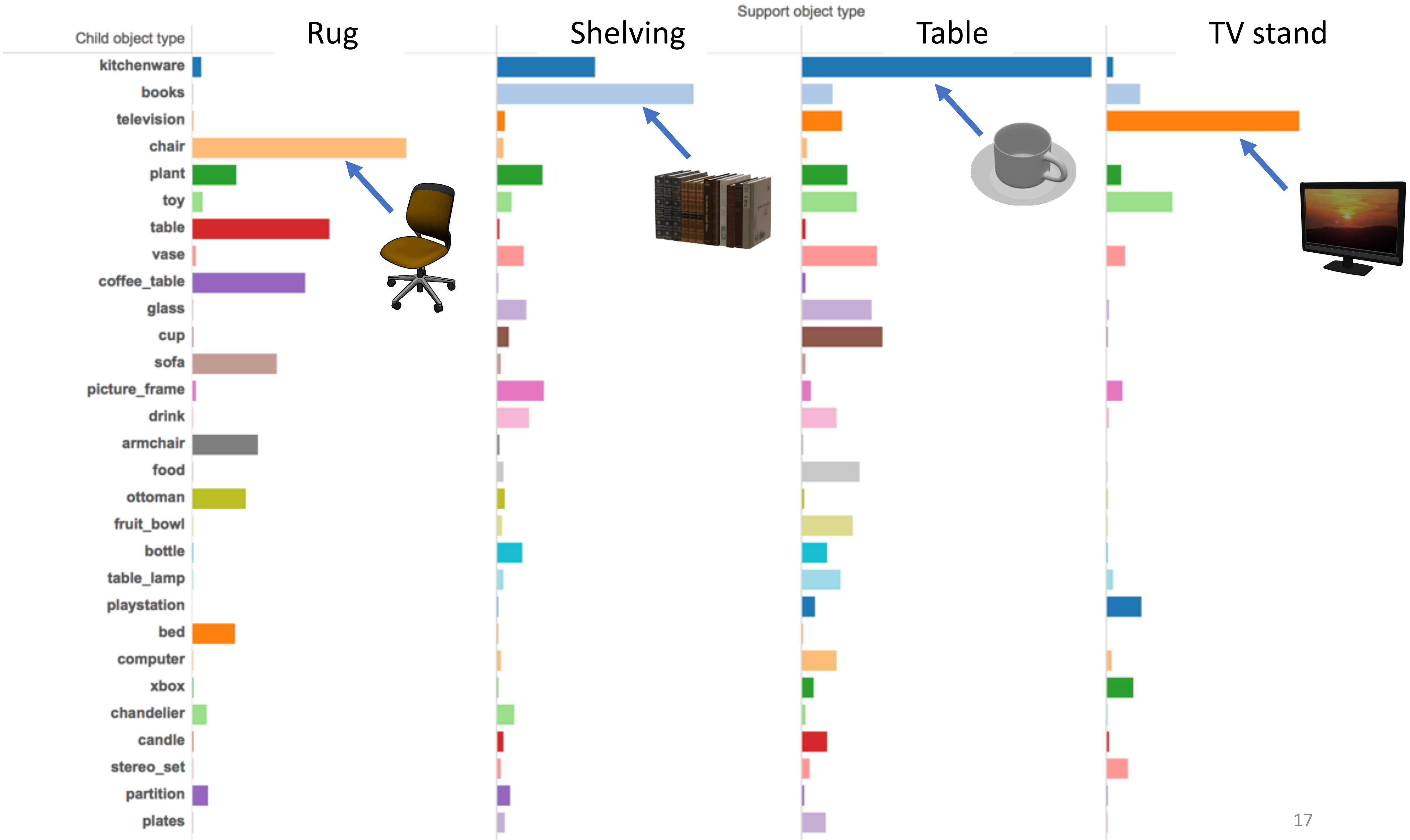
Matterport3D object statistics



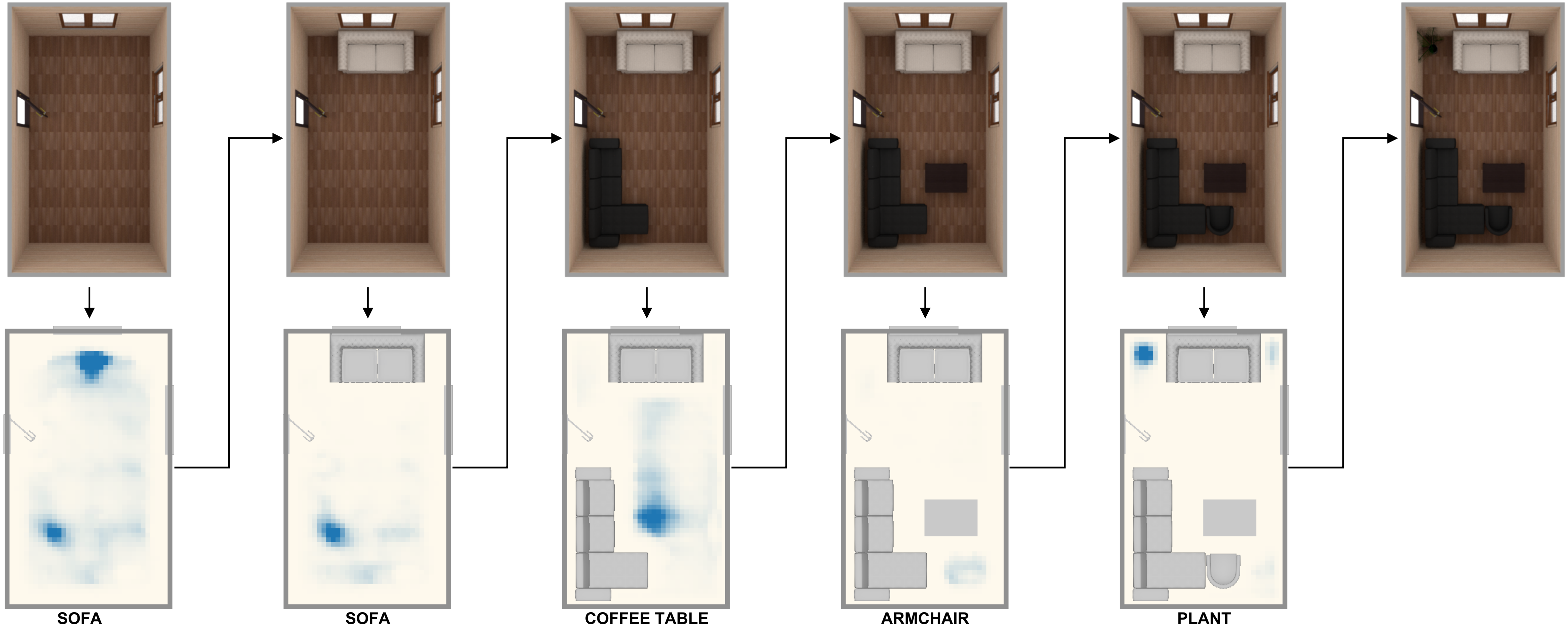




Ceiling
Floor
Wall
Child object type
window
door
chair
kitchen_cabinet
chandelier
shelving
plant
wardrobe_cabinet
rug
picture_frame
table
sofa
bed
toilet
armchair
stand
wall_lamp
sink
nightstand
coffee_table
kitchen_appliance
switch
curtain
desk
partition
column
household_appliance
double_bed
mirror
wall_shelf
office_chair
floor_lamp
tv_stand
hanging_kitchen_cabinet


Support object type
Child object type
Rug
Shelving
Table
TV stand
kitchenware
books
television
chair
plant
toy
table
vase
coffee_table
glass
cup
sofa
picture_frame
drink
armchair
food
ottoman
fruit_bowl
bottle
table_lamp
playstation
bed
computer
xbox
chandelier
candle
stereo_set
partition
plates

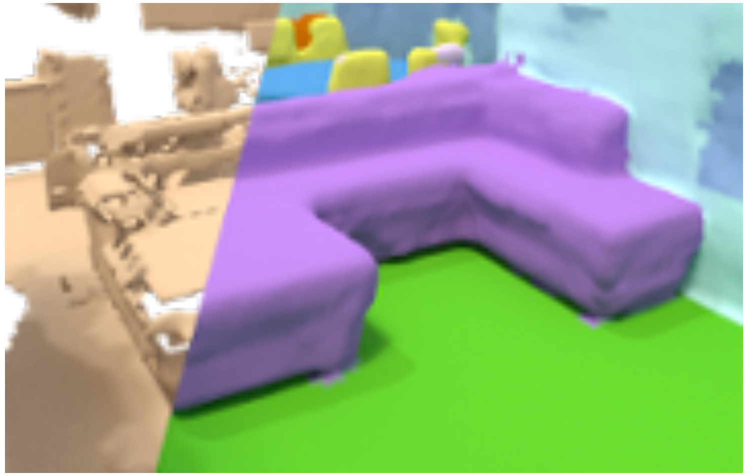
What goes in a living room and where?



