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

About me

PKU

OS Labs, Bachelor
Advisor: Prof. Yao Guo

VCLA@UCLA

Ph.D. of Computer Science
Advisor: Prof. Song-Chun Zhu

Amazon, Alexa AI

Research Intern
Mentor: Dr. Qing Ping

2016-2017 | 2017-2018 | 2018-2019 | 2020 | 2021 | 2023.02

PKU-UCLA JRI 3+2

Master of Computer Science
Advisor: Prof. Song-Chun Zhu

DMAI, Inc

Research Intern
Mentor: Dr. Tao Yuan

BIGAI

Research Scientist

What we (I) expected ☺

Favreau, J. (Director). (2008). Iron Man [Film]. Marvel Studios.

Favreau, J. (Director). (2010). Iron Man 2 [Film]. Marvel Studios.

What we have ☹

Favreau, J. (Director). (2008). Iron Man [Film]. Marvel Studios.

Apple Inc. Introducing Vision Pro (2023, June 5).

Fu et al., Mobile-ALOHA (2024)

Favreau, J. (Director). (2010). Iron Man 2 [Film]. Marvel Studios.

Embodied AI

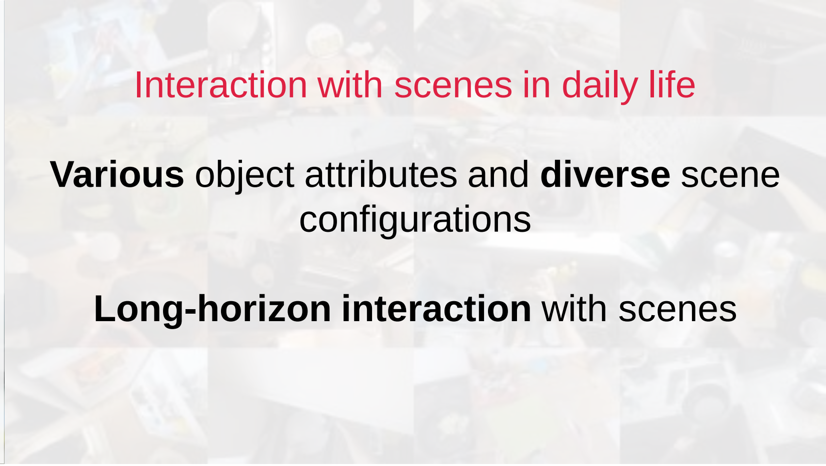
*"The embodiment hypothesis is the idea that **intelligence emerges in the interaction of an agent with an environment** and as a result of sensorimotor activity"*

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

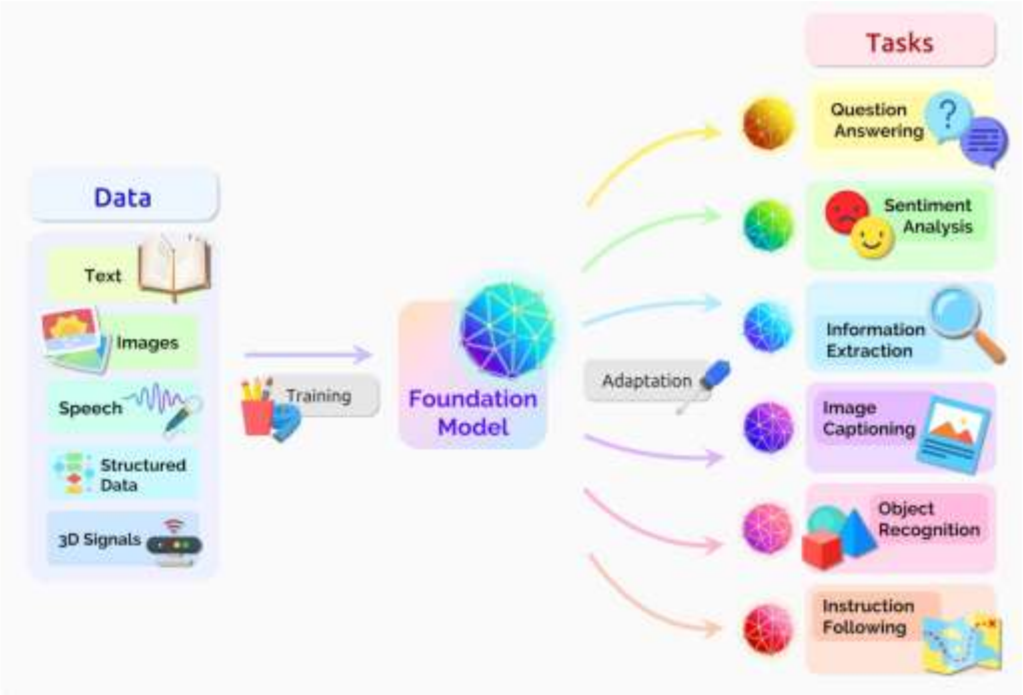
Damen et al., Scaling Egocentric Vision: The Epic-Kitchens Dataset, 2018

The need of generalization !!!

What we learned previously

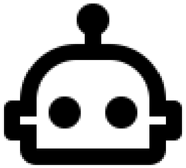
Data Data Data !!!

- ImageNet → Image Understanding
  - Million scale images
- GPT → Language modeling
  - Billion scale texts
- CLIP → Multi-modal alignment
  - Billion scale image-text pairs
- GPT-4V → More modalities
  - Unknown huge size (?)

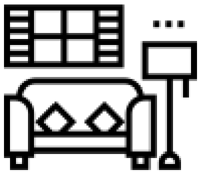


NVIDIA, What are foundation models, 2023
https://blogs.nvidia.com/blog/what-are-foundation-models/

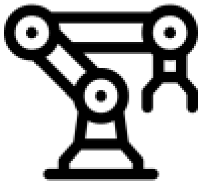
And in Embodied AI?



Perception



Grounding



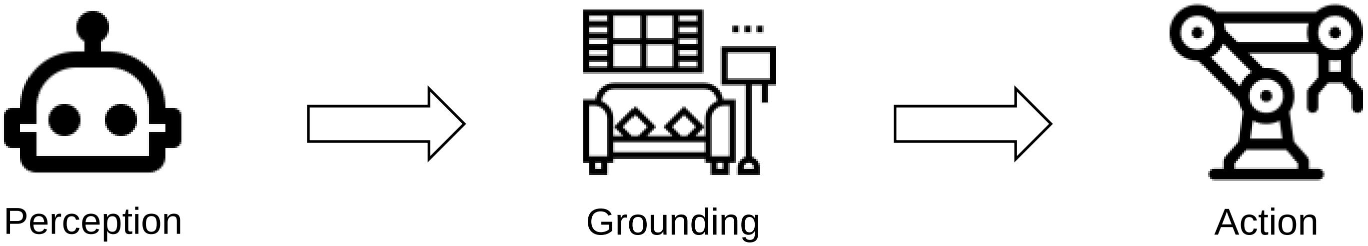
Action

- Object geometry / Physics
- Need to capture 3D
- Aligning captured data
- Representation efficiency
- …

- Object attributes / properties
- Spatial relationships
- Affordance & functionality
- Auto-pipeline / Quality control
- …

- Scene constraints
- Hardware prerequisites
- Data capturing efficiency
- Embodiment gap
- …

From the 3D scene perspective



- Object geometry / Physics
- Need to capture 3D
- Aligning captured data
- Representation efficiency
- …

- Object attributes / properties
- Spatial relationships
- Affordance & functionality
- Auto-pipeline / Quality control
- …

- Scene constraints
- Hardware prerequisites
- Data capturing efficiency
- Embodiment gap
- …

Q1: Is current data sufficient? Can we make full use of them?

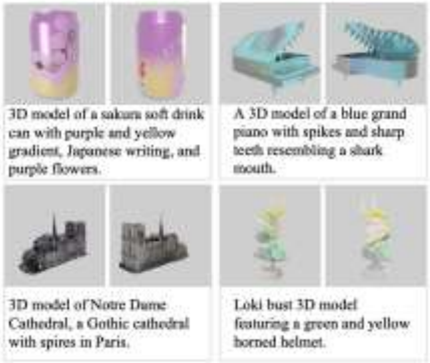
SceneVerse

Scaling 3D Vision-Language Learning for Grounded Scene Understanding

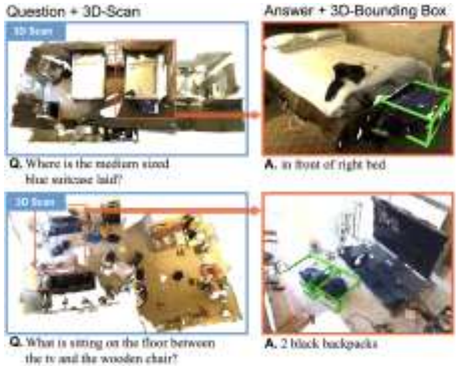
ECCV 2024

Existing Datasets for 3D-VL

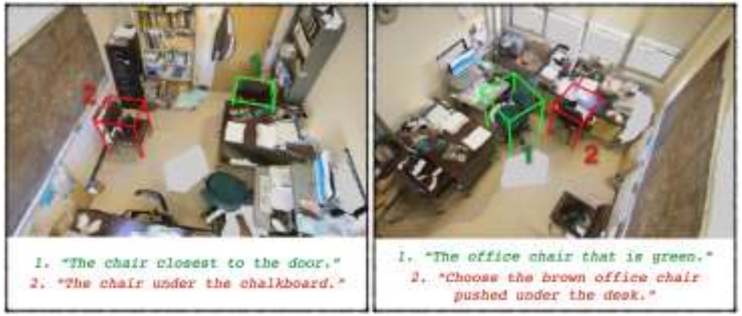
ScanRefer (Chen et al. 2020)



Cap3D (Luo et al. 2023)



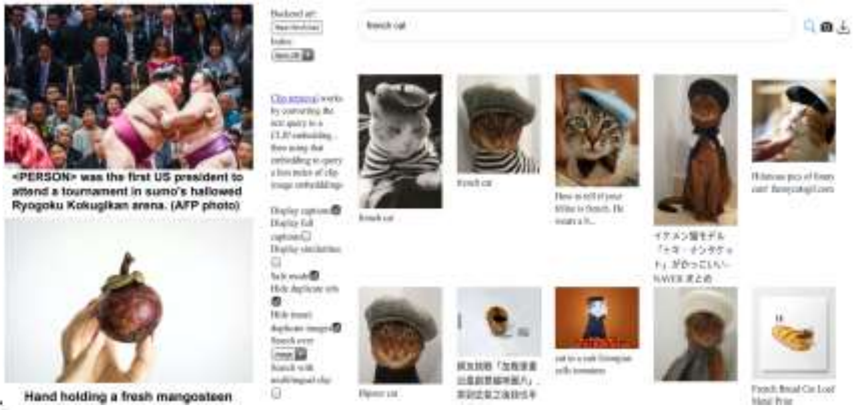
ScanQA (Azuma et al. 2022)



ReferIt3D (Achlioptas et al. 2020)

SQA3D (Ma et al. 2023)

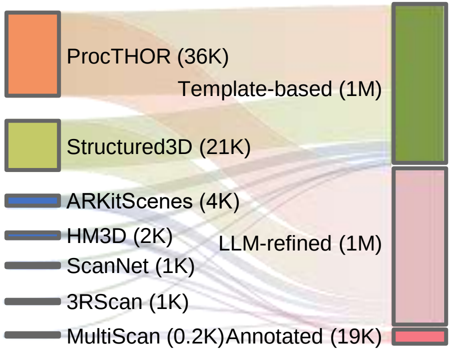
Existing Datasets for 3D-VL



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
SceneVerse	68K	1.5M	190K	2.3M	2.5M

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

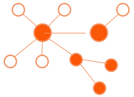
Scalable generation



Scene Caption

Sub-graph Context

{ 'scene_type': 'Bedroom',
'object_count': {'nightstand':2, ...},
'relation': {'nightstand', 'on', 'floor'},
{'backback', 'in front of', bed}, ...}



3D Sub-graph

Summary

Prompt: Provide a summary for a scene from a given scene graph delimited by triple backticks, …

Response: In this bedroom, there are two nightstands, ... The backpack is in front of the nightstand as well. The room appears to be functional, with the nightstands providing storage space and the telephone for communication.

Object Caption

BLIP2 Captions

1. A bed in a hotel room. (0.85)
2. A white comforter on a bed. (0.83)
3. A bed with a striped comforter. (0.83)

…

N. A picture of cat. (0.63)

Multiview Images

Summary

Prompt: Summarize the captions below. The summary should be a description of the {object}. Focus on the {object}'s attributes, like color, shape, material, etc.
Identify and correct the potential errors …

Response: The bed is in a hotel room with a striped comforter. It has a white comforter and a blanket on it.
The bed is also in a room with a bedside table.

