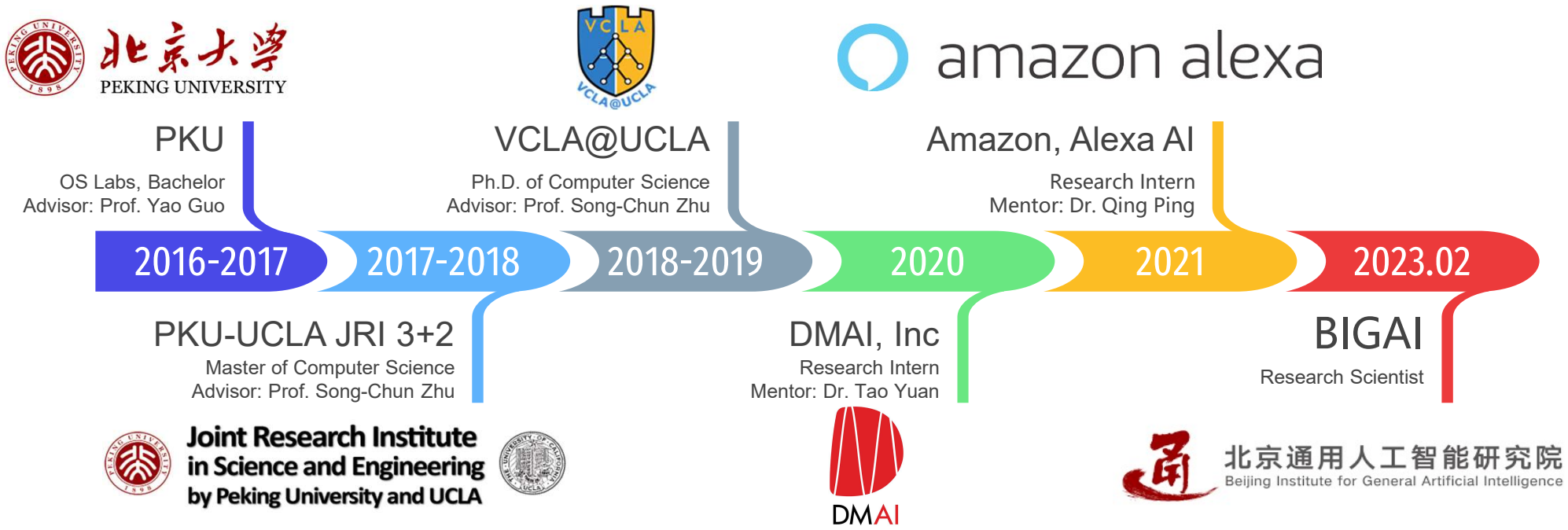
北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

GVL
General Vision Lab

Building Interactable 3D Scenes for Embodied AI

Baoxiong Jia
General Vision Lab, BIGAI

About me


北京大学
PEKING UNIVERSITY
VCLA
VCLA@UCLA
amazon alexa
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
Joint Research Institute
in Science and Engineering
by Peking University and UCLA
DMAI
北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

What we (I) expected ☺

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**

Walk, Run, Crawl, RL Fun | Boston Dynamics | Atlas, 2025
https://www.youtube.com/watch?v=I44_zbEwz_w