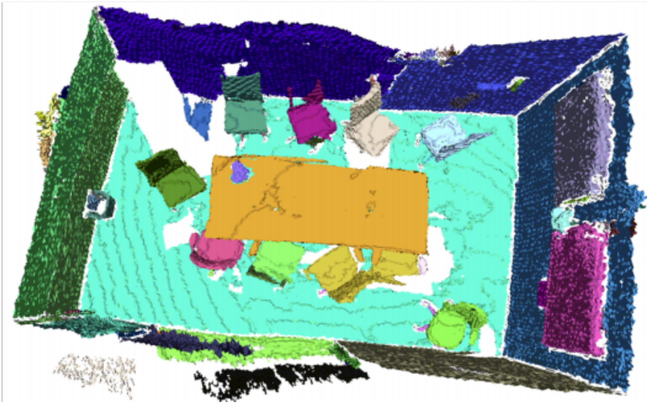
Deep Convolutional Priors for Scene Synthesis [Wang et al, 2018]

Progress in 3D deep learning

4D Spatio-Temporal ConvNets
[Choy et al. 2019]



ScanComplete
[Dai et al. 2018]



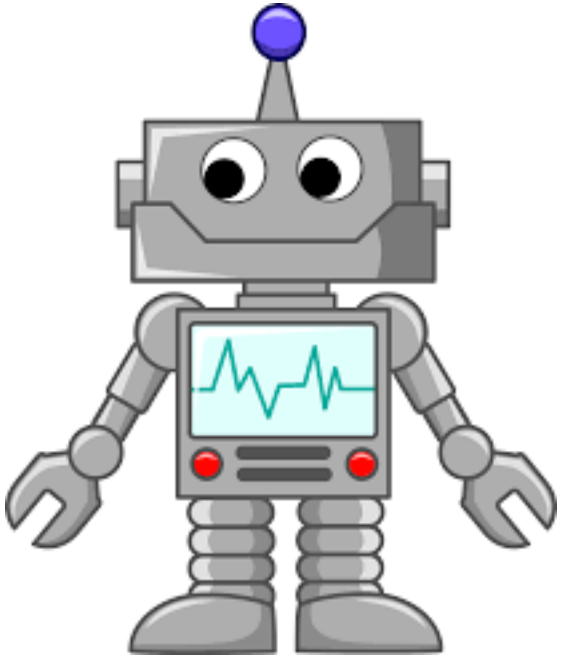
MASC
[Liu and Furukawa 2019]

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

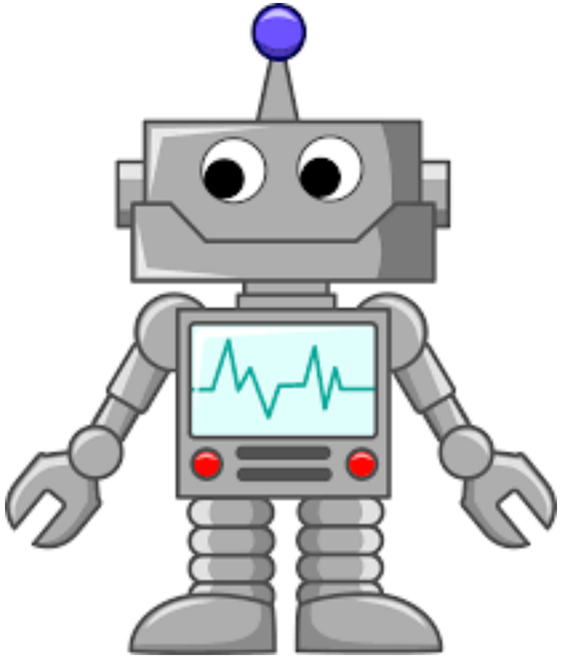
Vision-Language Navigation
[Anderson et al. 2018]

What can we do with language in 3D scenes?

Bring me my coffee cup

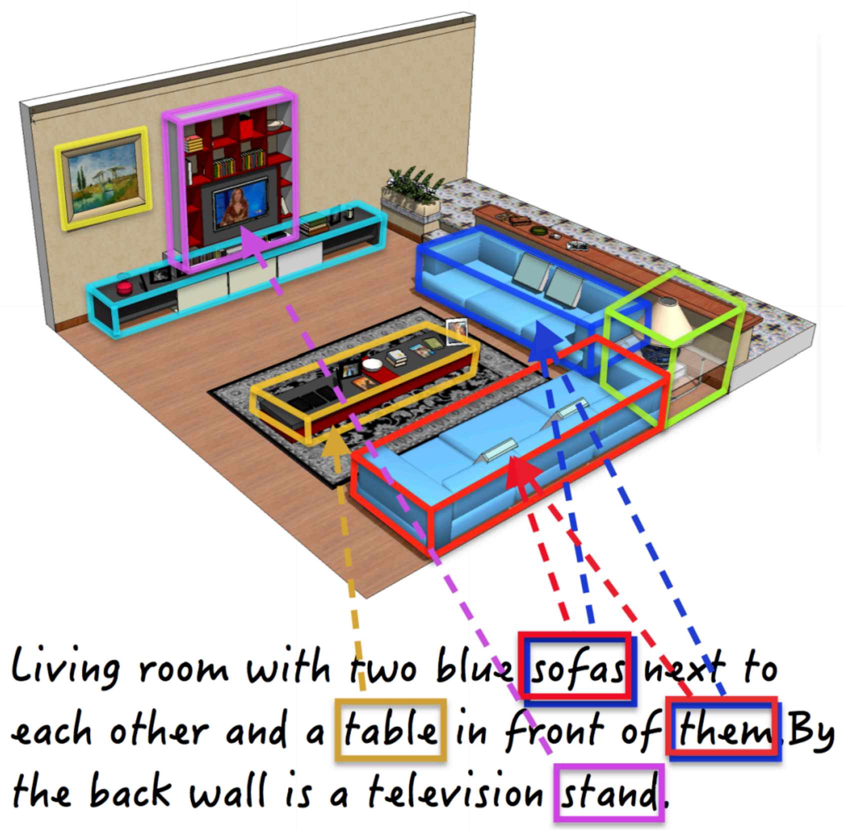


I left my notebook on couch, can you get it for me?



Fundamental task: identifying the object

REVERIE: Workshop Challenge



What are you talking about? Text-to-Image Coreference, Kong et al., CVPR 2014

REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments
Qi et al., CVPR 2020

ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language

Zhenyu (Dave) Chen[1], Angel Chang[2], Matthias Niessner[1]

(to appear ECCV 2020)