Object Referral

Relationship Triplets

1. {'table', 'chair', 'left'},
2. {'bed', ('lamp', 'mini fridge'), 'between'}

Template-based Referral

1. The table is to the left of the chair.
2. It's a bed in the middle of a lamp and the mini fridge.

Rephrasing

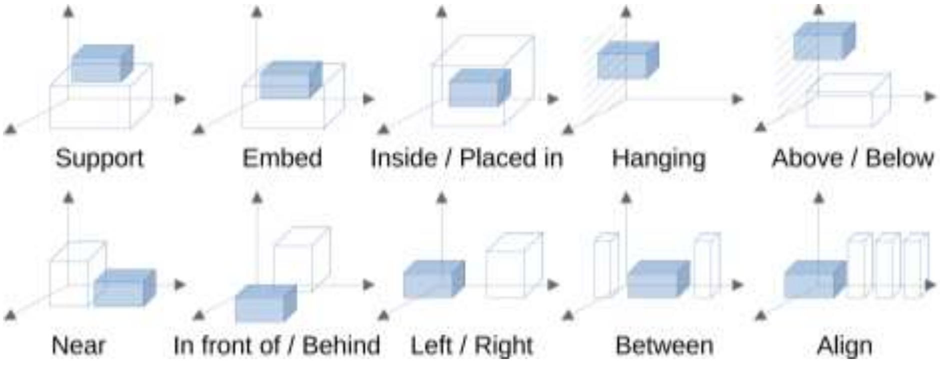
Prompt: Rewrite the following sentence using one random sentence structure. Focus on the location and relationships about the {target_object}, …

Response:

1. The table is situated to the left of the armchair.
2. The bed occupies the space between the lamp and the mini fridge, creating a cozy atmosphere.

Scalable generation

- Scene graph construction
  - Leverage instance annotation
  - Define relationship primitives
  - Identify relationship based on bbox



Scalable generation

- Scene graph construction
  - Leverage instance annotation
  - Define relationship primitives
  - Identify relationship based on bbox
- Language generation with templates
  - Pair-wise: *"There is a target-object that (is) spatial-relation the anchor-object."*
    Ex. "There is picture that is hanging on the wall."
  - Multi-objects: *"The target-object object is spatial-relation with anchor-object1 and anchor-object2."*
    Ex. "There is a cabinet that is between the sofa and TV."
  - Star reference: *"The target-object object is spatial-relation1 with anchor-object1, spatial-relation2 with anchor-object2, …"*
    Ex. "The table is next to the counter, supporting plates, between chair-1 and chair-2."
  - **View-dependent relationships !!!**
    - Left / right, in front of / behind, …
    - *"Facing the sofa"*, "*Facing the TV*", "Facing the bookshelf"
    - "Facing the table"?

Scalable generation

- Refinement with LLMs
  - Generate natural and diverse descriptions
  - Avoid predictions / revisions errors
  - Focus on commonsense information like attributes, spatial relationships, functionality, affordance, etc.

Description type	Prompt
Object caption	Summarize caption below. The summary should be a description of the target-object. Focus on the target-object's attribute, like color, shape and material, *etc*. Identify and correct the potential errors. caption: *A bed in a hotel room. A white comforter on a bed. A bed with a striped comforter...* target-object: *Bed*
Object referral	Rewrite the following caption using one random sentence structure. You should give me only one rewritten sentence without explanation. caption: *The bed is between desk and nightstand.*
	Rewrite the following caption. You should give me only one rewritten sentence about target-object without explanation. Make sure target-object is the subject of the sentence, not anchor-object(s). If the sentence is in full inversion, keep the inversion. caption: *The armchair is next to the sofa.* target-object: *Armchair* anchor-object(s): *Sofa*
	Rewrite the following caption using one random sentence structure. You need to focus on the location and relations of the target-object that appears in the sentence. If multiple target-object appear in the sentence, you need to focus on the first target-object that appears. You can also add the target-object's function and comfort level based on the sentence, e.g., how the objects can be used by humans and human activities in the scene. You should give me only one rewritten sentence without explanation. caption: *Far from the bowl and peppershaker, the vase is to the left, it is also on the top of countertop.* target-object: *Vase*
Scene captioning	Your task is to provide a summary for a scene from a given scene graph. The scene contains some objects, which compose a scene graph in json format. There are 3 types of descriptions in scene graph: "scene type" denotes the type of the scene. "objects count" then listed the objects in the scene and their quantity, it should be noted that the actual objects in the room may be more than listed. "objects relations" describe the spatial relations with objects. Also describe the scene concerning commonsense, e.g., how the objects can be used by human and human activity in the scene. The description should conform to the given scene graph. The spatial relations between objects can only be inferred from the "objects relations" in scene graph. Don't describe each object in the scene, pick some objects of the scene for summary. Don't describe each relations in the scene, pick some relations of the scene for summary. You can also summarize the room's function, style, and comfort level based on the arrangement and count of objects within the room. The summary should be about the object types, object attributes, relative positions between objects. Your summary must not exceed 80 words. You must write using one random sentence structure. scene graph: `{ 'scene_type': 'Bedroom', 'object_count': {'nightstand':2, ...}, 'relation': {'nightstand', 'on', 'floor'}, {'backback', 'in front of', bed}, ...}`

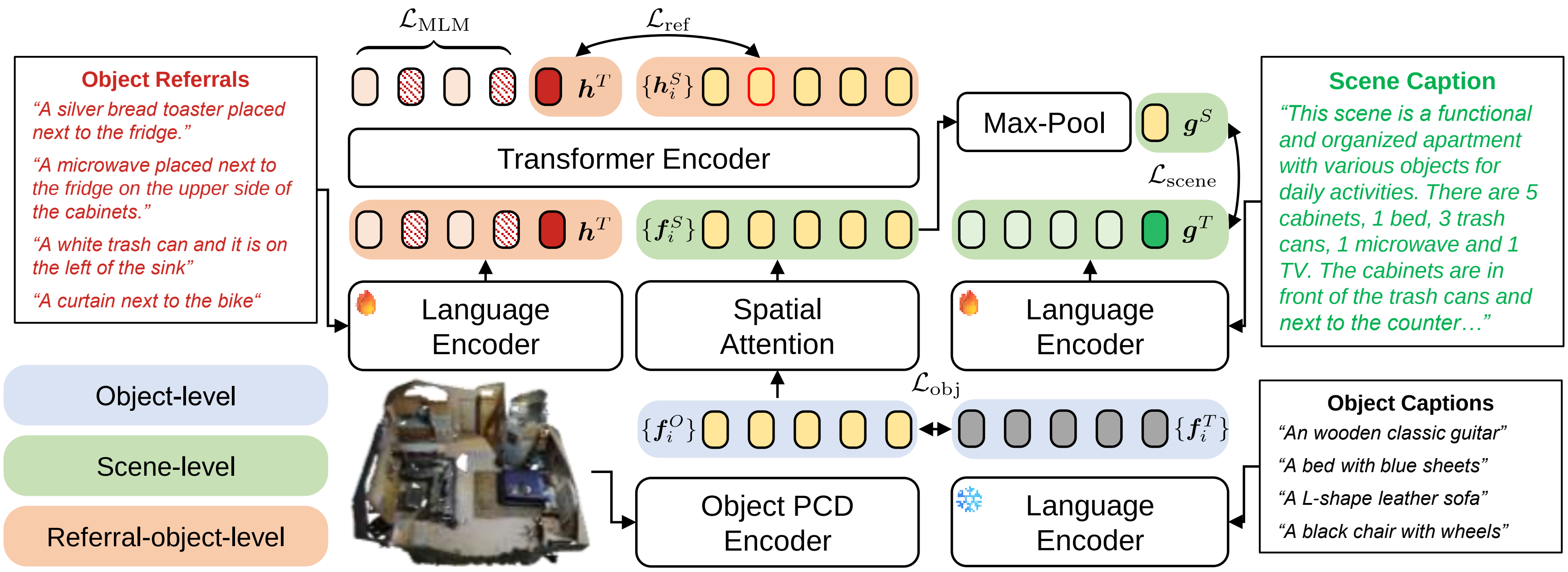
Scene Captioning

Scene Captioning

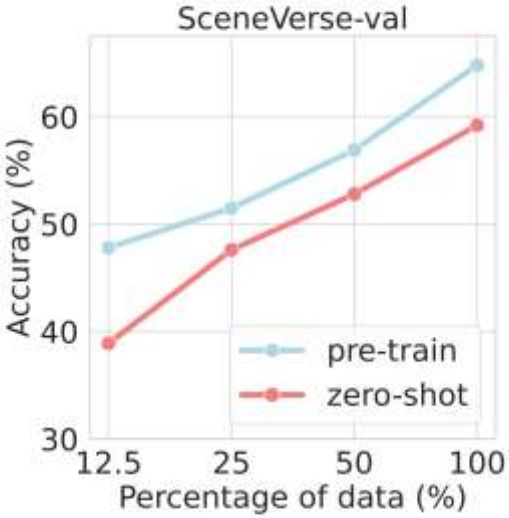
Scene Captioning

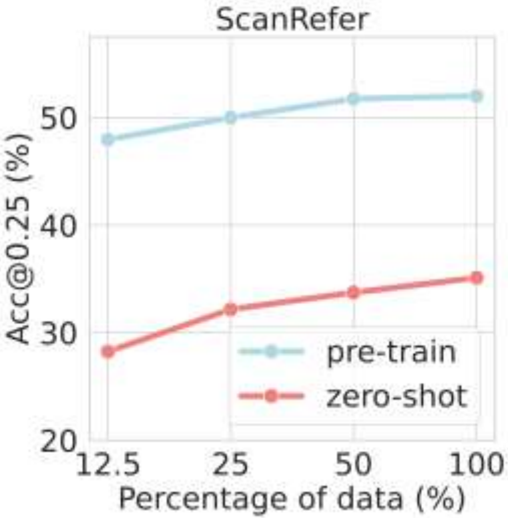
Scene Captioning

Grounded Pre-training for Scenes



Data scaling with SceneVerse





Method	Overall	Easy	Hard	V-Dep.	V-Indep.
3D-VisTA (*scratch*)	40.7	53.1	21.6	37.3	44.3
3D-VisTA (*zero-shot*)	52.9	59.6	35.4	53.7	52.2
3D-VisTA (*zero-shot text*)	58.1	70.0	39.6	52.5	64.1
Ours (*scratch*)	38.5	50.2	20.8	33.7	43.9
Ours (*zero-shot*)	59.2	69.4	44.0	53.1	66.3
Ours (*zero-shot text*)	60.6	70.9	45.1	54.8	67.3

3D Object grounding

- *Zero-shot:* pre-train then test on unseen scenes and texts
- *Zero-shot text:* pre-train then test on seen scenes and unseen texts

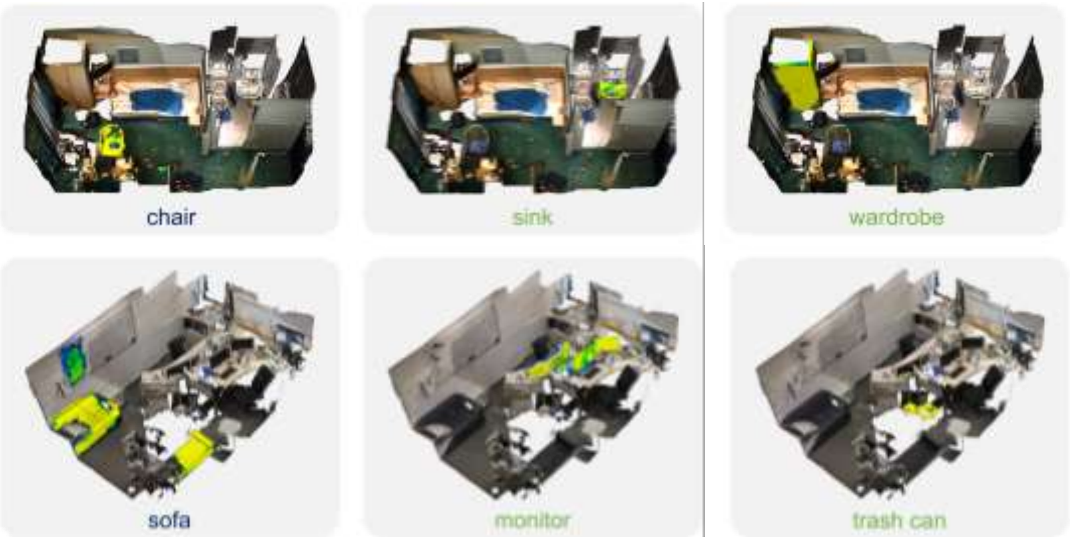
Data scaling with SceneVerse

Model	ScanQA val	ScanQA w/obj	ScanQA w/o obj	SQA3D
ScanRefer+MCAN [5]	18.6	20.6	19.0	-
ScanQA [5]	20.3	23.5	20.9	46.6
SQA3D [59]	-	-	-	47.2
3D-VisTA [101]	22.4	**27.0**	23.0	48.5
3D-LLM [39]	20.5	19.1	-	-
Ours	**22.7**	25.0	**23.5**	**49.9**

Question answering

Model	Network	mIoU	Δ	mAcc	Δ
OpenScene [66]	SPUNet16	57.2	-	69.9	-
PLA [29]	SPUNet16	17.7	-	33.5	-
RegionPLC [87]	SPUNet16	56.9	-	75.6	-
RegionPLC+SceneVerse	SPUNet16	**58.2**	+1.7%	**77.3**	+2.2%
OpenScene [66]	SPUNet32	57.8	-	70.3	-
PLA [29]	SPUNet32	19.1	-	41.5	-
RegionPLC [87]	SPUNet32	59.6	-	77.5	-
RegionPLC+SceneVerse	SPUNet32	**61.0**	+2.3%	**79.7**	+2.8%

Open-vocabulary Segmentation



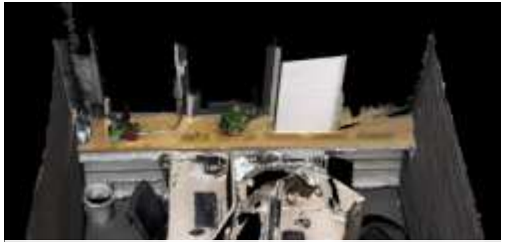
Qualitative visualization of open-vocabulary 3D segmentation prediction

Limitations

- Modality gaps
  - Gap between synthetic and real data exists
- Language quality
  - Quality of language matters
- Scaling scene is still necessary
  - Many tail classes in ScanNet 607

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


"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."