Interaction with scenes in daily life

Various object attributes and **diverse** scene configurations

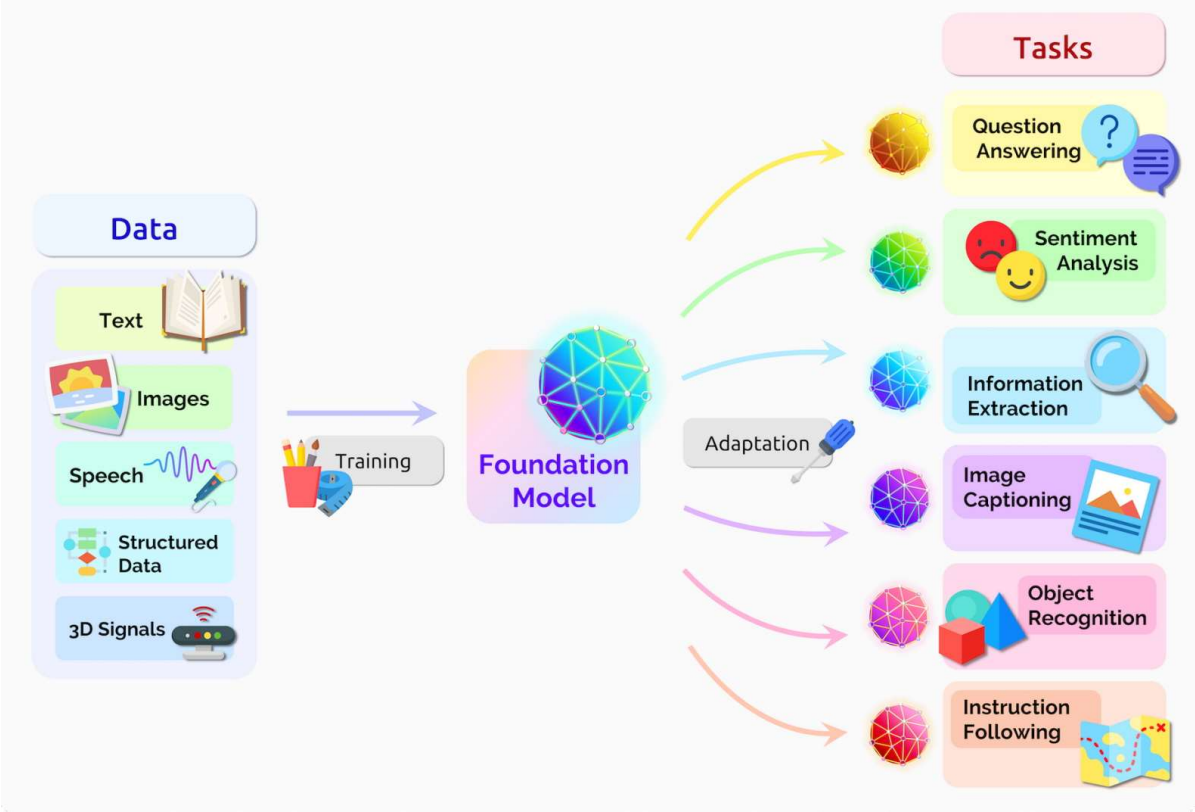
Long-horizon interaction with scenes

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

What we learned previously

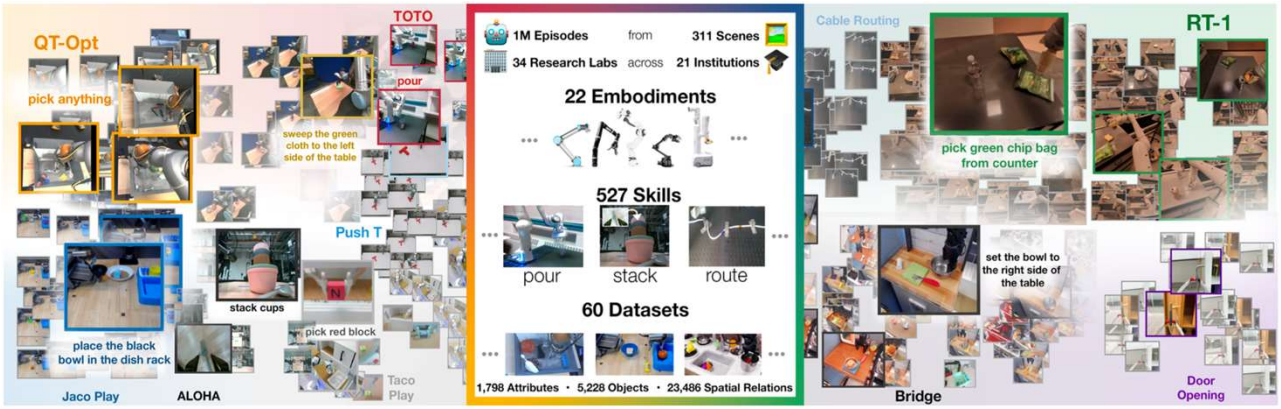
Data Data Data !!!

- 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/

Data for robotics?



Open-X-Embodiment (O'Neill et al., 2024)

Bridge Data V2 (Walke et al., 2023)

AgiBot World Colosseo (AgiBot, 2025)

Droid (Khazatsky et al., 2024)

Training Generalist Policies

- Pre-trained with large-scale manipulation data
- Leveraging large-scale pre-trained VLMs
- Starting to show generalizability on complex daily life tasks

How to close the gap between generalist and scene specific tasks?