1 TECHNISCHE UNIVERSITÄT MÜNCHEN

2 SFU SIMON FRASER UNIVERSITY

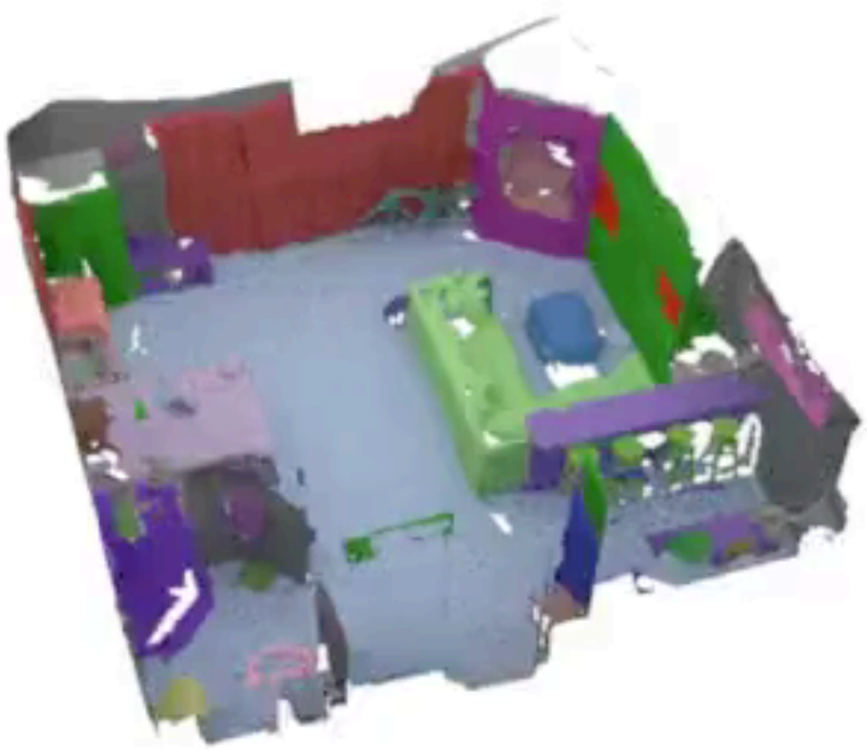
Task

Input

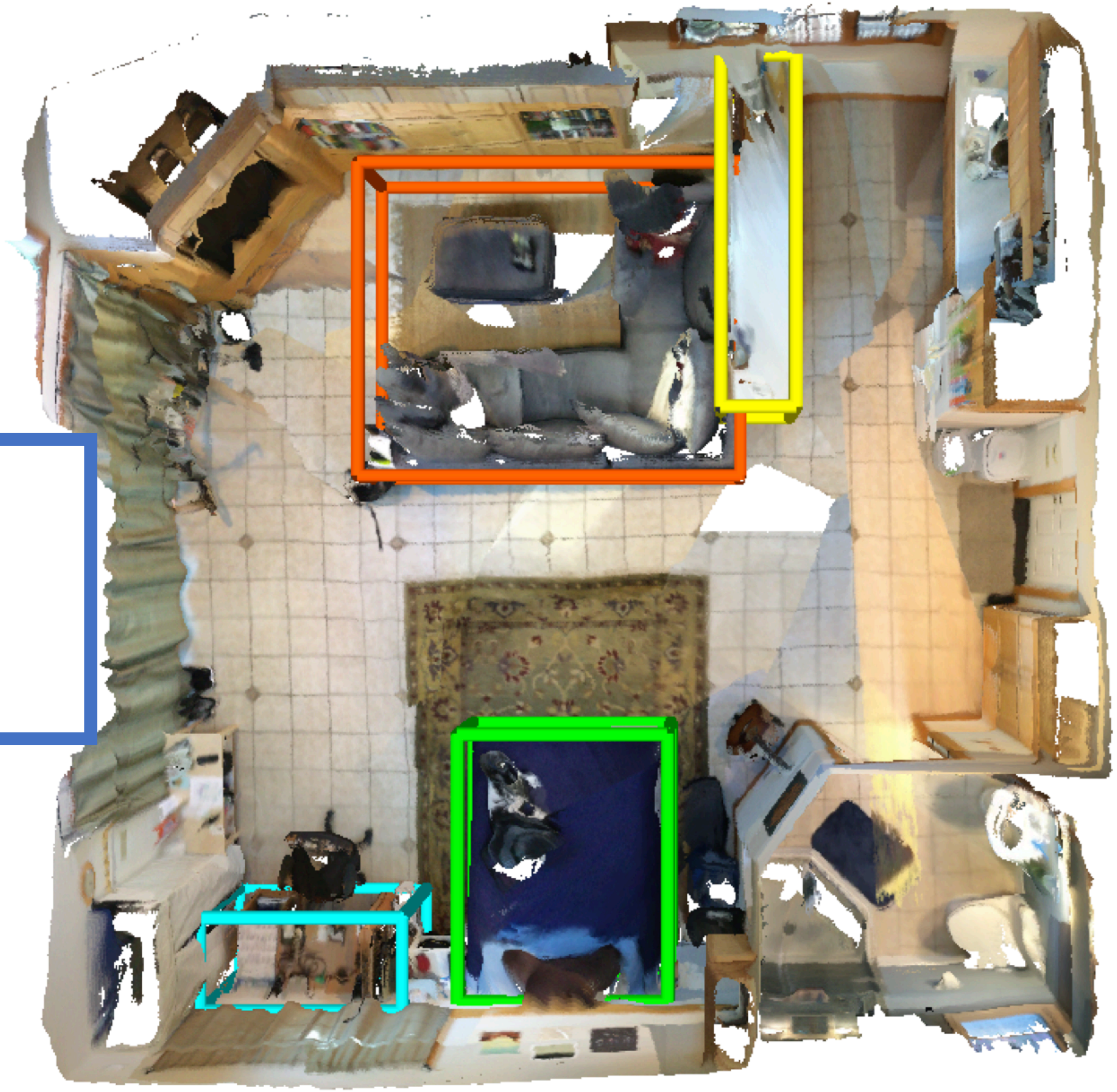
A black chair in the corner. It is next to a table.