Takeaways

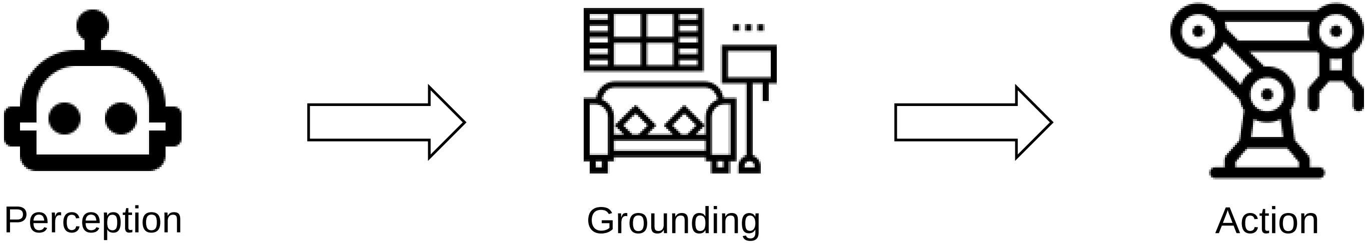
Good:

- Scaling works! We see signs of generalization capabilities.
- Auto-generated data also works! Maybe we can expect more from them.

Bad:

- **Ambiguities in language descriptions** (e.g. left/right) needs resolving.
- **Modality gap** between real-world scenes, real and synthetic scenes.
- Scaling language is easy, **scaling scenes is still hard**.

From the 3D scene perspective



- Object geometry / Physics
- Need to capture 3D
- Aligning captured data
- Representation efficiency
- …

- Object attributes / properties
- Spatial relationships
- Affordance & functionality
- Auto-pipeline / Quality control
- …

- Scene constraints
- Hardware prerequisites
- Data capturing efficiency
- Embodiment gap
- …

Q2: How to scalably obtain “real” scenes with correct physics and fine details?

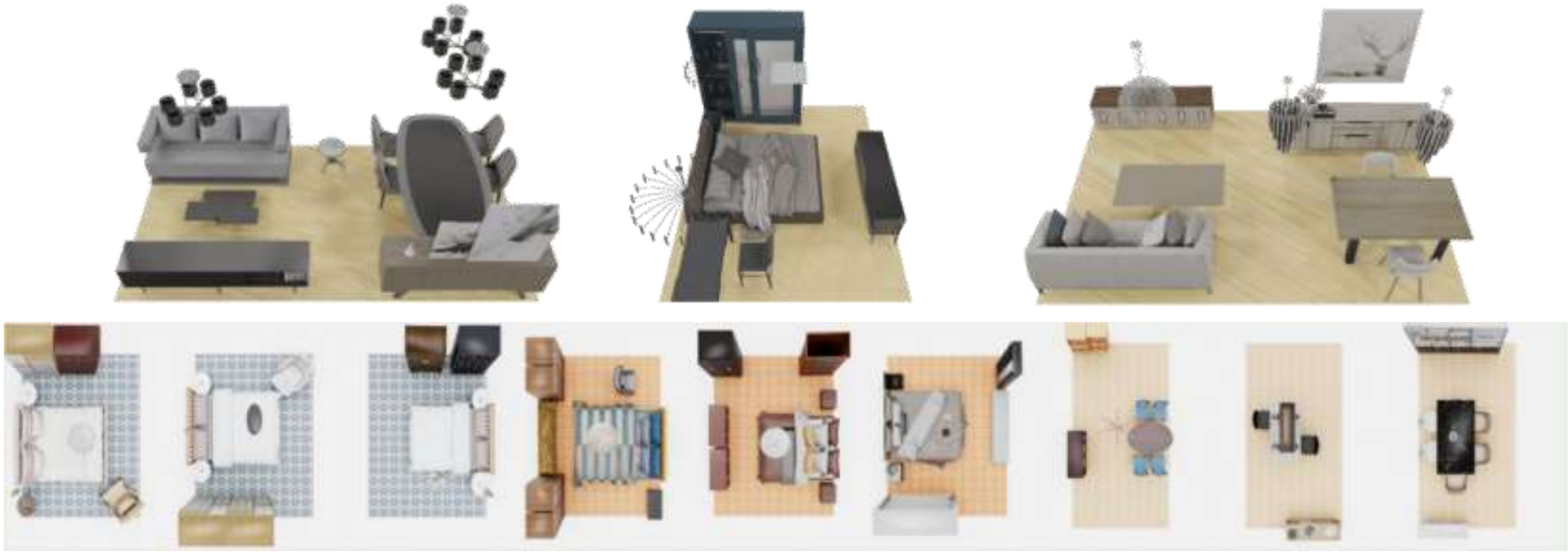
PhyScene

Physically Interactable 3D Scene Synthesis for Embodied AI

CVPR 2024 Highlight

Scene generation

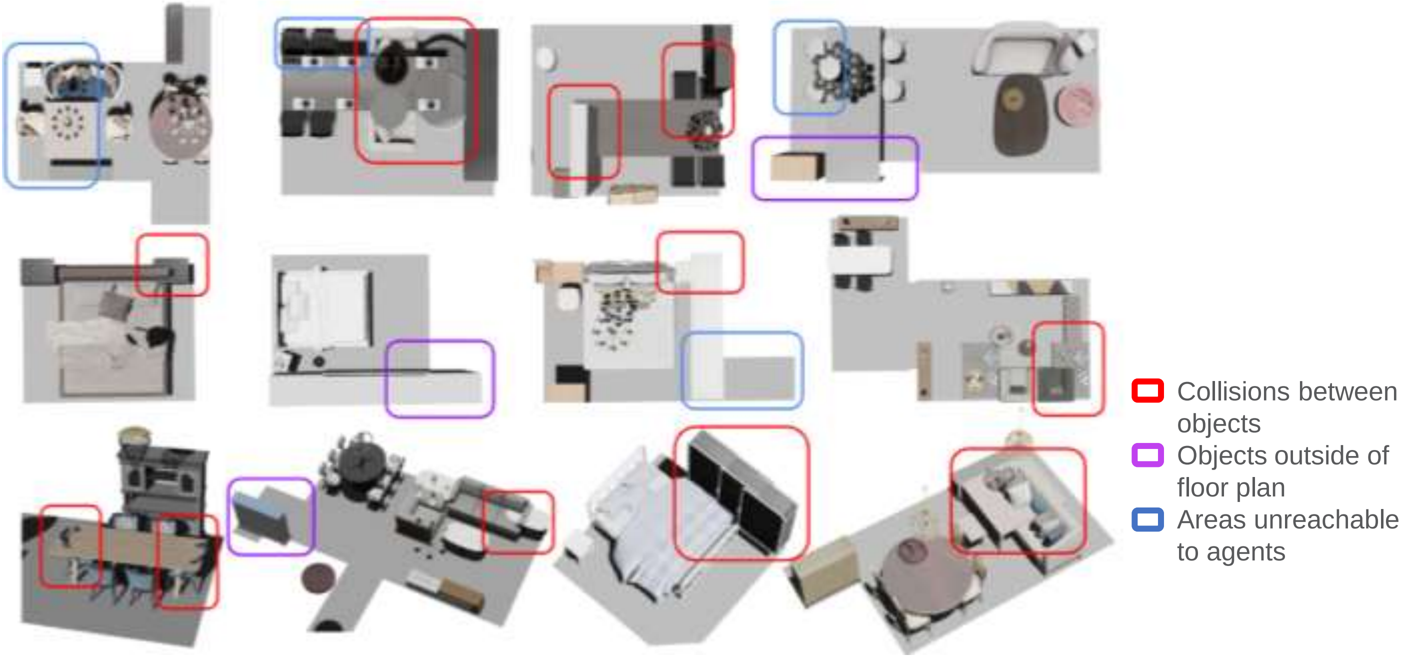
- Generate layouts with artist designed furniture assets to facilitate indoor design



Paschalidou et al., ATISS: Autoregressive Transformers for Indoor Scene Synthesis, NeurIPS 2018