Adapting to Your Specific Scene

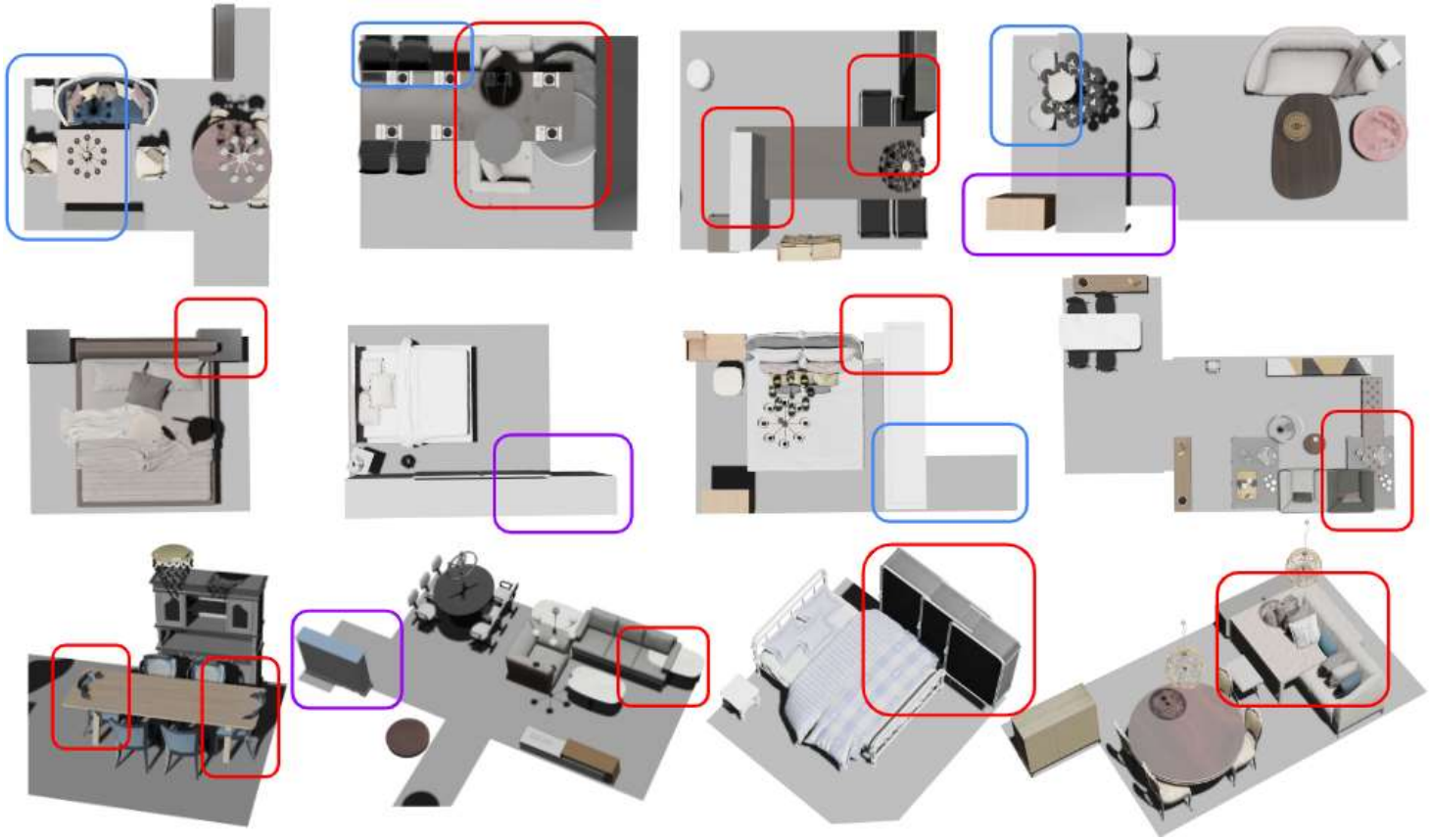
- Can only afford few-shot demonstrations
- Sensitive to capturing modalities and viewpoints
- Rolling out “almost” successful trajectories but hard to improve