Goal

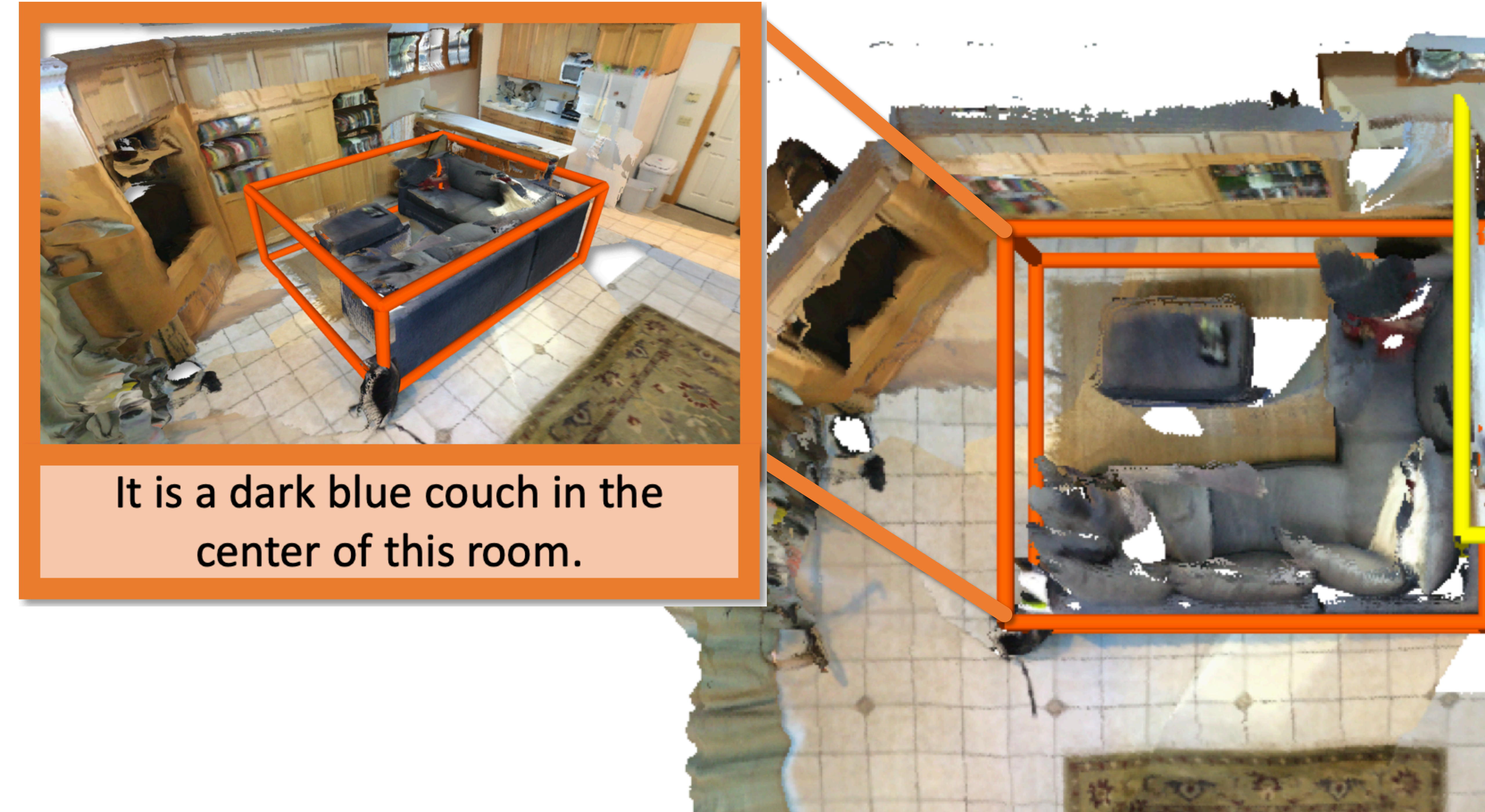
3D Scene



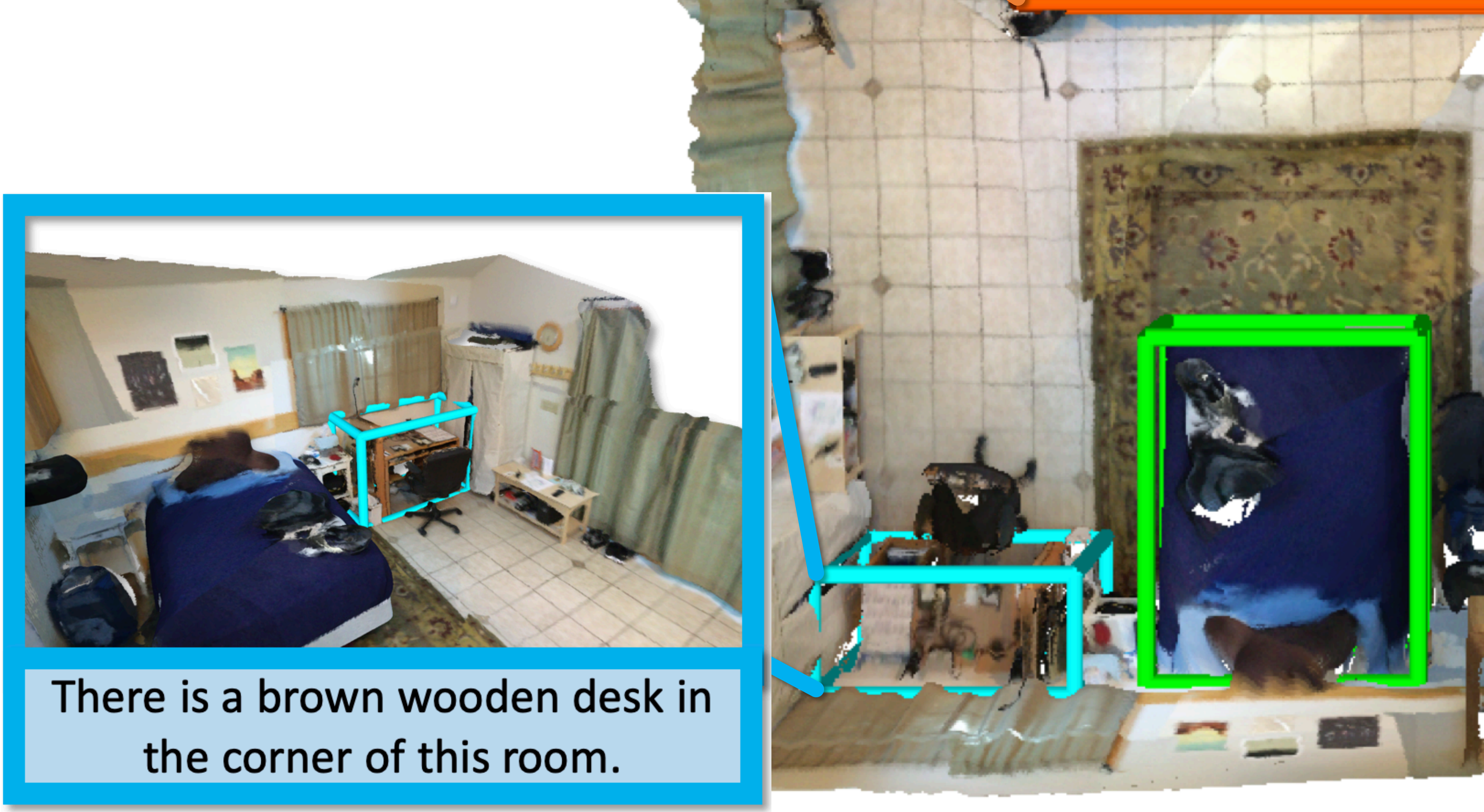
Dataset



Collect descriptions for objects


It is a dark blue couch in the
center of this room.


This is a long bar table behind stools.

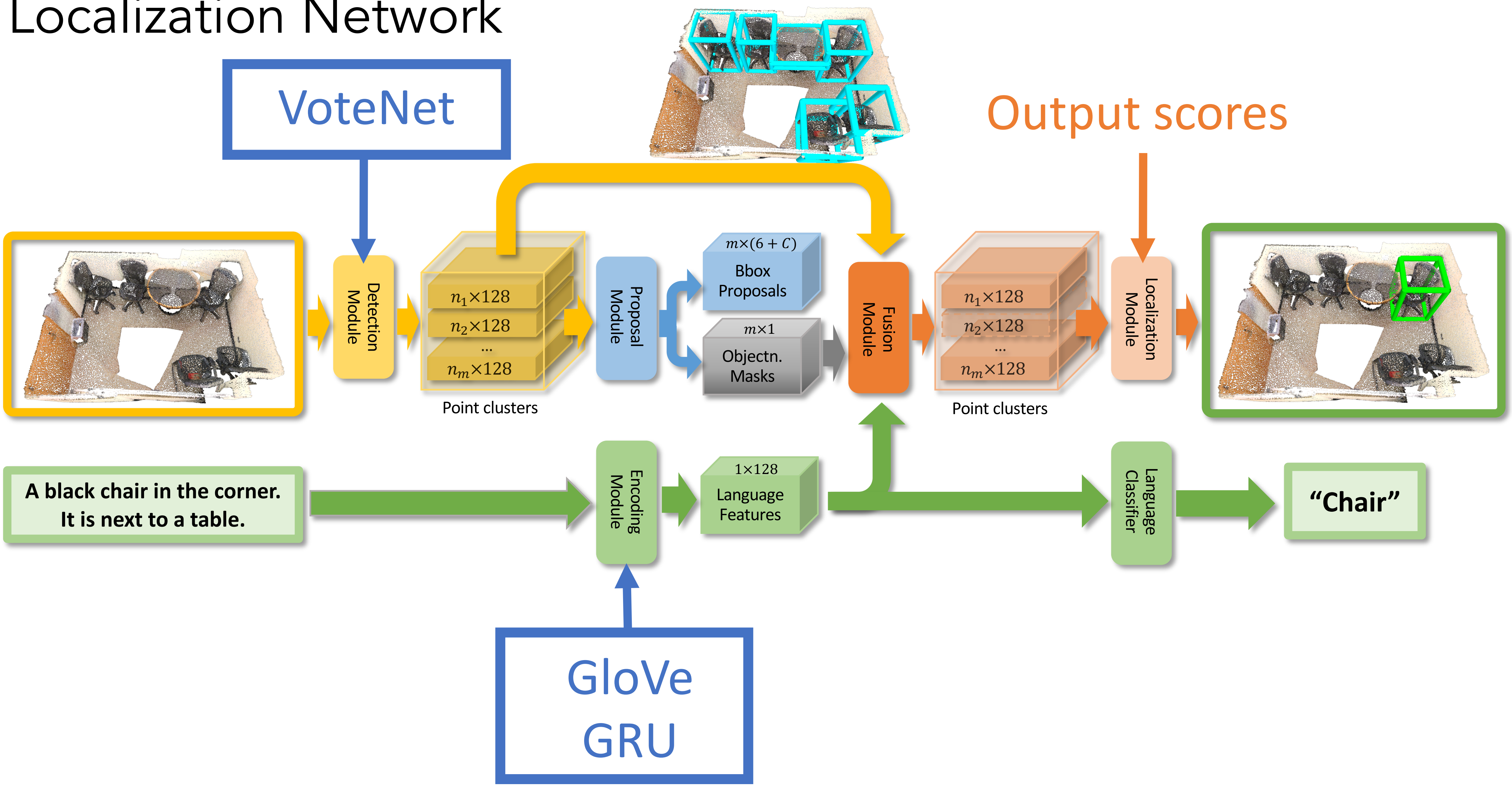

There is a brown wooden desk in the corner of this room.


It is a dark blue couch in the center of this room.

703 scenes
9,976 objects
~5 descriptions per object

49,006 descriptions

Localization Network



VoteNet, Qi et al., CVPR 2019

GloVe, Pennington et al., EMNLP 2014

Training

Overall Loss

$$\mathcal{L} = \alpha\mathcal{L}_{\text{loc}} + \beta\mathcal{L}_{\text{det}} + \gamma\mathcal{L}_{\text{cls}}$$

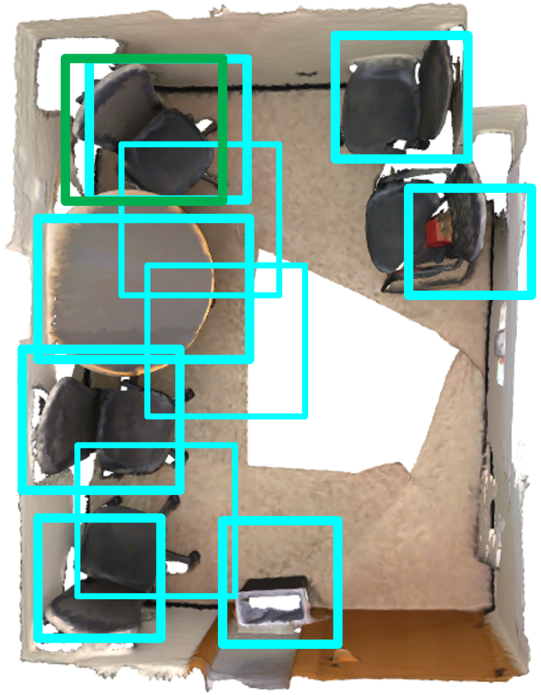
Localization Loss

$$\mathcal{L}_{\text{loc}} = -\sum_{i=1}^{M}[w_{\text{neg}}(1-t_i)\log(1-s_i) + w_{\text{pos}}t_i\log(s_i)]$$

Object Detection Loss

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{vote-reg}} + 0.5\mathcal{L}_{\text{objn-cls}} + \mathcal{L}_{\text{box}} + 0.1\mathcal{L}_{\text{sem-cls}}$$

Semantic Class Loss



Select proposal with highest IoU with ground truth box as target

Can we successfully localize objects using natural language in 3D?