Sadly…

- Too much effort needed for satisfying physics constraints

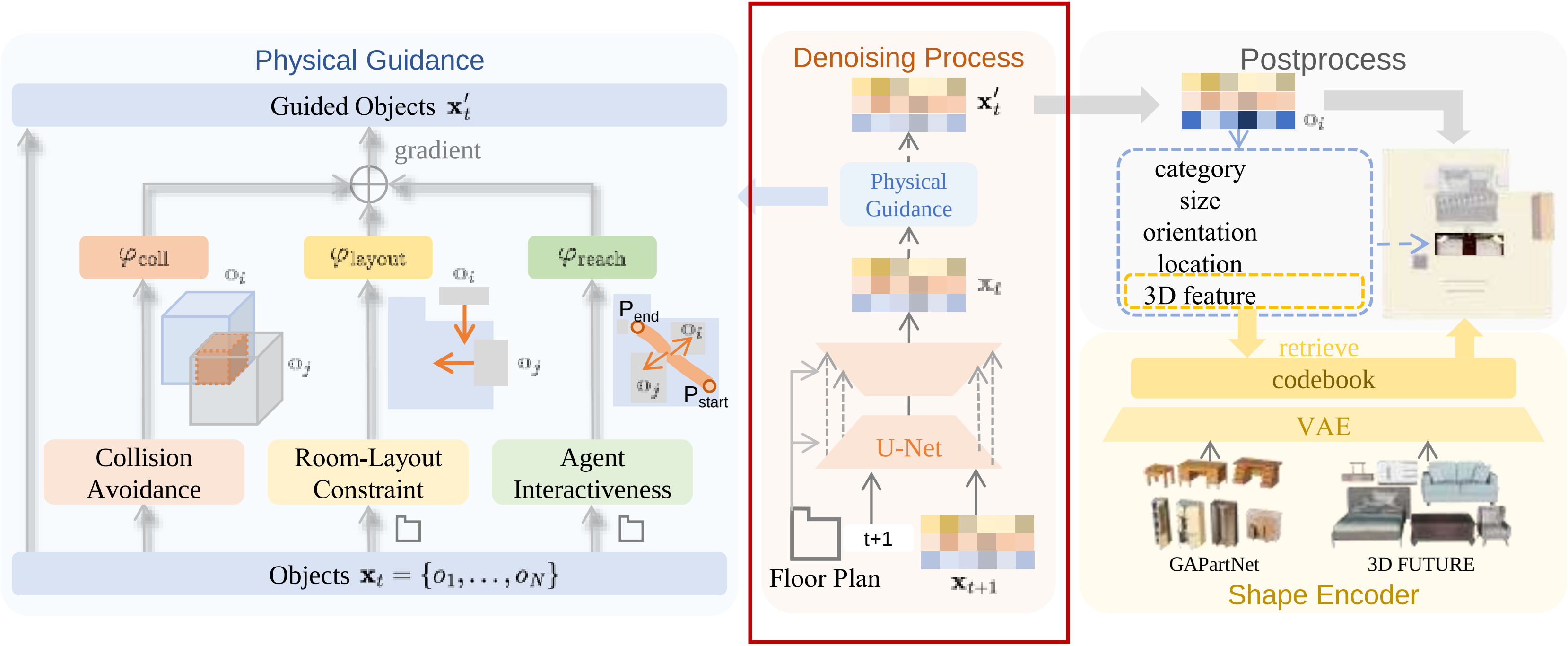


Fu et al., 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics, ICCV 2021

PhyScene

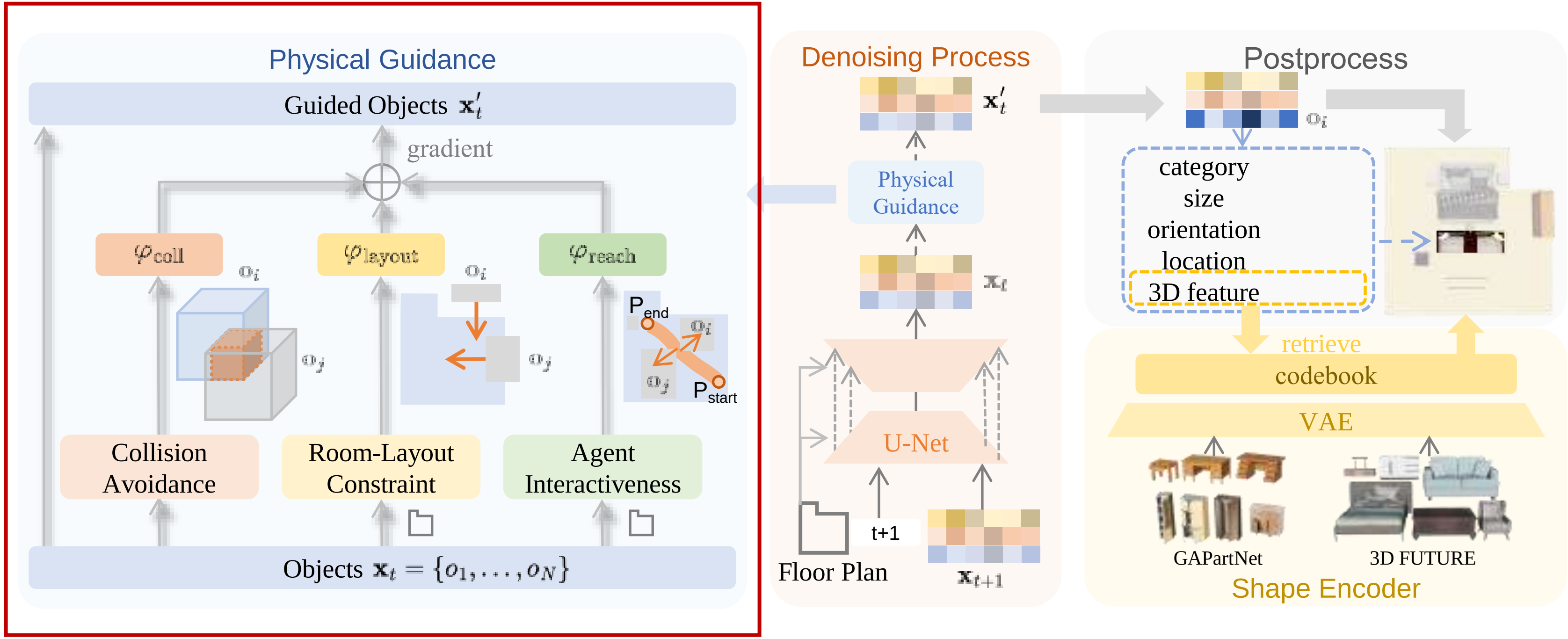
- A guided diffusion model for **physically interacble** scene synthesis with realistic layout and **interactable objects**.