Scalable Generation of Synthetic Scenes

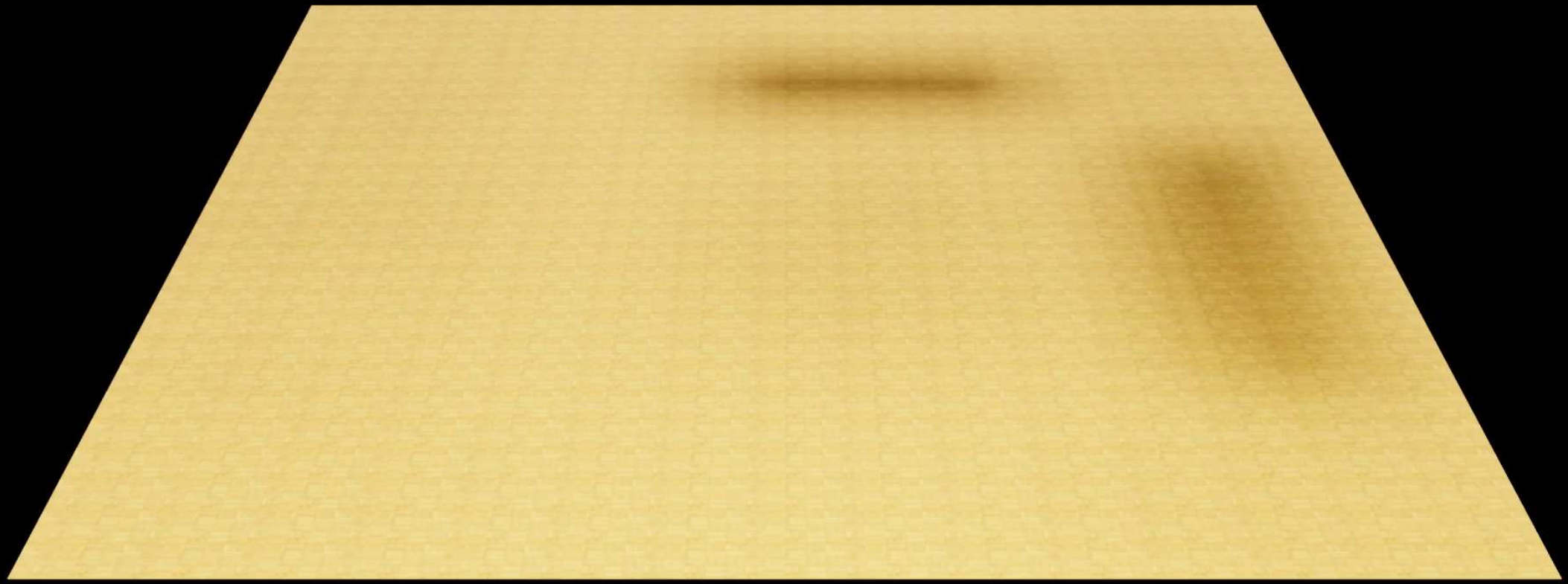
PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI

CVPR 2024

Synthetic Scenes to the Rescue?