Baseline methods

Semantic segmentation + language features

- based on PointNet++ [Qi et al, NIPS 2017]
- no notion of object instances

PointRefNet

Object detection network + random

- based on deep 3D hough voting [Qi et al, ICCV 2019]
- predicted object categories
- select one at random that matches category

VoteNetRand

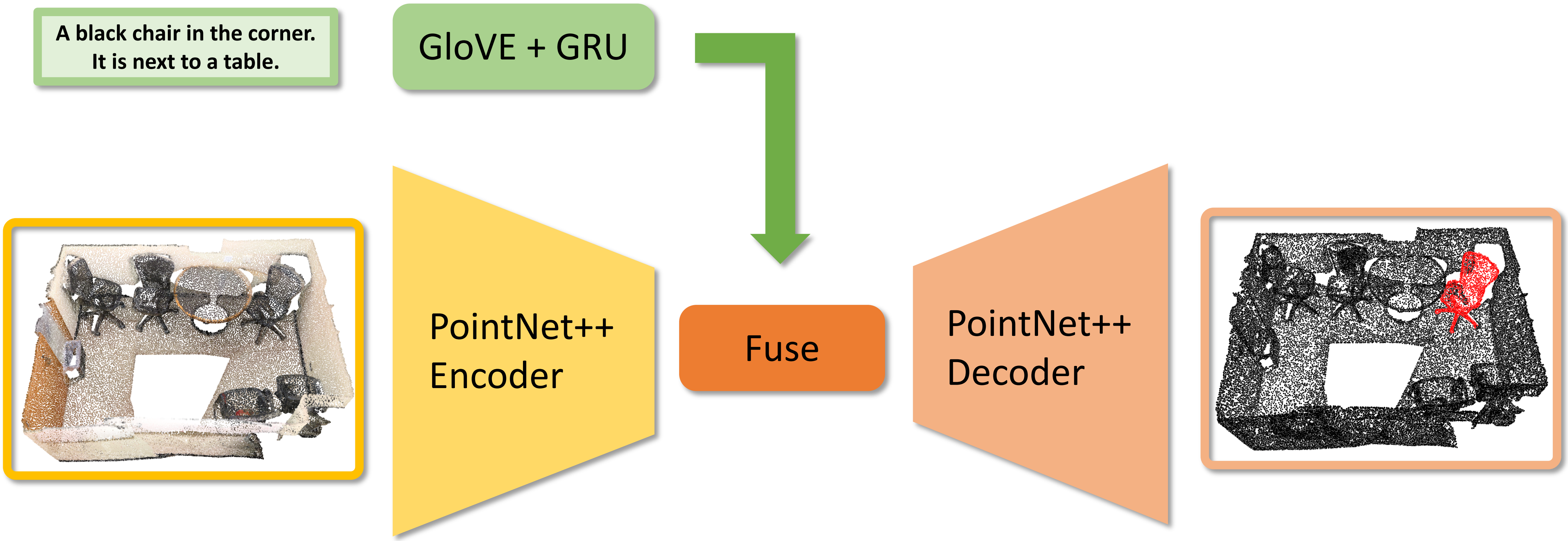
2D referring expression baselines

- SCRC based on [Hu et al, CVPR 2016]
- One stage based on [Yang et al, CVPR 2019]

Best prediction from several views back projected into 3D

2D Projection

Baselines: PointRefNet



PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, Qi et al., CVPR2017

Baselines: VoteNetRand

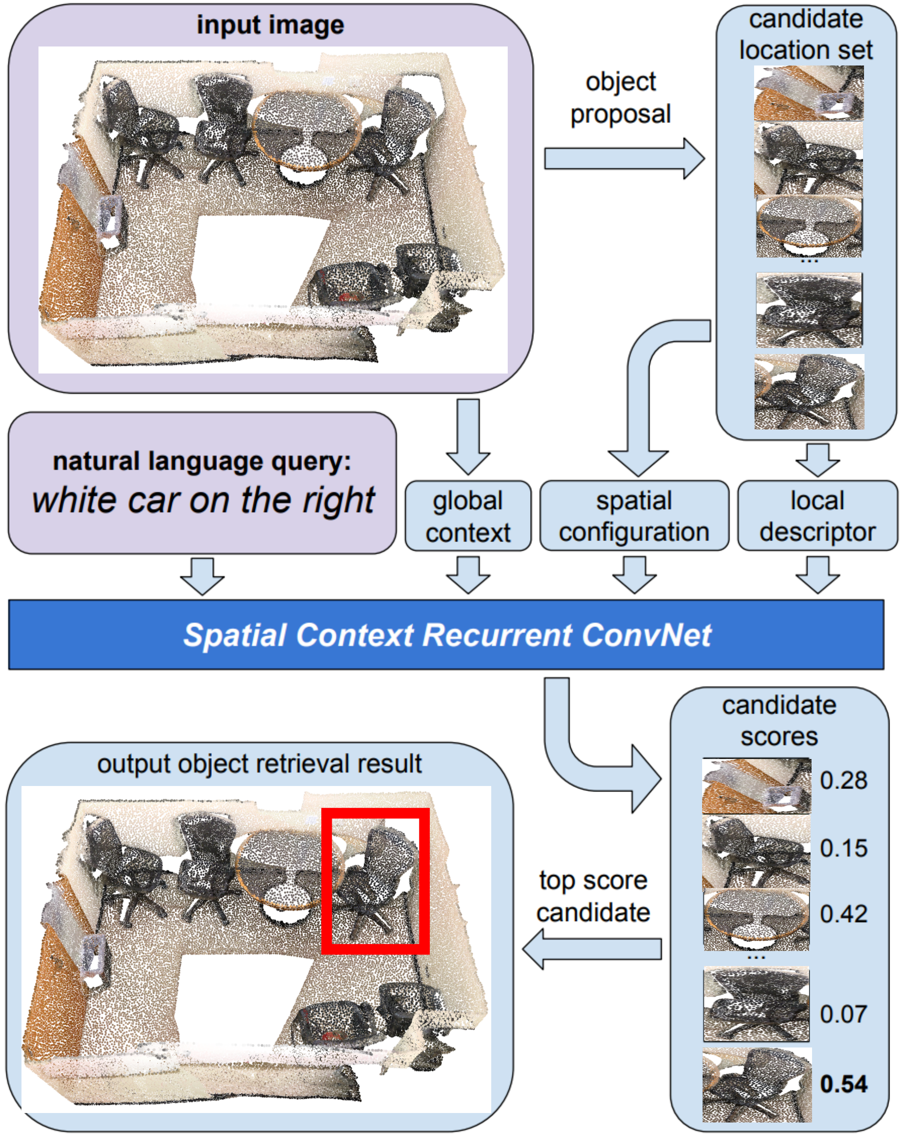
Random Selection

Among correct labels

Proposals

Output

Baselines: 2D referring expression + Projection



Natural Language Object Retrieval, Hu et al., CVPR2016

A Fast and Accurate One-Stage Approach to Visual Grounding, Yang et al., CVPR2019

Evaluation

Precision at IOU of 0.5

	P@0.5
PointRefNet	5.92
VoteNetRand	6.28
2D Proj (SCRC)	6.45
2D Proj (One-Stage)	9.04
Ours (all features)	22.39

PointRefNet: Semantic segmentation network (based on PointNet++ [Qi et al, NIPS 2017]) with language features (no notion of object instances)

VoteNetRand: Object detection network (based on deep 3D hough voting [Qi et al, ICCV 2019]) with predicted object categories, select one at random

2D referring expression baselines (SCRC based on [Hu et al, CVPR 2016] and One stage based on [Yang et al, CVPR 2019]), with best prediction from several views back projected into 3D

Unique

Multiple

This is a white trash can.
It is behind a short white trash can.