Modeling



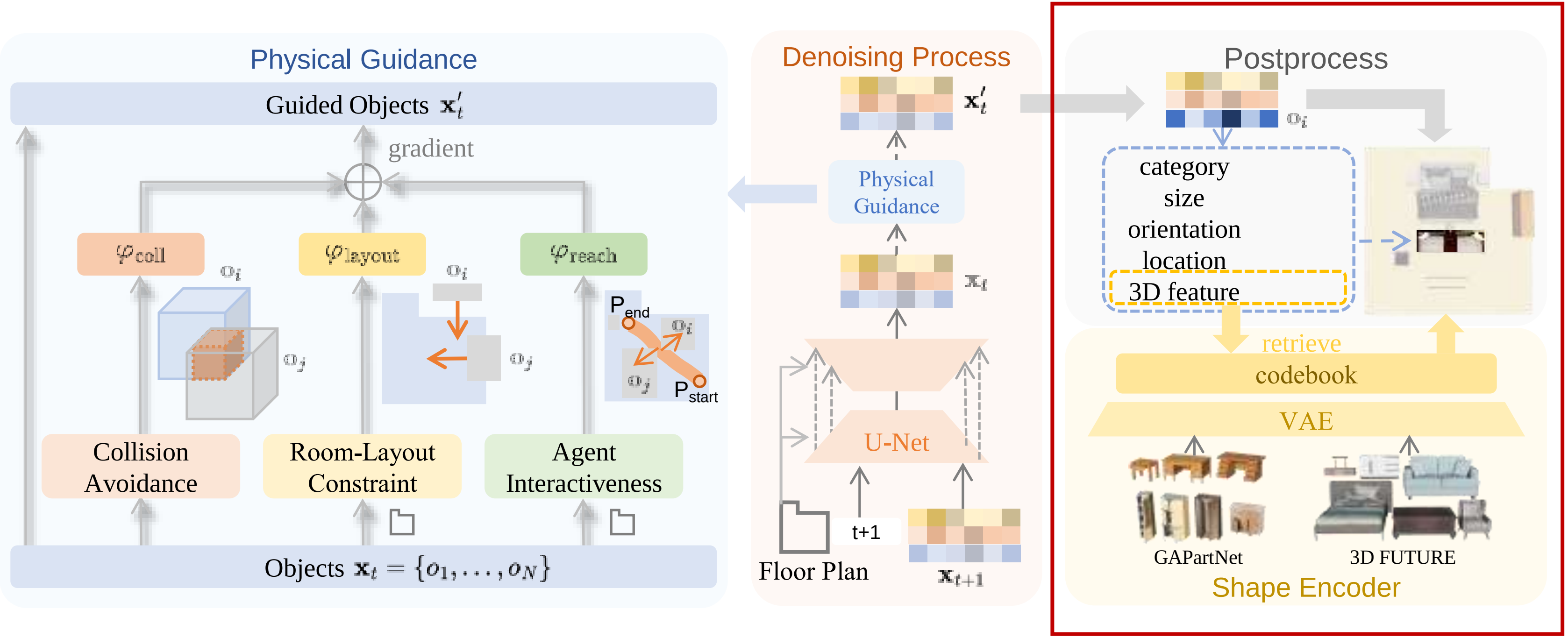
Diffusion model

Modeling



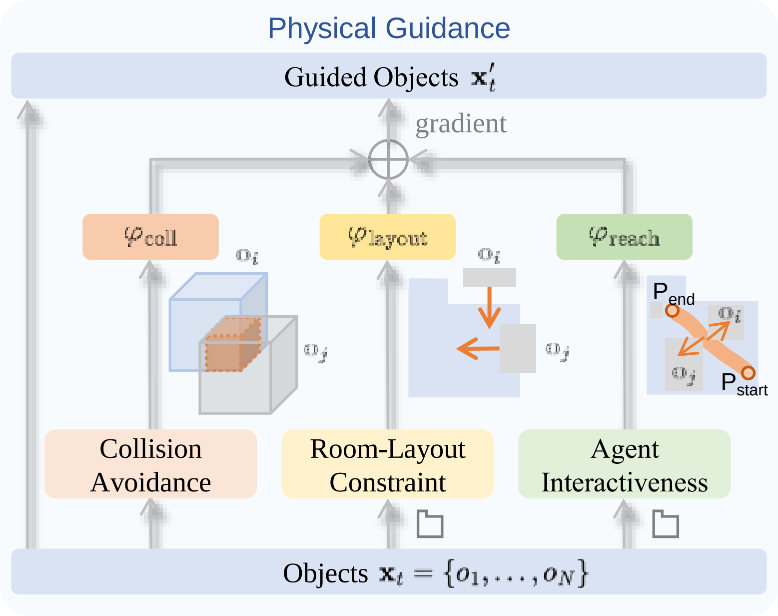
Physical guidance

Modeling

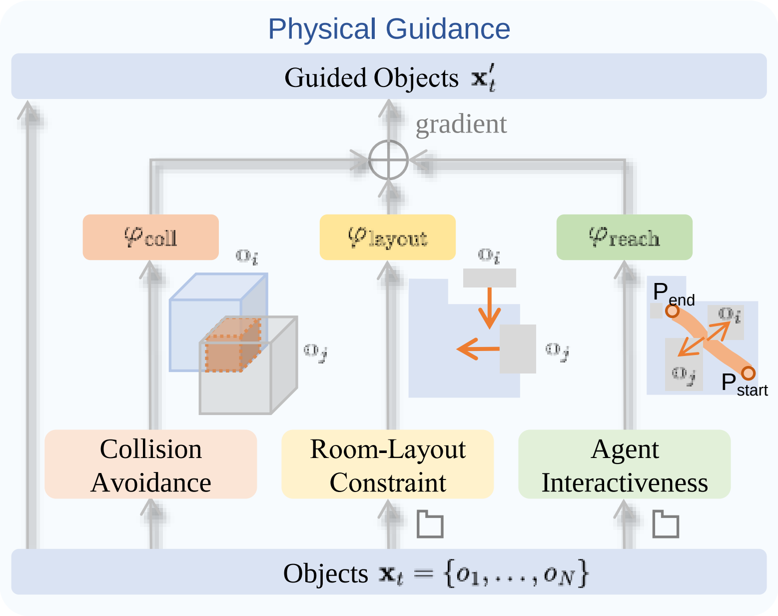


Articulated objects

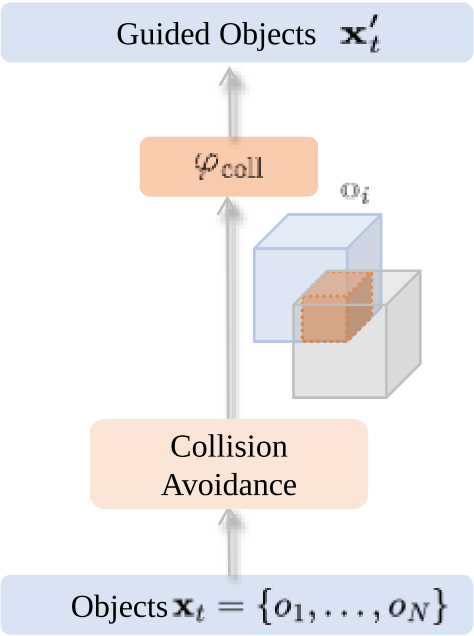
Physical guidance

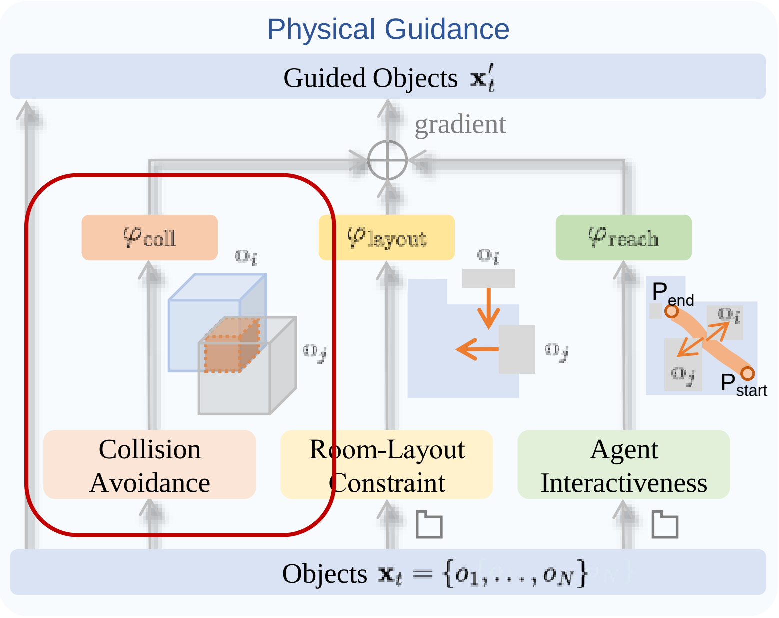


Physical guidance

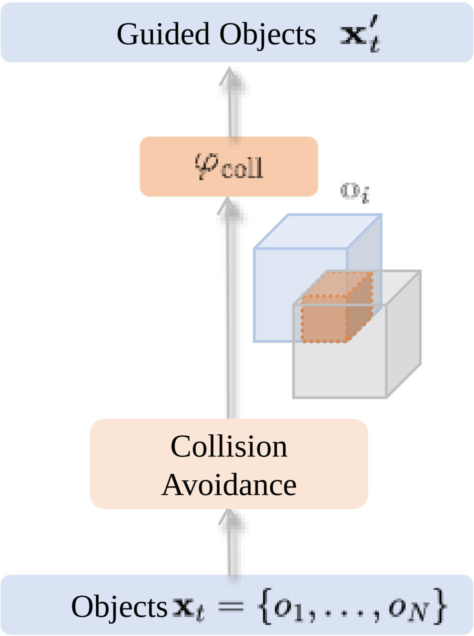


Physical guidance: collision avoidance

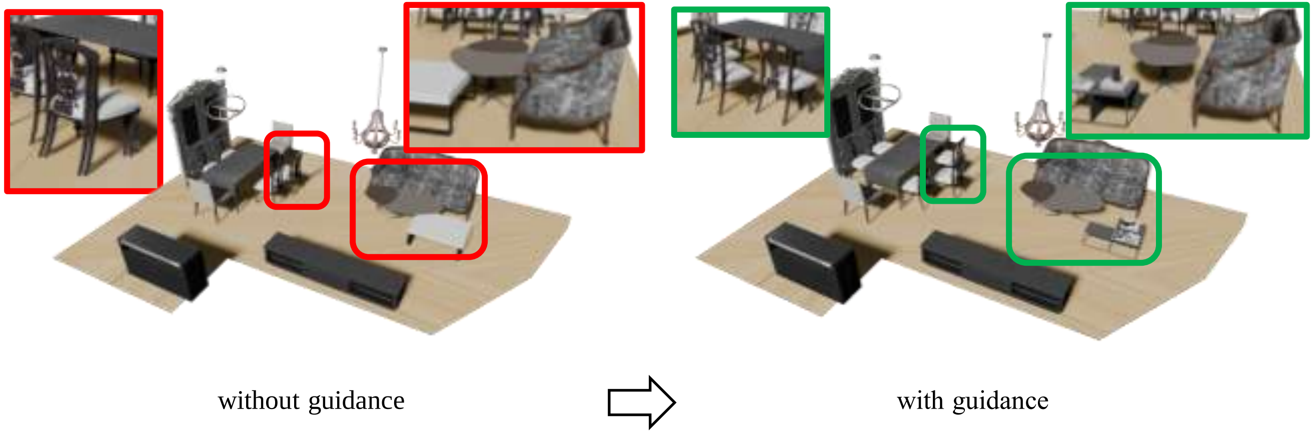




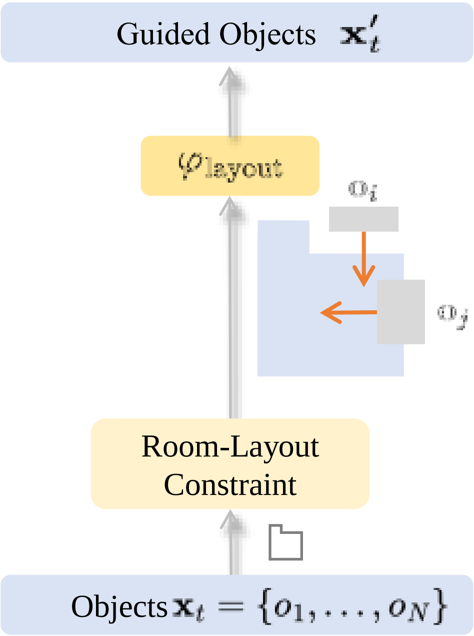
Physical guidance: collision avoidance

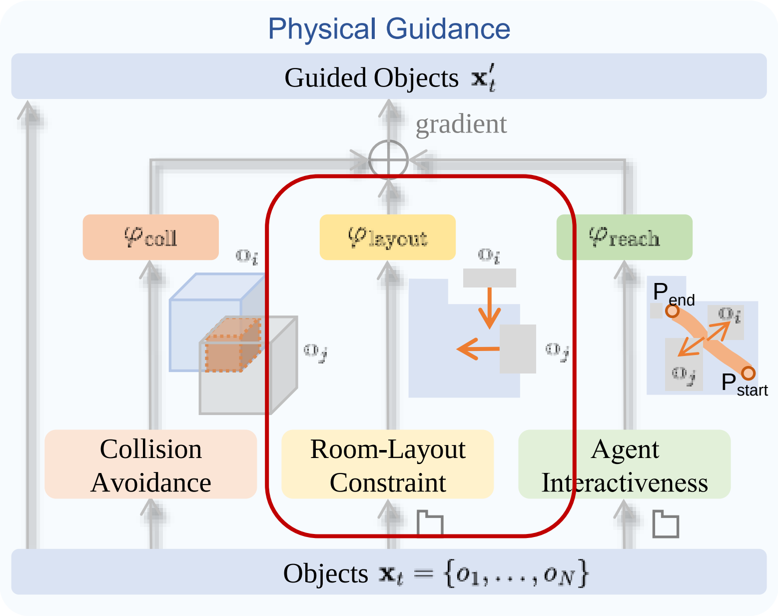


$$\varphi_{\text{coll}}(\mathbf{x}) = -\sum_{i,j,i\neq j} \mathbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j),$$

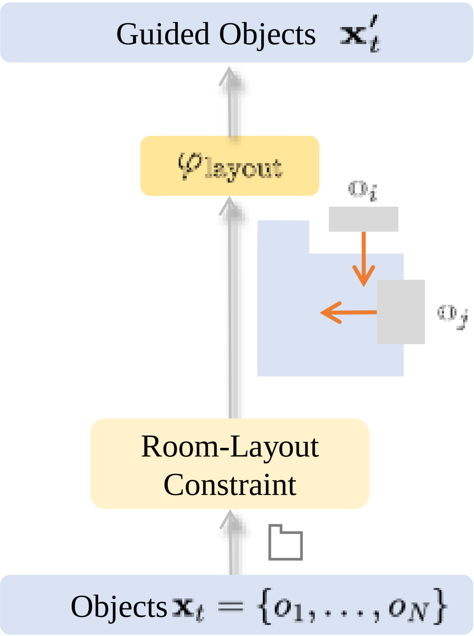


Physical guidance: room-layout constraint





Physical guidance: room-layout constraint

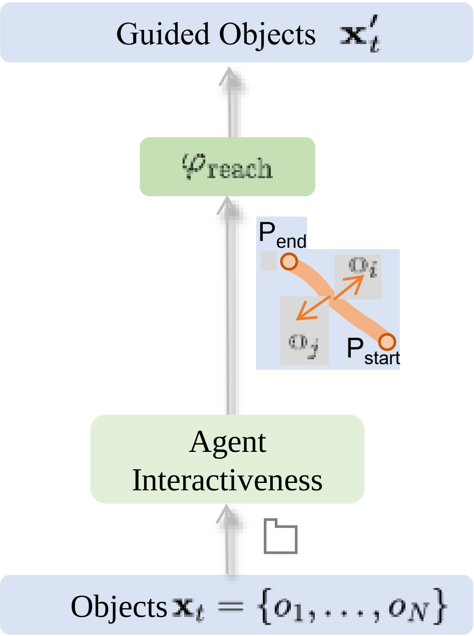


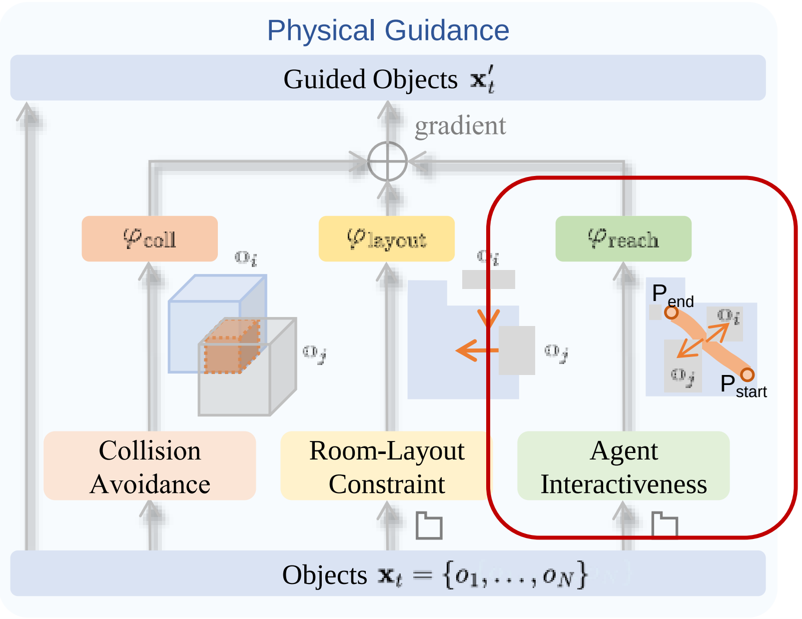
$$\varphi_{\text{layout}}(\mathbf{x}|\mathcal{F}) = -\sum_{i=1}^{N}\sum_{j=1}^{W} \mathbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j^{\text{wall}}).$$

without guidance

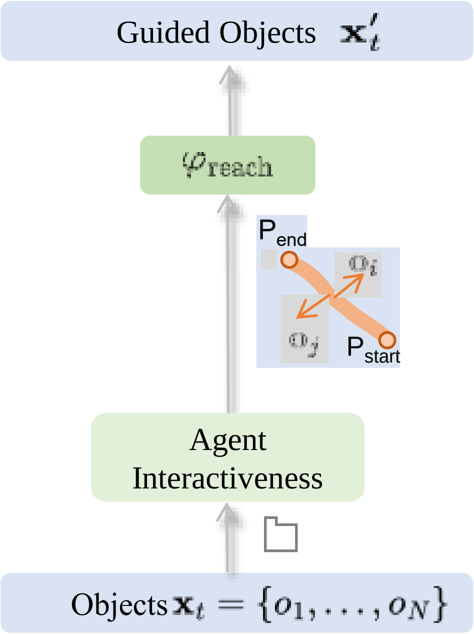
with guidance

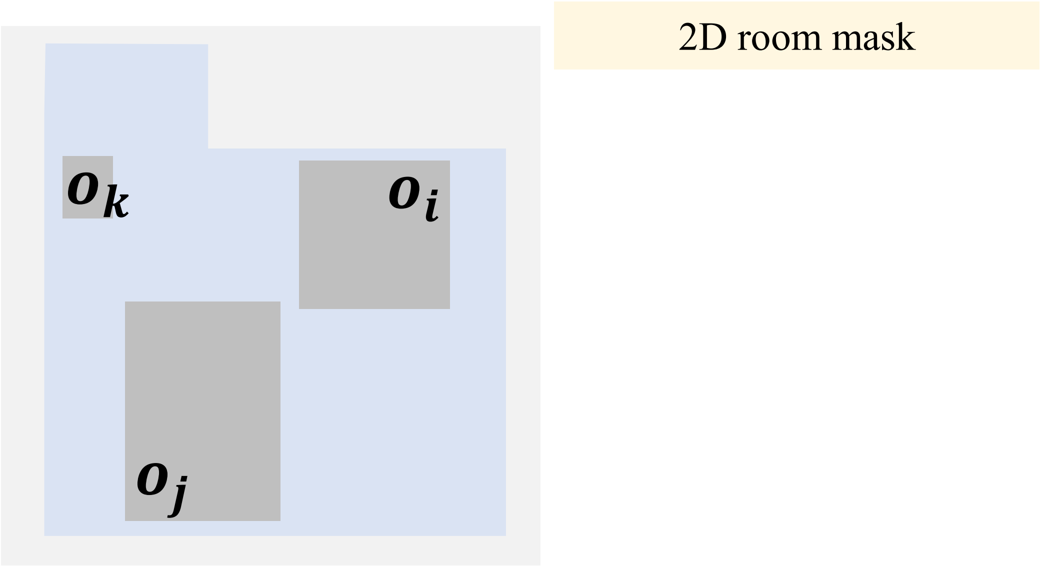
Physical guidance: reachability guidance