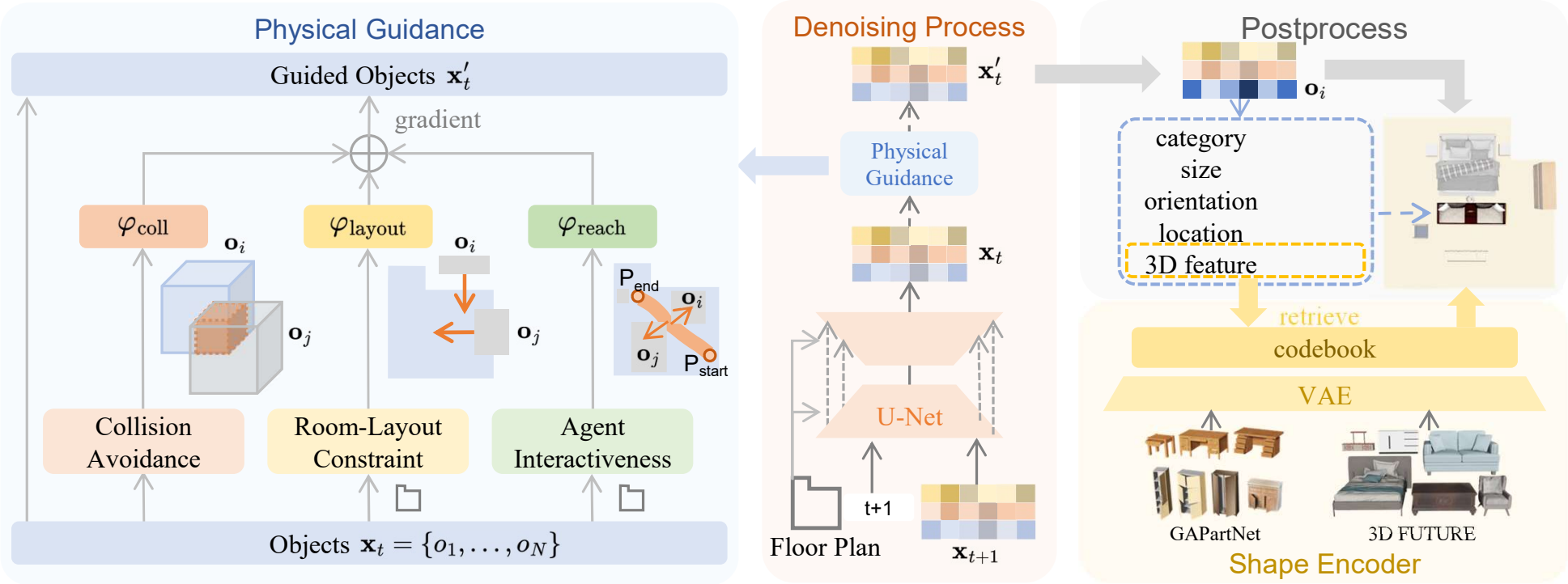


Fu et al., 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics, ICCV 2021



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

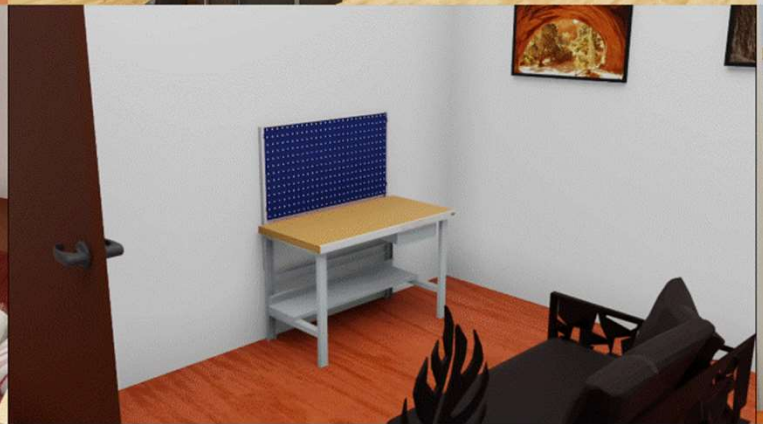
PhyScene


Physical Guidance
Guided Objects $\mathbf{x}'_t$
gradient
φ_{coll}
φ_{layout}
φ_{reach}
$\mathbf{o}_i$
$\mathbf{o}_j$
P_{end}
P_{start}
Collision Avoidance
Room-Layout Constraint
Agent Interactiveness
Objects $\mathbf{x}_t = \{o_1, \ldots, o_N\}$
Denoising Process
$\mathbf{x}'_t$
Physical Guidance
$\mathbf{x}_t$
U-Net
t+1
Floor Plan
$\mathbf{x}_{t+1}$
Postprocess
$\mathbf{o}_i$
category
size
orientation
location
3D feature
retrieve
codebook
VAE
GAPartNet
3D FUTURE
Shape Encoder