This is a trash can with no lid. It is in front of a trash can with a lid.

It is a white refrigerator in a kitchen with brown cabinets. Next to it are two white trash cans.

Precision at IOU of 0.5

	Unique	Multiple	Overall
PointRefNet	12.85	4.71	5.92
VoteNetRand	23.04	3.35	6.28
2D Proj (One-Stage)	22.82	6.49	9.04
Ours (all features)	**39.95**	**18.17**	**22.39**

PointRefNet: Semantic segmentation network with language features (no notion of object instances)
VoteNetRand: Object detection network with predicted object categories, select one at random
One stage: 2D referring expression baseline with best prediction from several views back projected into 3D

Baselines: OracleCatRand (upper baseline)

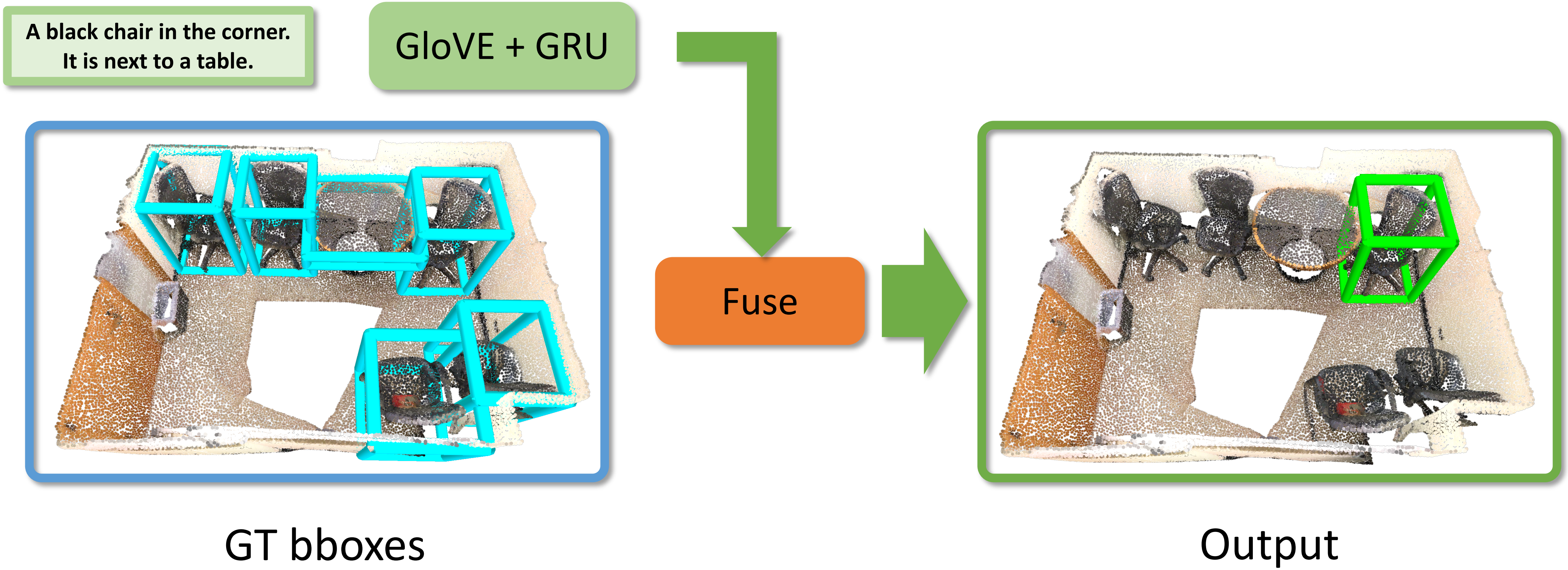
Random
Selection

Among
correct labels

GT bboxes

Output

Baselines: OracleRefer (upper baseline)



Precision at IOU of 0.5

	Unique	Multiple	Overall
OracleCatRand	100.00	17.84	29.76
OracleRefer	73.55	32.00	40.06
PointRefNet	12.85	4.71	5.92
VoteNetRand	23.04	3.35	6.28
2D Proj (One-Stage)	22.82	6.49	9.04
Ours (all features)	39.95	18.17	22.39

OracleCatRand: Perfect bounding boxes and known object categories, select one at random
OracleRefer: Perfect bounding boxes, use language features fused with pointnet features to match
PointRefNet: Semantic segmentation network with language features (no notion of object instances)
VoteNetRand: Object detection network with predicted object categories, select one at random
One stage: 2D referring expression baseline with best prediction from several views back projected into 3D

Precision at IOU of 0.5

	Unique	Multiple	Overall
OracleCatRand	100.00	17.84	29.76
OracleRefer	73.55	32.00	40.06
VoteNetRand	23.04	3.35	6.28
Ours (all features)	39.95	18.17	22.39

Drops significantly

OracleCatRand: Perfect bounding boxes and known object categories, select one at random
OracleRefer: Perfect bounding boxes, use language features fused with pointnet features to match
VoteNetRand: Object detection network with predicted object categories, select one at random

Precision at IOU of 0.5

	Unique	Multiple	Overall
OracleCatRand	100.00	17.84	29.76
OracleRefer	need better object detection	32.00	40.06
VoteNetRand		3.35	6.28
Ours (all features)	39.95	18.17	22.39

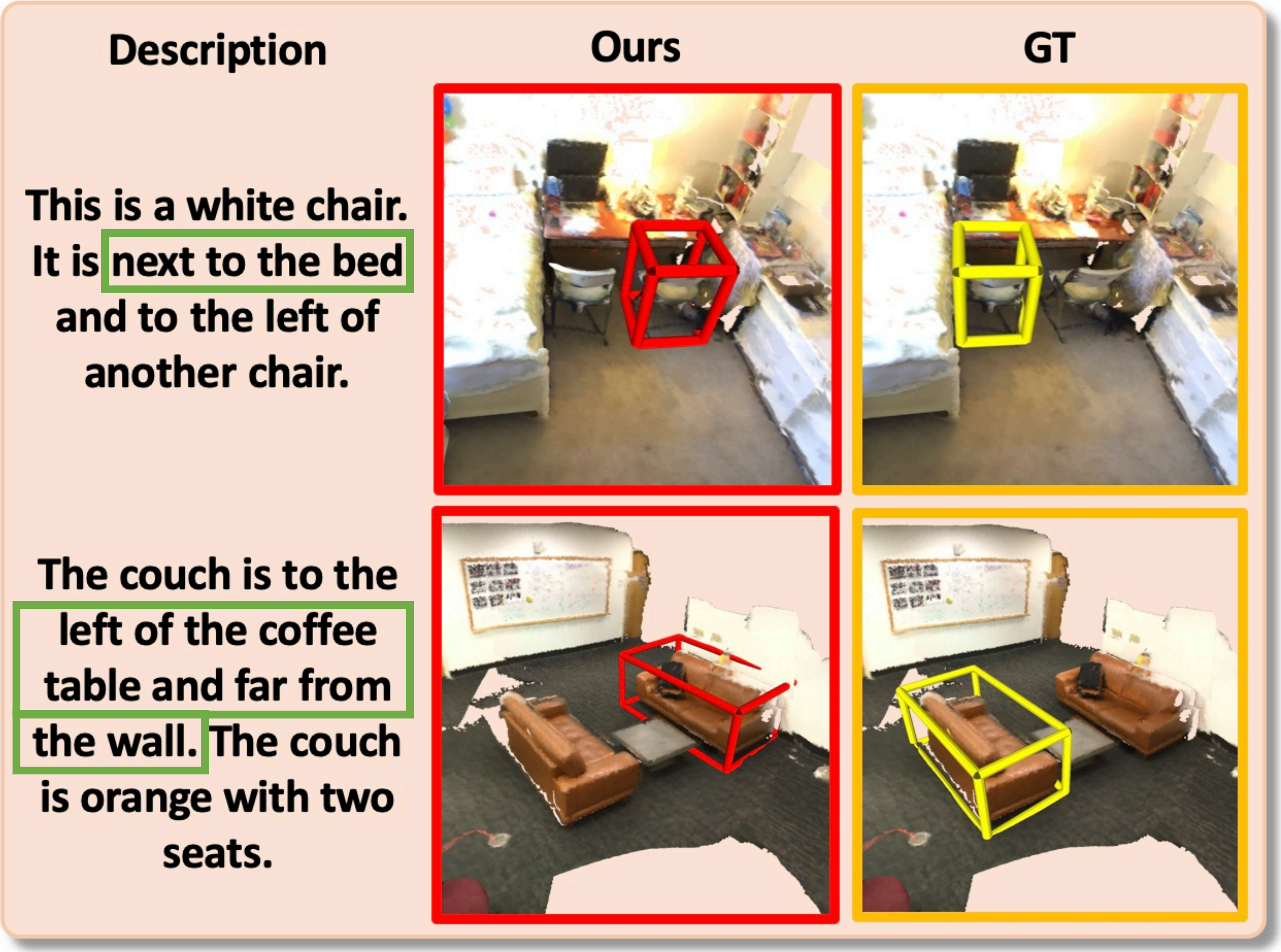
OracleCatRand: Perfect bounding boxes and known object categories, select one at random
OracleRefer: Perfect bounding boxes, use language features fused with pointnet features to match
VoteNetRand: Object detection network with predicted object categories, select one at random