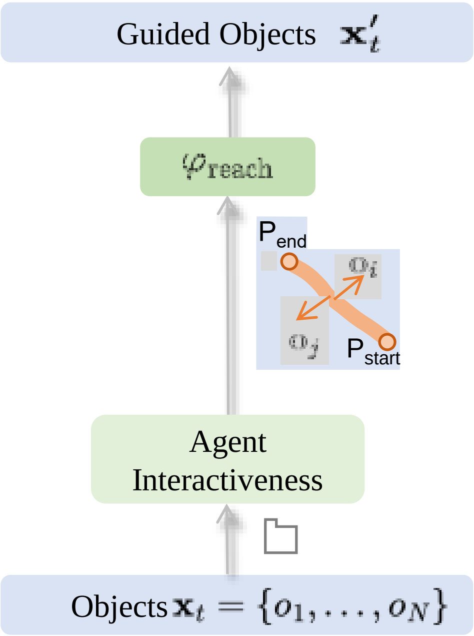


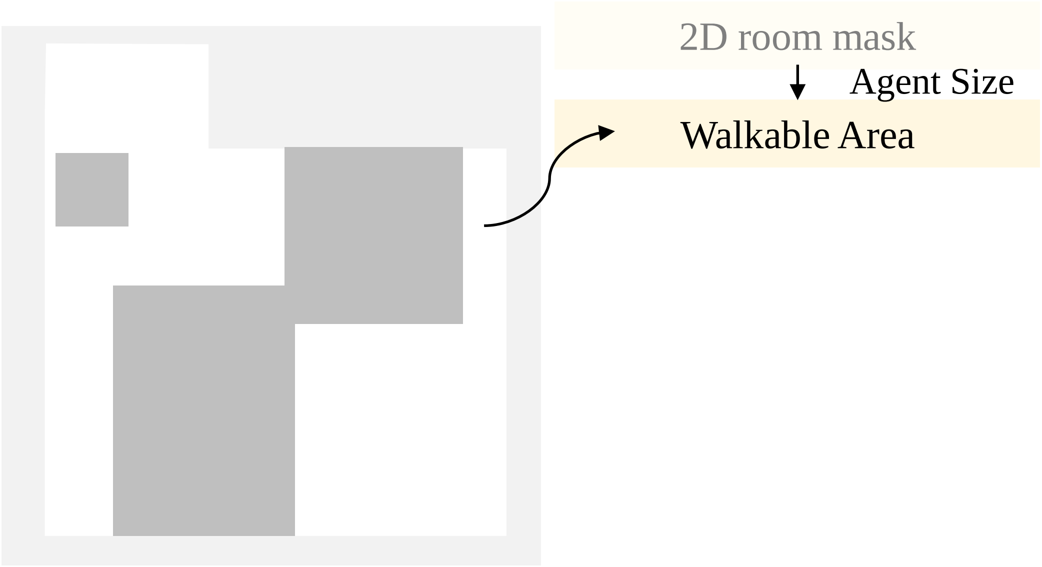
Physical guidance: reachability guidance



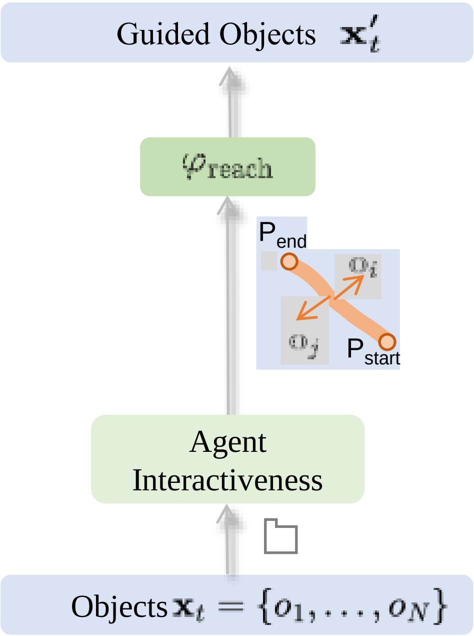


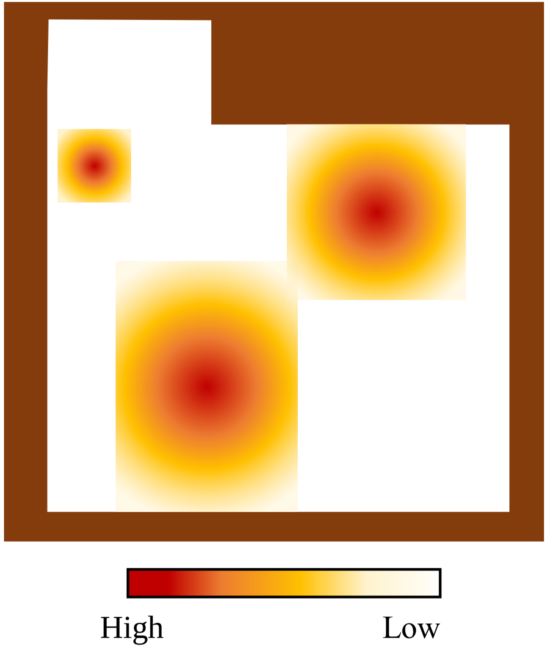
Physical guidance: reachability guidance





Physical guidance: reachability guidance





2D room mask

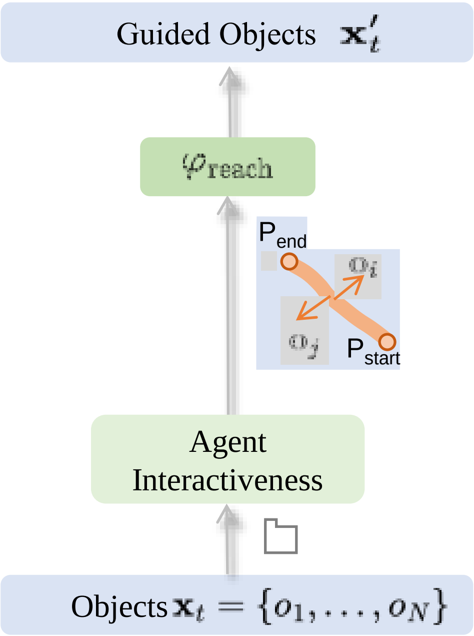
↓ Agent Size

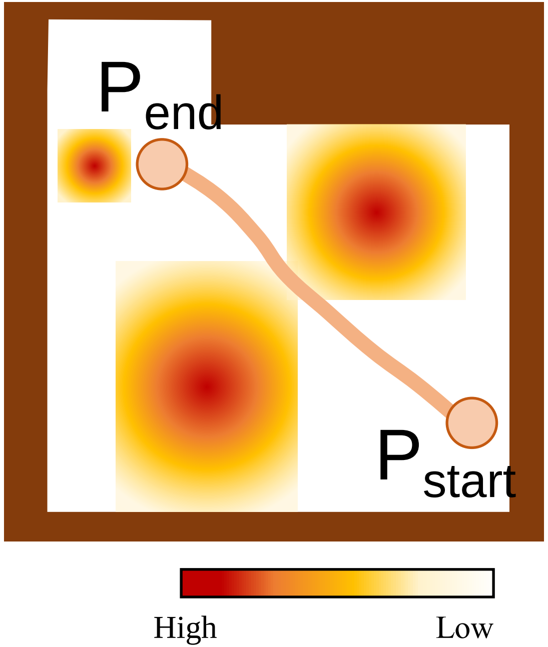
Walkable Area

↓ + Gaussian

Cost Map

Physical guidance: reachability guidance





2D room mask

↓ Agent Size

Walkable Area

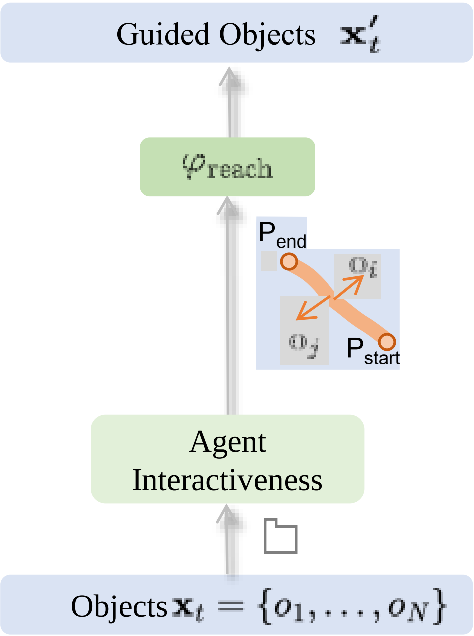
↓ + Gaussian

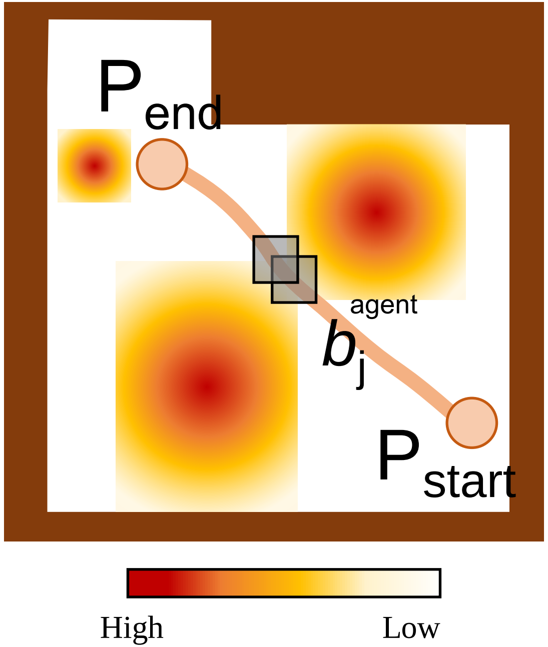
Cost Map

↓ + A* algorithm

Shortest Path

Physical guidance: reachability guidance





2D room mask

↓ Agent Size

Walkable Area

↓ + Gaussian

Cost Map

↓ + A* algorithm

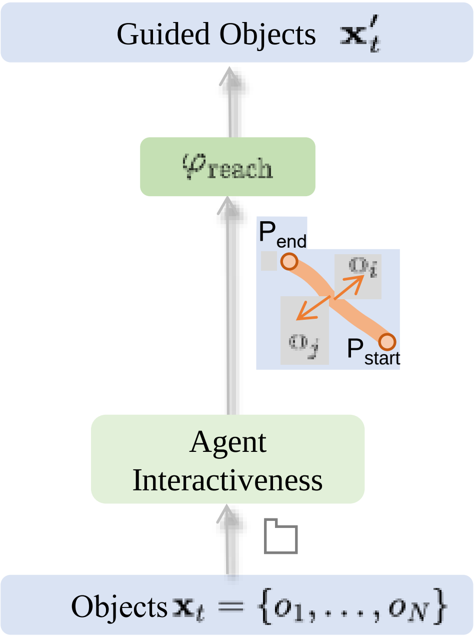
Shortest Path

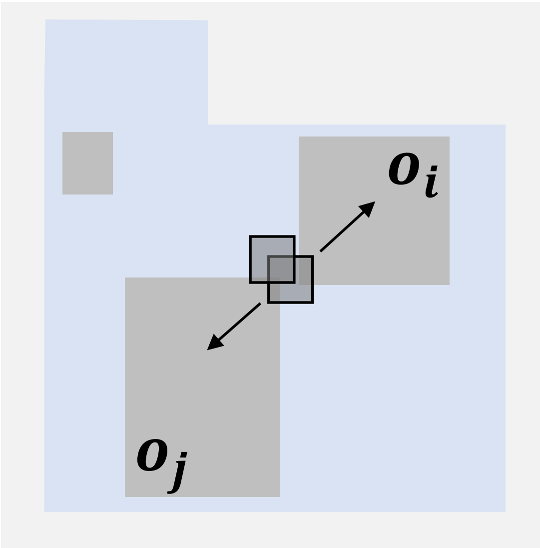
↓ Agent Size

Agent Bboxes on Path

Physical guidance: reachability guidance

$$\varphi_{\text{reach}}(\mathbf{x}|\mathcal{F}) = -\sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j^{\text{agent}}).$$