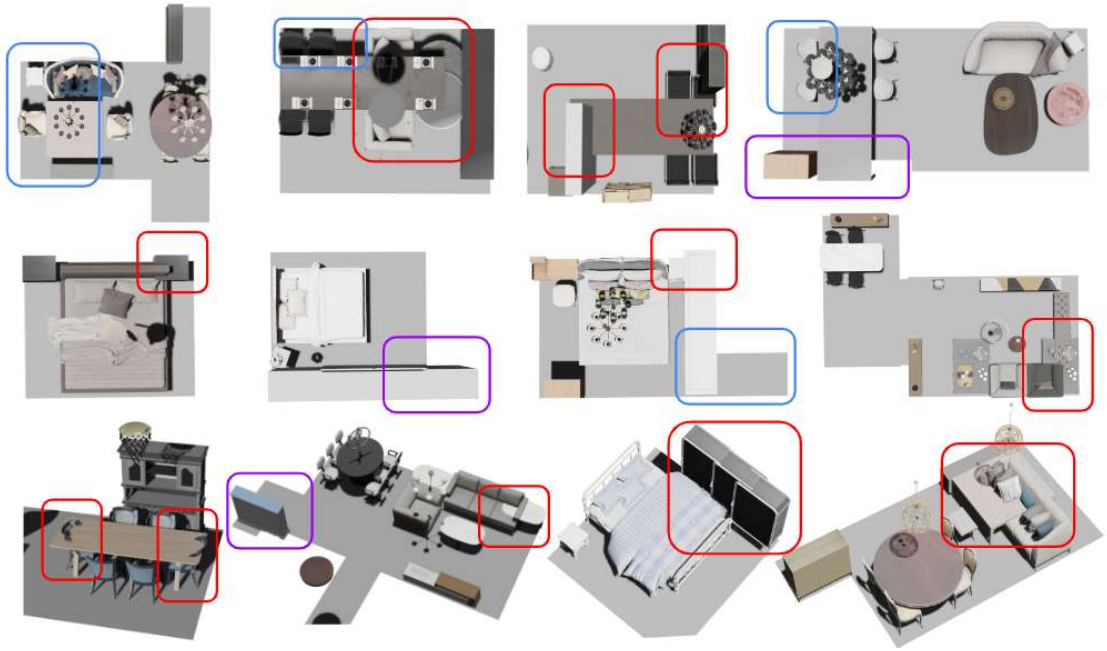



Limitations

Not enough scale / diversity

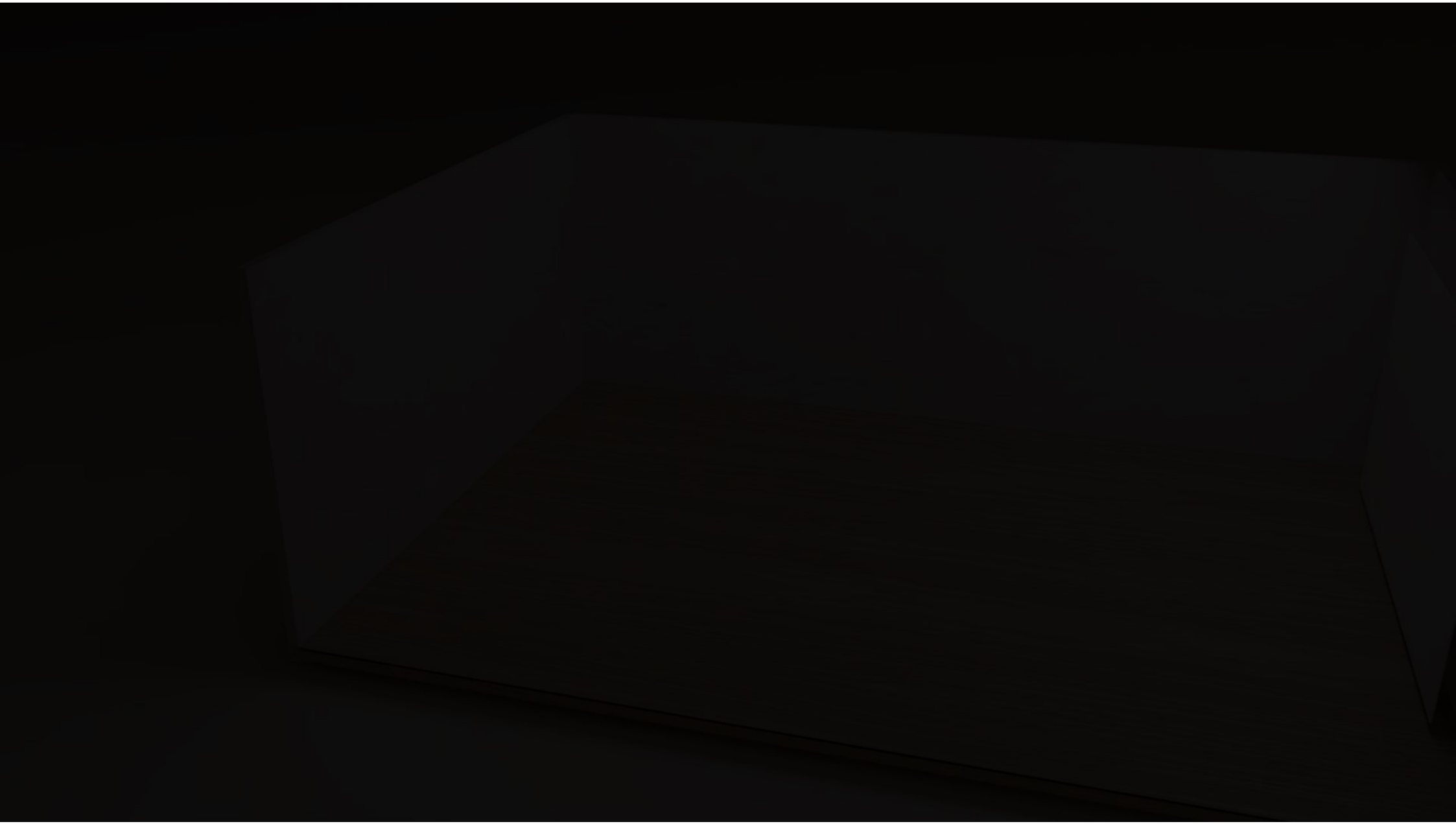
- No small objects
- Limited articulated objects
- Three room types available
- Limited scale (thousands)