Precision at IOU of 0.5

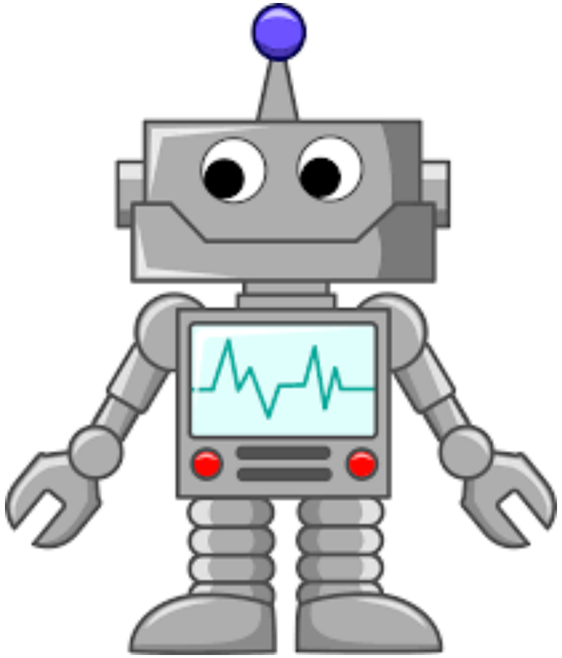
	Unique	Multiple	Overall
OracleCatRand	100.00	17.84	29.76
OracleRefer	73.55	32.00	40.06
VoteNetRand	23.04	3.35	6.28
Ours (all features)	39.95	18.17	22.39

need better disambiguation

OracleCatRand: Perfect bounding boxes and known object categories, select one at random
OracleRefer: Perfect bounding boxes, use language features fused with pointnet features to match
VoteNetRand: Object detection network with predicted object categories, select one at random


Description
Ours
GT
This is a white chair. It is next to the bed and to the left of another chair.
The couch is to the left of the coffee table and far from the wall. The couch is orange with two seats.

I left my notebook on couch, can you get it for me?



Building large-scale interactive environments for grounded language learning

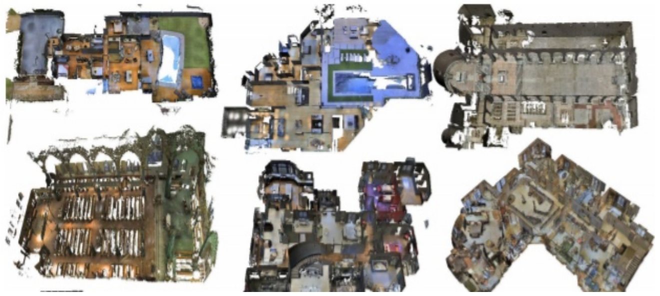
Datasets for semantic understanding in 3D

3D scenes

3D shapes

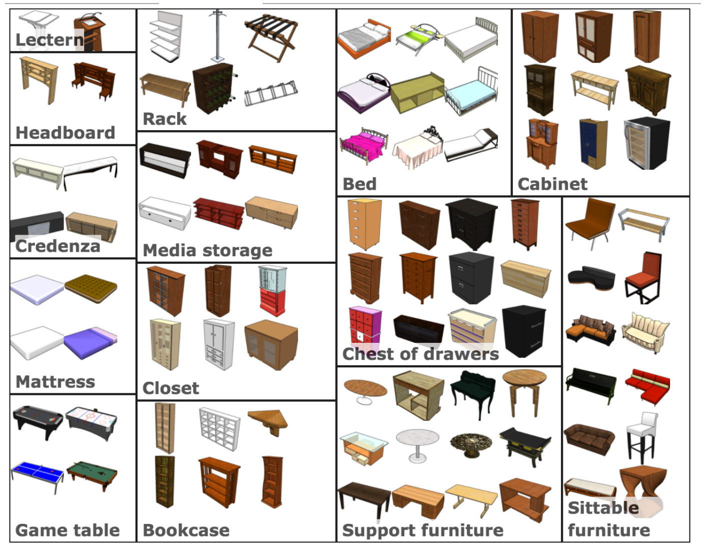
ScanNet

[Dai et al. 2017]



Matterport3D

[Chang et al. 2017]



ShapeNet

[Chang et al. 2015]

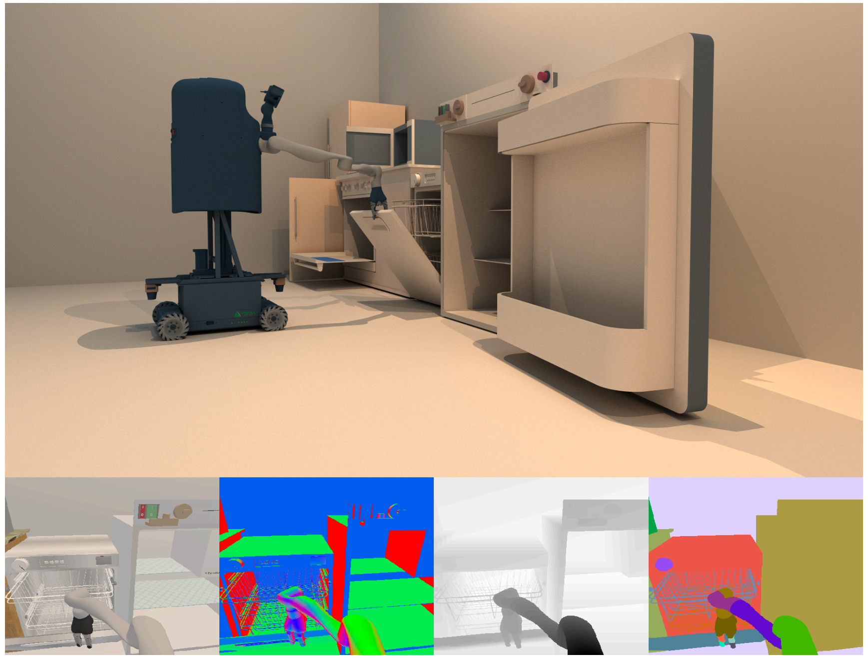
PRINCETON UNIVERSITY

TECHNISCHE UNIVERSITÄT MÜNCHEN

Stanford University

Simulation Environments

MINOS
https://minosworld.github.io/
[Savva et al. 2017]



SAPIEN
https://sapien.ucsd.edu/
[Xiang et al. CVPR 2020]