2D room mask

↓ Agent Size

Walkable Area

↓ + Gaussian

Cost Map

↓ + A* algorithm

Shortest Path

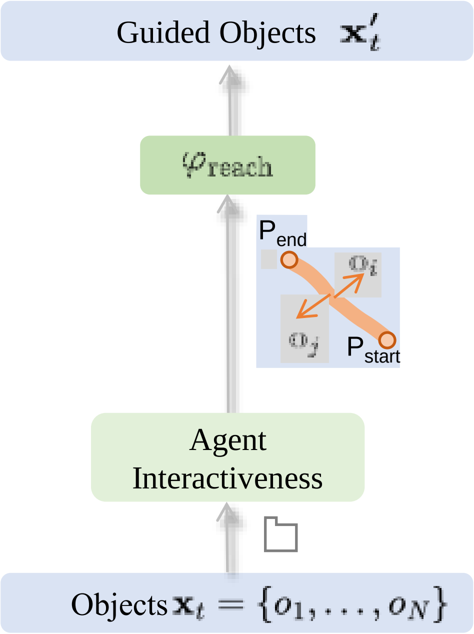
↓ Agent Size

Agent Bboxes on Path

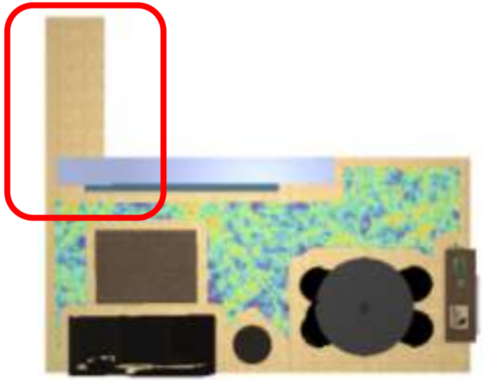
↓

Compute **IoU** between Objects and Agents

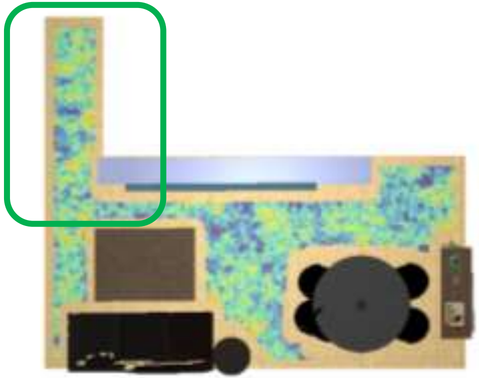
Physical guidance: reachability guidance



$$\varphi_{\text{reach}}(\mathbf{x}|\mathcal{F}) = -\sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{IoU}_{3D}(\boldsymbol{b}_i, \boldsymbol{b}_j^{\text{agent}}).$$

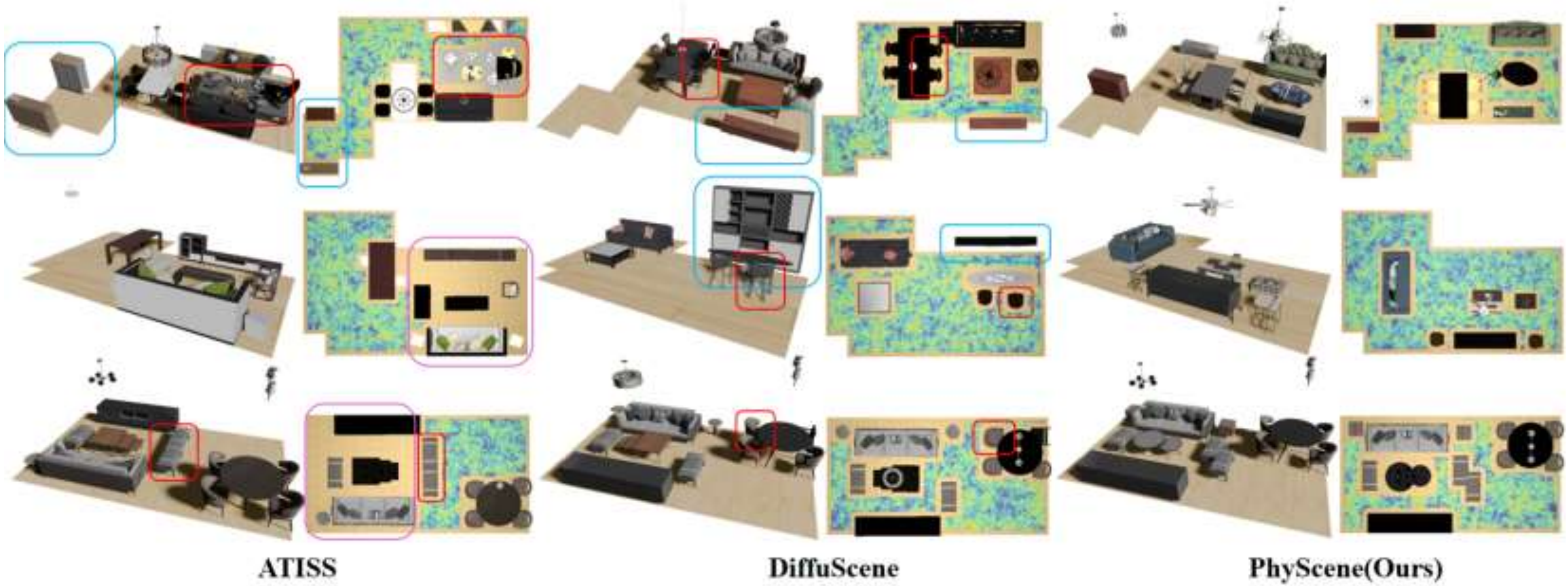


without guidance



with guidance

Comparisons

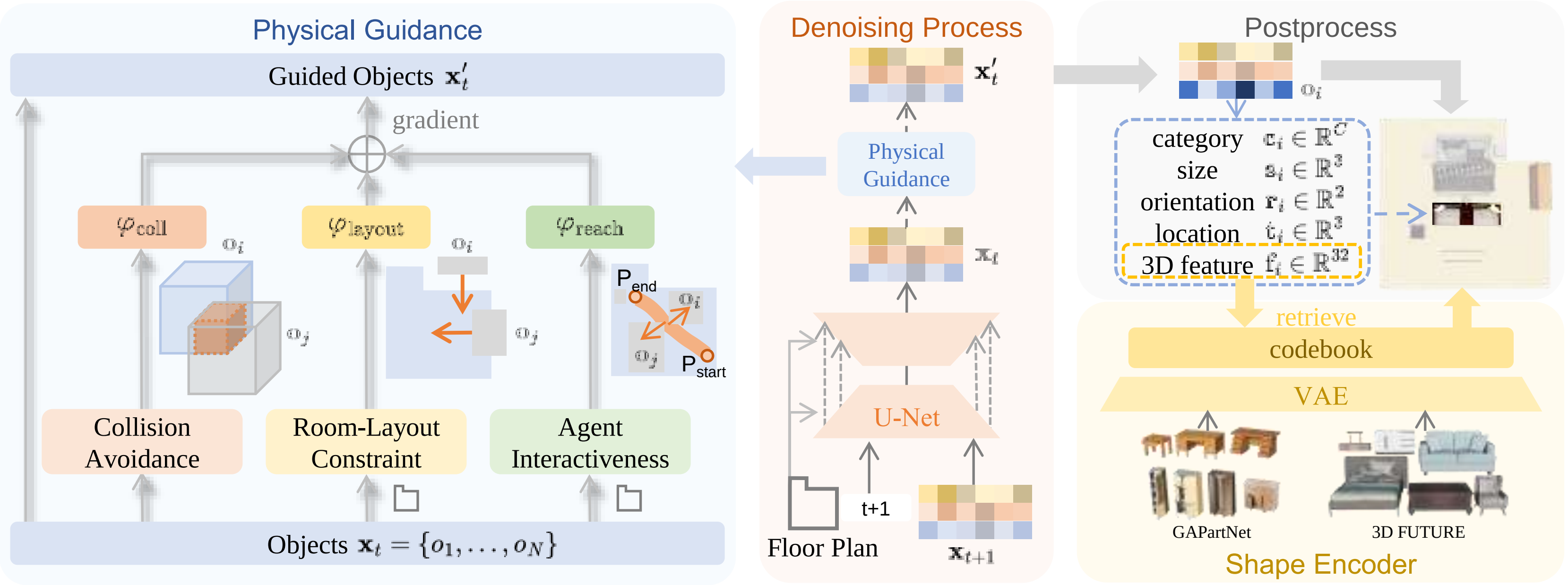


Collisions between objects Objects outside the floor plan Unreachable area to the embodied agent

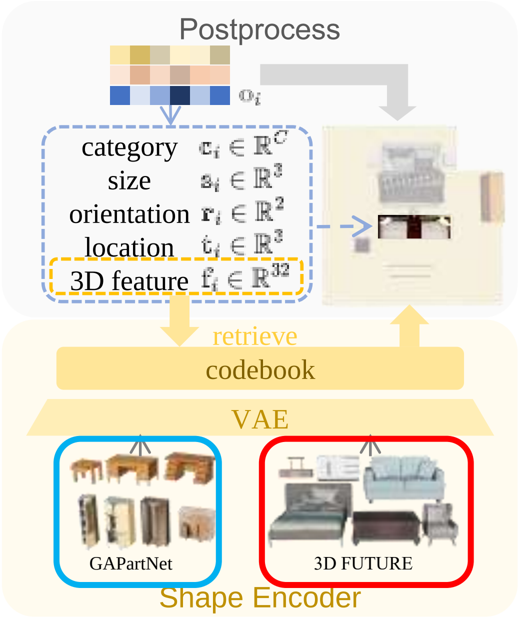
Comparisons

Room Type	Method	Perceptual Metrics				Physical Plausibility Metrics				
		FID ↓	KID ↓	SCA ↓	CKL ↓	Col_{obj} ↓	Col_{scene} ↓	R_{out} ↓	$R_{walkable}$ ↑	R_{reach} ↑
Bedroom	ATISS	30.19	0.0010	49.14	0.0028	0.248	0.46	0.286	0.839	0.736
	DiffuScene	**25.00**	**0.0004**	51.78	0.0031	0.228	0.43	0.272	0.827	0.755
	PhyScene (Ours)	25.52	0.0006	**50.10**	**0.0025**	**0.187**	**0.36**	**0.245**	**0.865**	**0.762**
Living Room	ATISS	45.66	0.0035	**51.64**	0.0016	0.316	0.85	**0.136**	0.814	**0.791**
	DiffuScene	**38.69**	**0.0012**	54.06	0.0017	0.198	0.69	0.238	0.790	0.756
	PhyScene (Ours)	43.33	0.0031	53.50	**0.0015**	**0.191**	**0.63**	0.219	**0.815**	0.771
Dining Room	ATISS	41.66	0.0039	64.57	0.0040	0.591	0.96	**0.132**	**0.874**	**0.848**
	DiffuScene	**38.31**	**0.0020**	60.19	**0.0013**	0.160	0.55	0.244	0.787	0.847
	PhyScene (Ours)	39.90	0.0026	**60.00**	**0.0013**	**0.151**	**0.53**	0.217	0.852	0.789

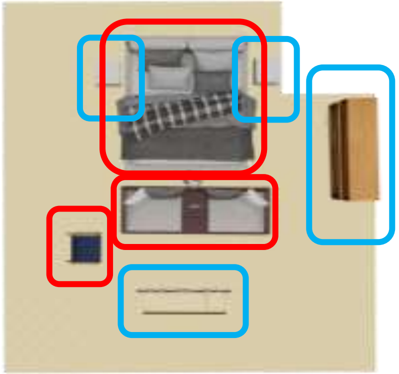
Embedding articulated objects



Embedding articulated objects

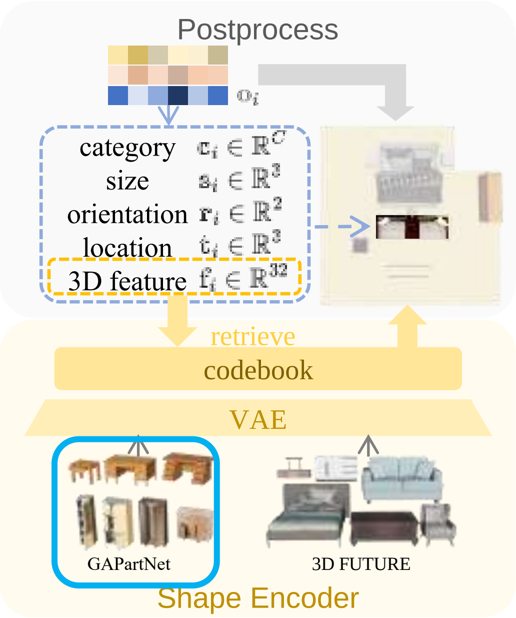


+Articulated Object

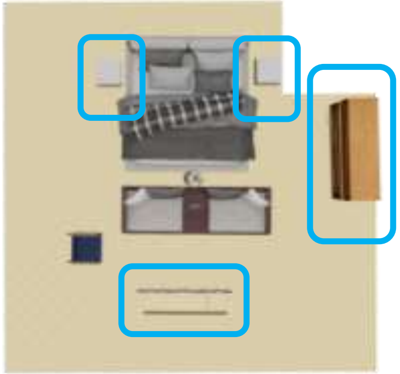


Embedding articulated objects

+ Guidance



+Articulated Object



Embedding articulated objects

通境

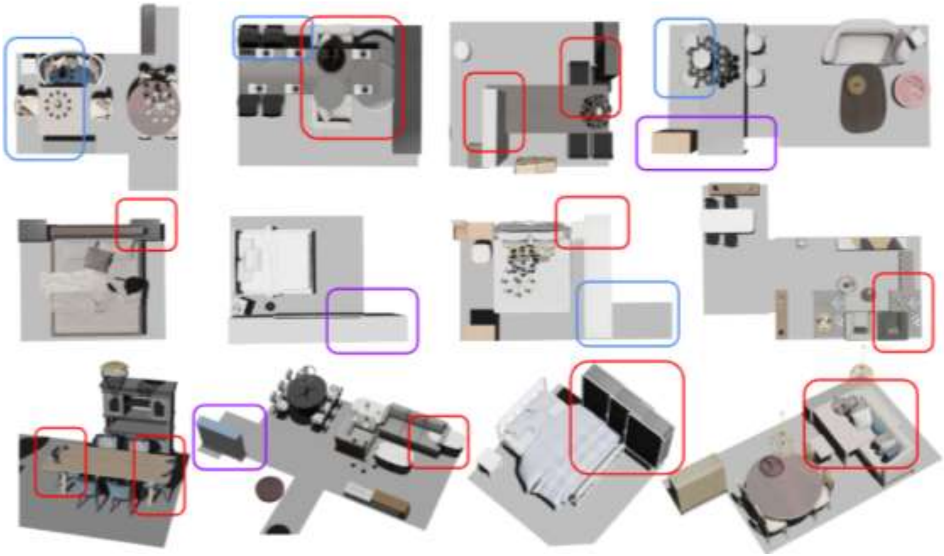
TongVerse

全球首个支持具身机器人物理交互的多场景室内训练靶场

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

Limitations

- Training data quality
  - Physically incorrect training scenes
- Conflicting guidance functions
  - Collision pushing objects apart
- Not enough scale / diversity
  - No small objects
  - Limited articulated objects
  - Three room types available
  - Limited scale (thousands)