Bringing Real Scenes into Simulation

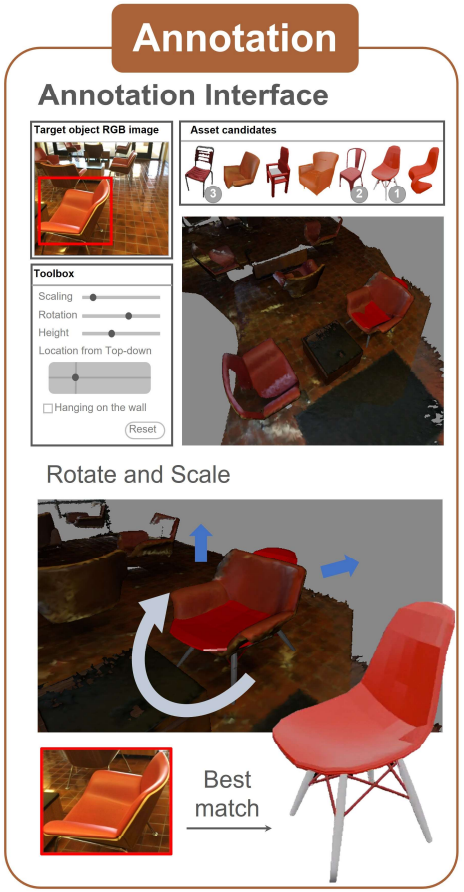
MetaScenes: Towards Automated Replica Creation for Real-world 3D Scans

CVPR 2025

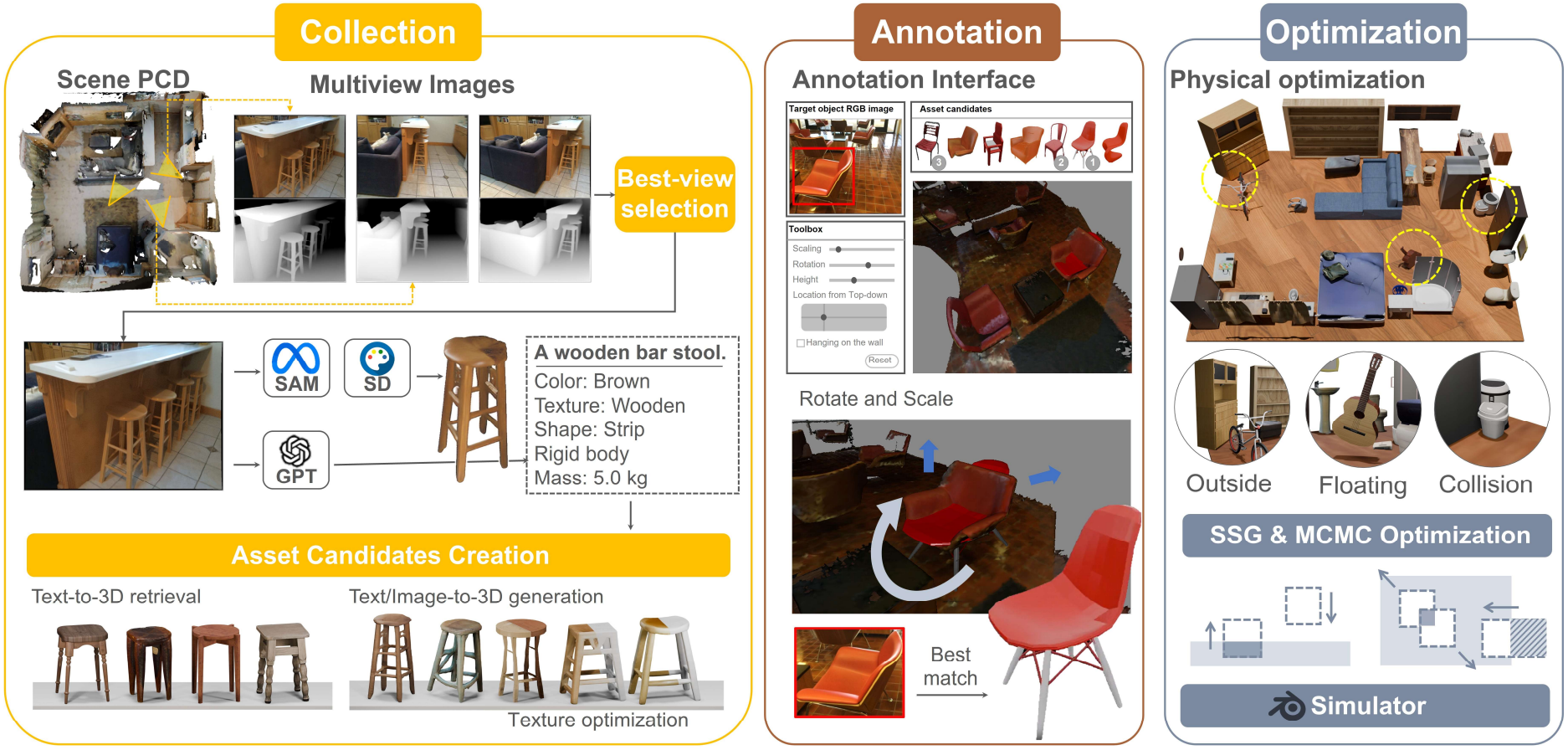


MetaScenes creation

MetaScenes creation



MetaScenes creation



MetaScenes for EAI

Table 5. **Cross-domain embodied navigation.** **METASCENES** improves generalization in unseen real scenes.