PRINCETON UNIVERSITY
intel
UC San Diego
Stanford University

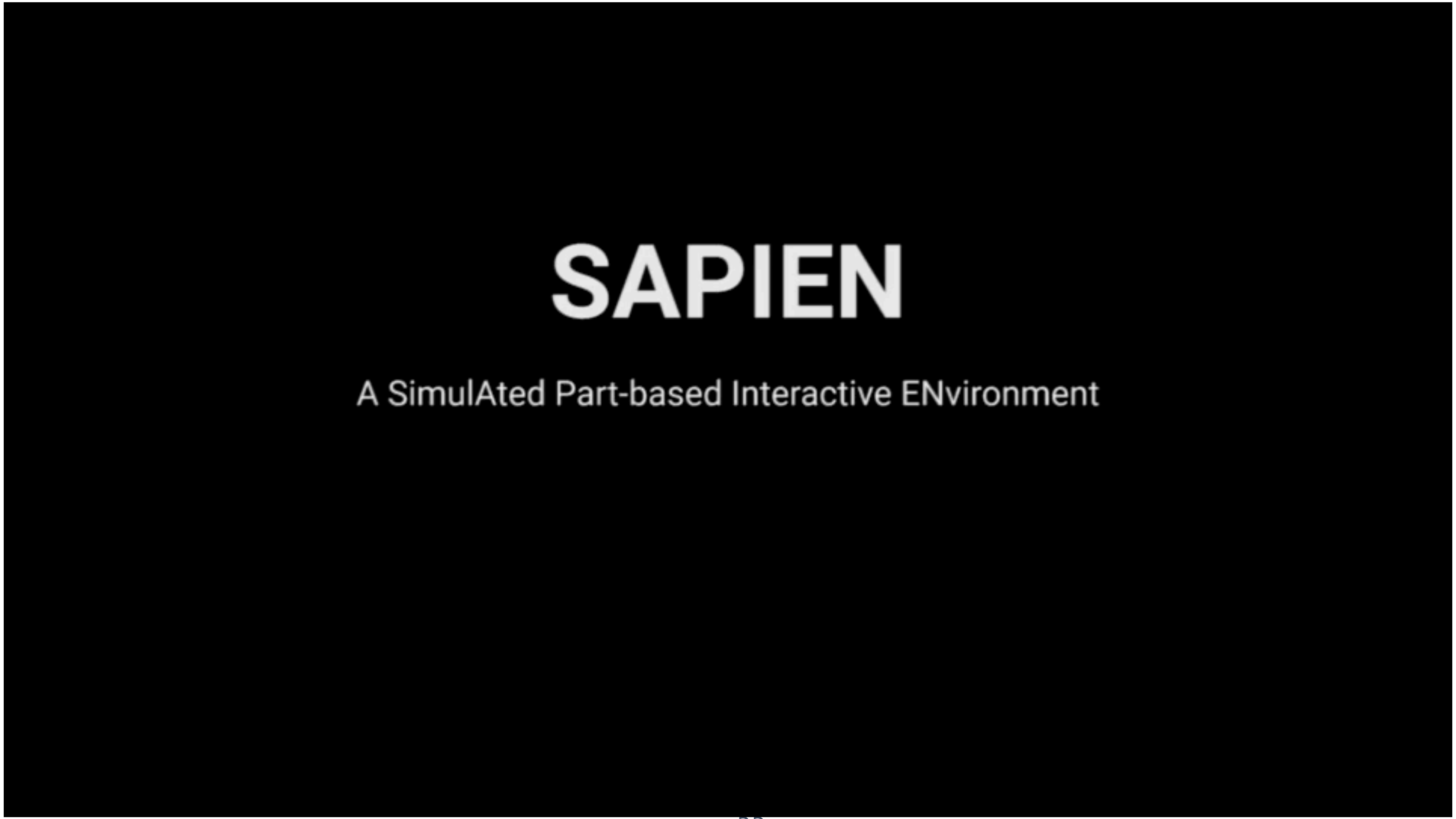
SAPIEN + PartNet Mobility dataset

2,345 objects
46 categories
14,068 moveable parts

PhyX based simulation
framework in c++ and python

Interactions in SAPIEN Demo Video

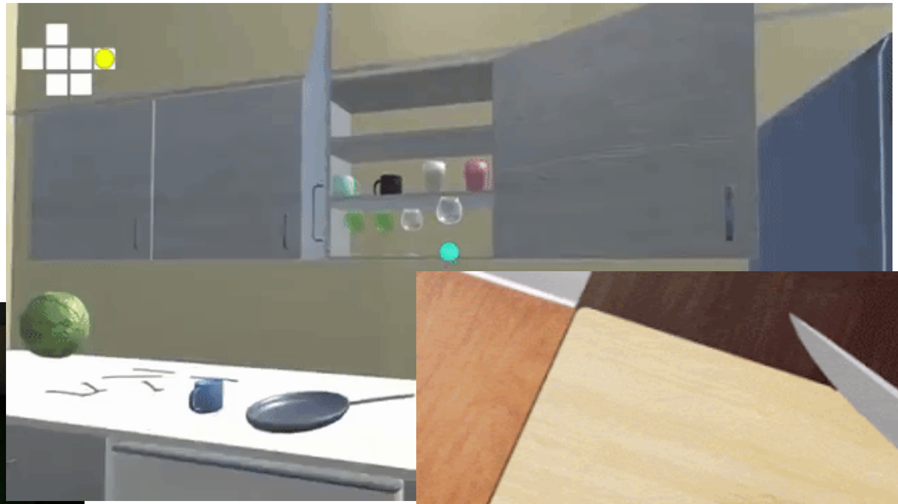
https://www.youtube.com/watch?v=K2yOeJhJXzM&feature=youtu.be



3D environments for interaction

AI2-THOR
[Kolve et al. 2017]

Cornell CHALET
[Yan et al. 2018]



VirtualHome
[Puig et al. 2018]

VRKitchen
[Gao et al. 2019]

3D environments for interaction

AI2-THOR

[Kolve et al. 2017]

ALFRED

[Shridhar et al. 2020]

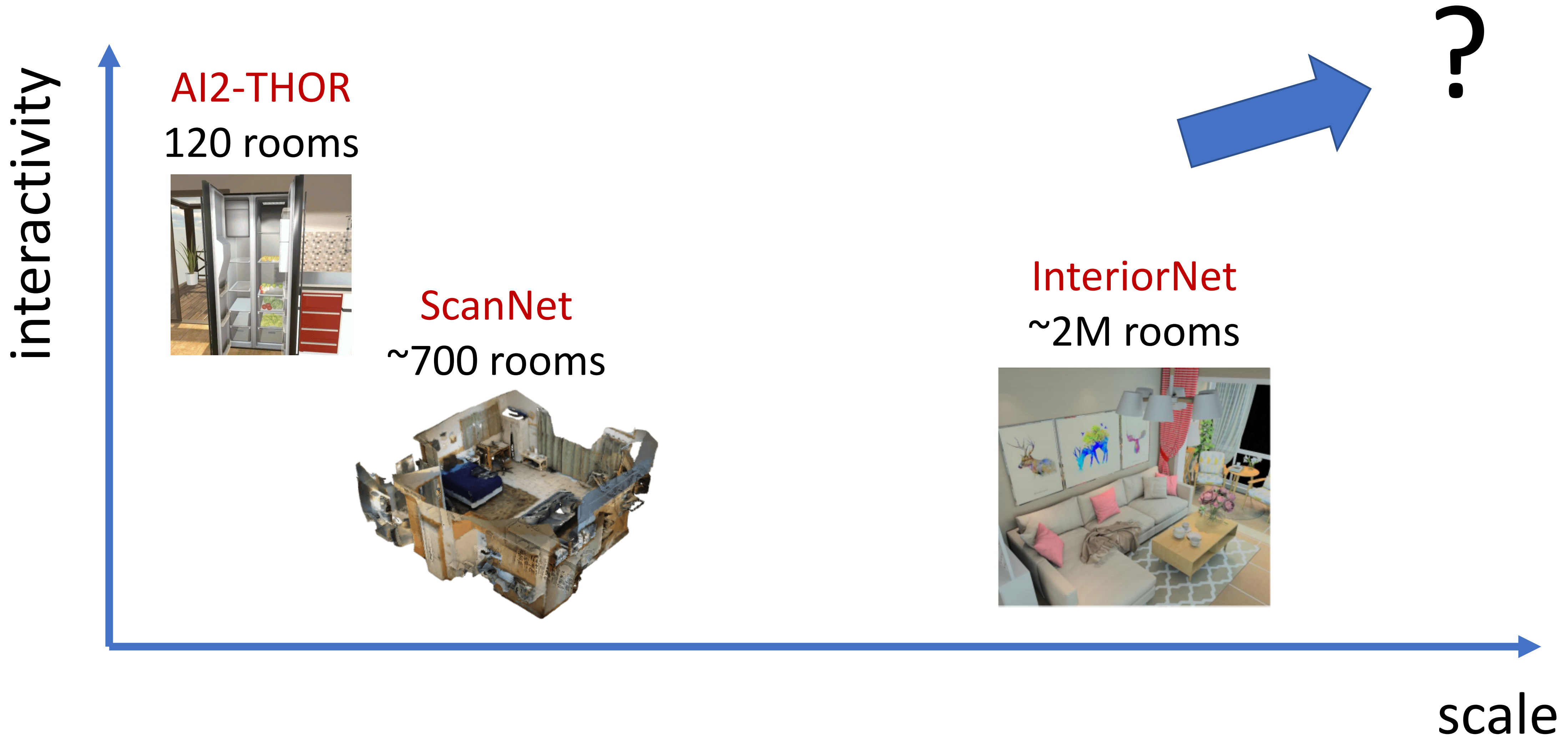
T

3]

VRKitchen

ao et al. 2019]

Scale is limited compared to static datasets



Takeaway messages

- Understanding language requires common sense
- Much of common sense is spatial, relies on anticipation of "what will happen if I do this?"
- 3D representations allow simulation for connecting language with egocentric perception & "world mental model" building

Thank you!