Takeaways

Good:

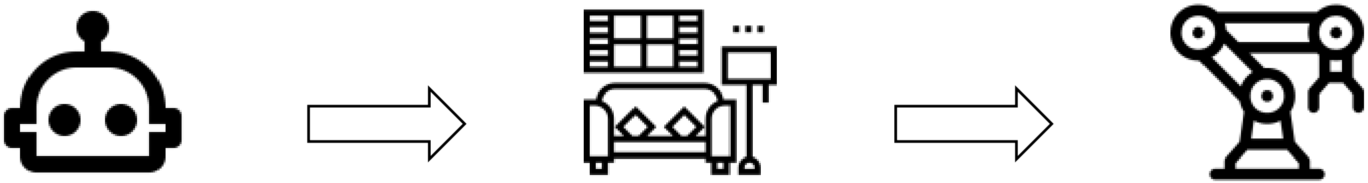
- We can optimize the scene generation process to make them physically plausible.
- No worries on fine details of 3D scenes, can put them into simulators and train agents.

Bad:

- **Post-optimization** that ensures **naturalness** and **realism** might be difficult.
- **Limited training data scale / quality / diversity** for data-driven approaches.

Better priors from 2D? from real world scenes? from language?

From the agent perspective



Perception | Grounding | Action

- Object geometry / Physics
- Need to capture 3D
- Aligning captured data
- Representation efficiency
- …

- Object attributes / properties
- Spatial relationships
- Affordance & functionality
- Auto-pipeline / Quality control
- …

- Scene constraints
- Hardware prerequisites
- Data capturing efficiency
- Embodiment gap
- …

Q3: How to scalably obtain interaction with the scenes following instructions?
What's next with these 3D Data?

Interaction with Scenes

Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance

CVPR 2024 Highlight

COME-Robot: Closed-Loop Open-Vocabulary Mobile Manipulation with GPT-4V

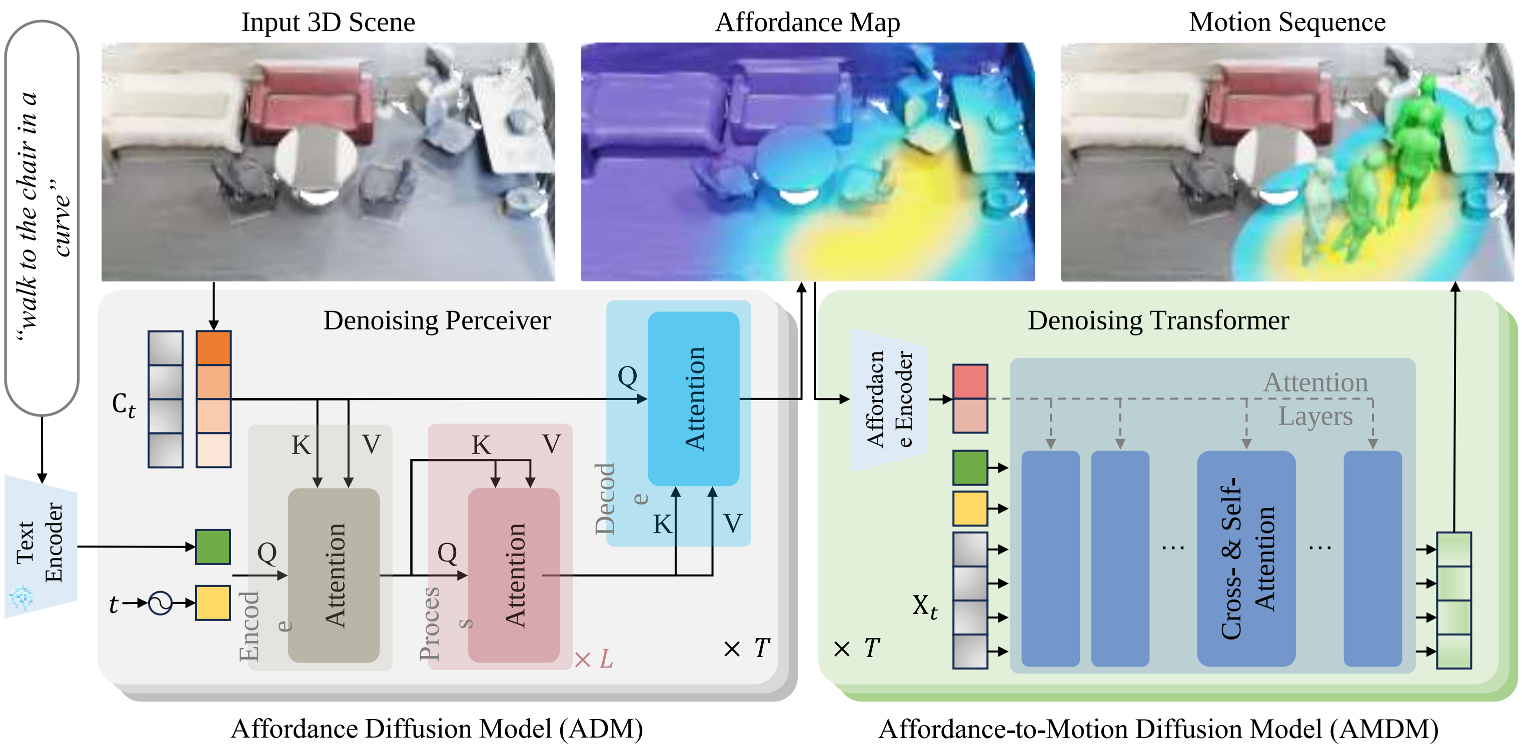
ArXiv 2024

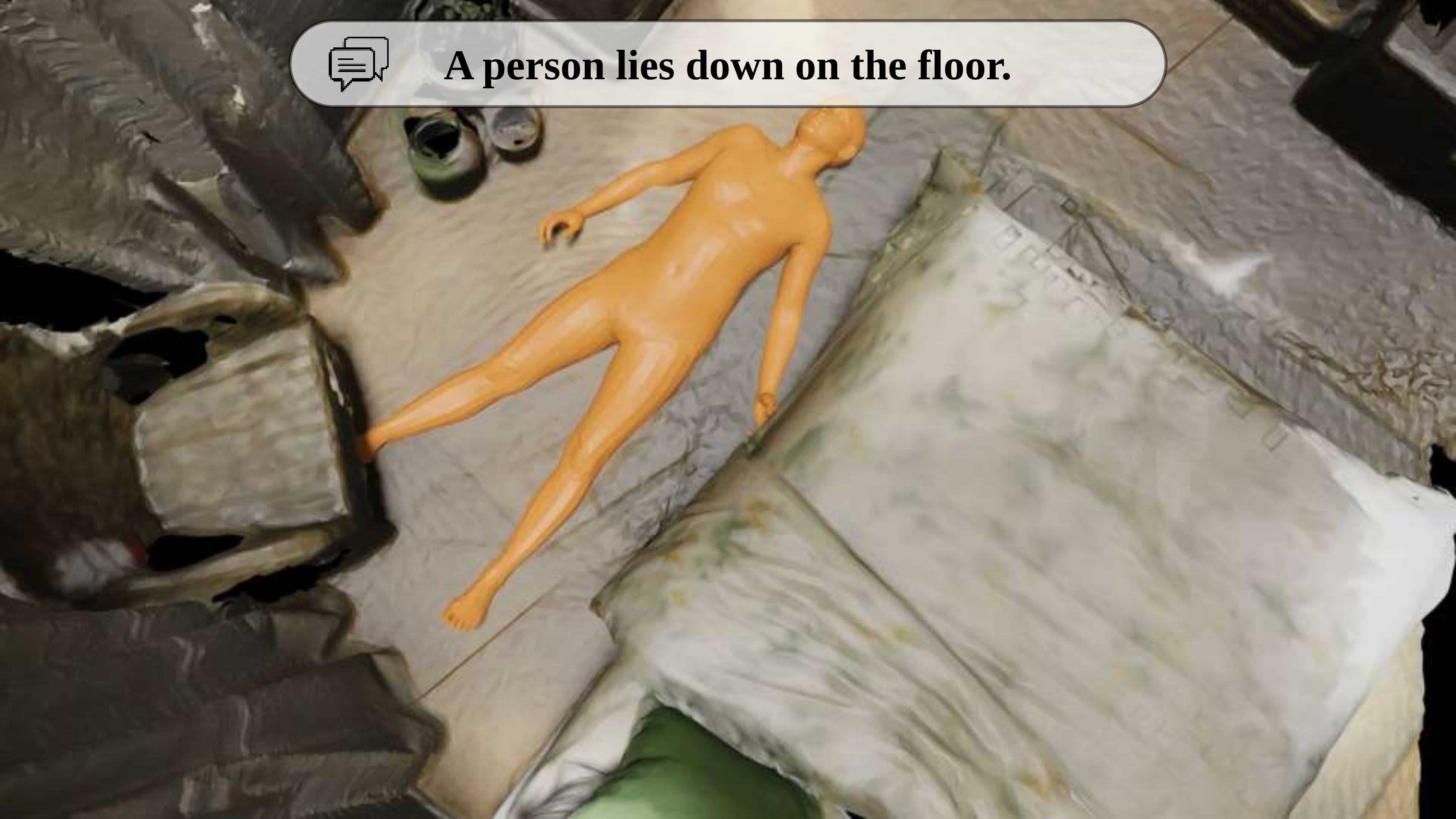
A person waves with his left hand.


Walk to the chair


The man walks to the chair in a curve.

AffordMotion




A person lies down on the floor.


I am hungry. Could you give me some food? And pass me a cup of juice.
15x
Active
Perception

COME-Robot

I am thirsty. Put a cup into the plate, I will pour water into it.
Vision Language Model
GPT-4V
Active Perception
Reason:
I need to perform a global exploration to identify the table.
Code:
global_object_map = robot.explore_global())
Global object map
Reason:
I find two tables. I will navigate to the first table to start a local search for cups and plate.
Code:
robot.navigate_to(table_0)
table_object_map = robot.explore_local()
Local object map
Situated Commonsense Reasoning
Reason:
The table contains two cups. I will observe both cups to confirm which is suitable.
Code:
robot.report_observation(cup_0)
robot.report_observation(cup_1)
Full cup_0
Empty cup_1
Reason:
The image shows that cup_0 is not empty and cup_1 is empty and suitable for the task.
Code:
robot.grasp(cup_1)
robot.report_observation('stay')
Try grasping
Recover From Failure
Reason:
The image shows the cup is not in the robot's gripper. I will attempt to grasp cup_1 again.
Code:
robot.grasp(cup_1)
robot.report_observation('stay')
Grasp failed
Grasp succeeded
Reason: This image shows the cup is in the robot's gripper. I will place the cup into the plate.
Code:
robot.place(plate_0)
robot.report_observation('stay')
Place succeeded

Overall

From the 3D scene perspective

- Scaling works, grounding might be solved reasonably well shortly
- Scaling scenes is difficult, no matter captured or generated
- Ensuring physics is a must for embodied AI but really challenging
- Need more priors from other modalities for generalization

…

From the agent perspective

- Intermediate representations to the rescue for generalization
- Effort-less data collection is critical, either automated or human shadowed

…

More to come at BIGAI

Project Page
https://scene-verse.github.io/

Project Page
https://physcene.github.io/

Project Page
https://afford-motion.github.io/

Project Page
https://come-robot.github.io/

Thank you!