Benchmark	Data Source	SR(%)↑	EL↓	Curvature↓	SEL↑	SPL↑
In-domain Scenes	ProcTHOR [13]	52.43	25.34	0.38	50.00	43.81
	METASCENES	58.00	23.40	0.17	55.00	51.39
	Both	**59.07**	**22.78**	**0.21**	**55.94**	**52.28**
Heldout Scenes	ProcTHOR [13]	51.21	25.73	0.33	48.43	43.82
	METASCENES	**52.64**	**25.57**	**0.14**	**49.62**	**45.55**
	Both	51.36	25.58	0.22	48.33	44.78
Heldout Domains	ProcTHOR [13]	45.33	28.56	0.38	42.90	37.58
	METASCENES	**50.67**	**26.56**	**0.25**	**47.78**	**44.33**
	Both	46.67	26.95	0.27	43.43	41.51

Table A4. Comparison on VLN experiments with HSSD

Benchmark	Data Source	SR(%)↑	EL↓	Curvature↓	SEL↑	SPL↑
10 scenes from Replica CAD	HSSD	27.00	33.77	0.39	26.77	23.32
	METASCENES	32.00	33.71	0.46	31.56	26.91

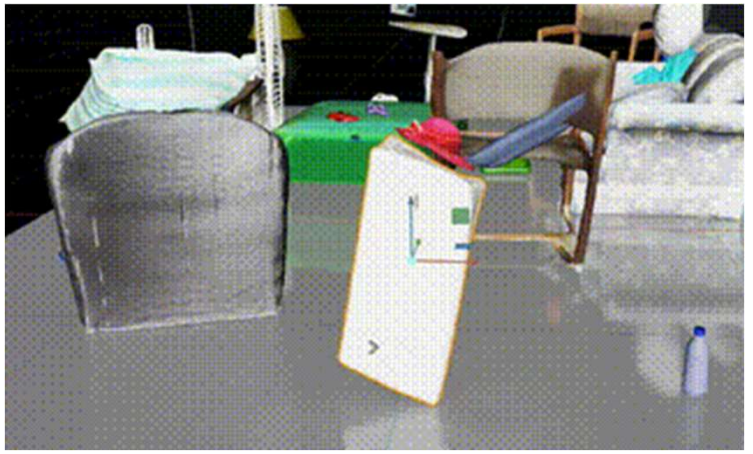
Discussion

Physical Plausibility

- Reconstructed / Generated objects
- Precise locations and physics
- Require additional manual post-optimization

Interactability

- Missing articulated objects
- Largely depending on available asset libraries
- Currently only for navigation, and potentially for pick & place

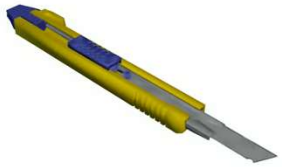


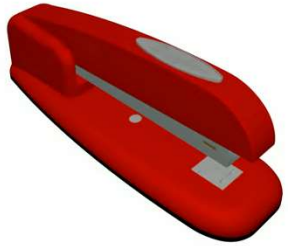
Reconstruction of Interactable Objects

Building Interactable Replicas of Complex Articulated Objects via Gaussian Splatting

ICLR 2025

Articulated objects

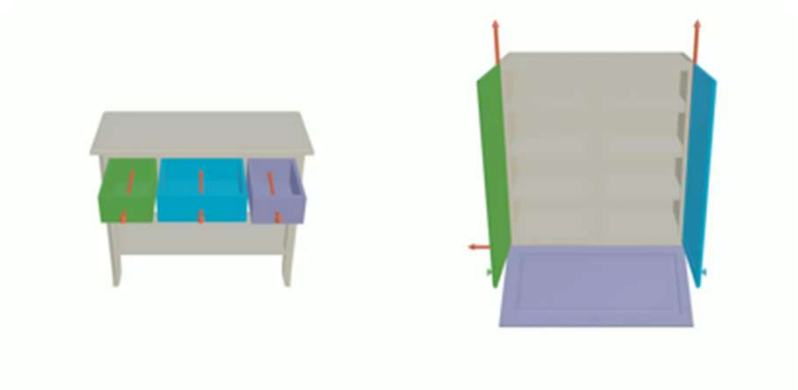




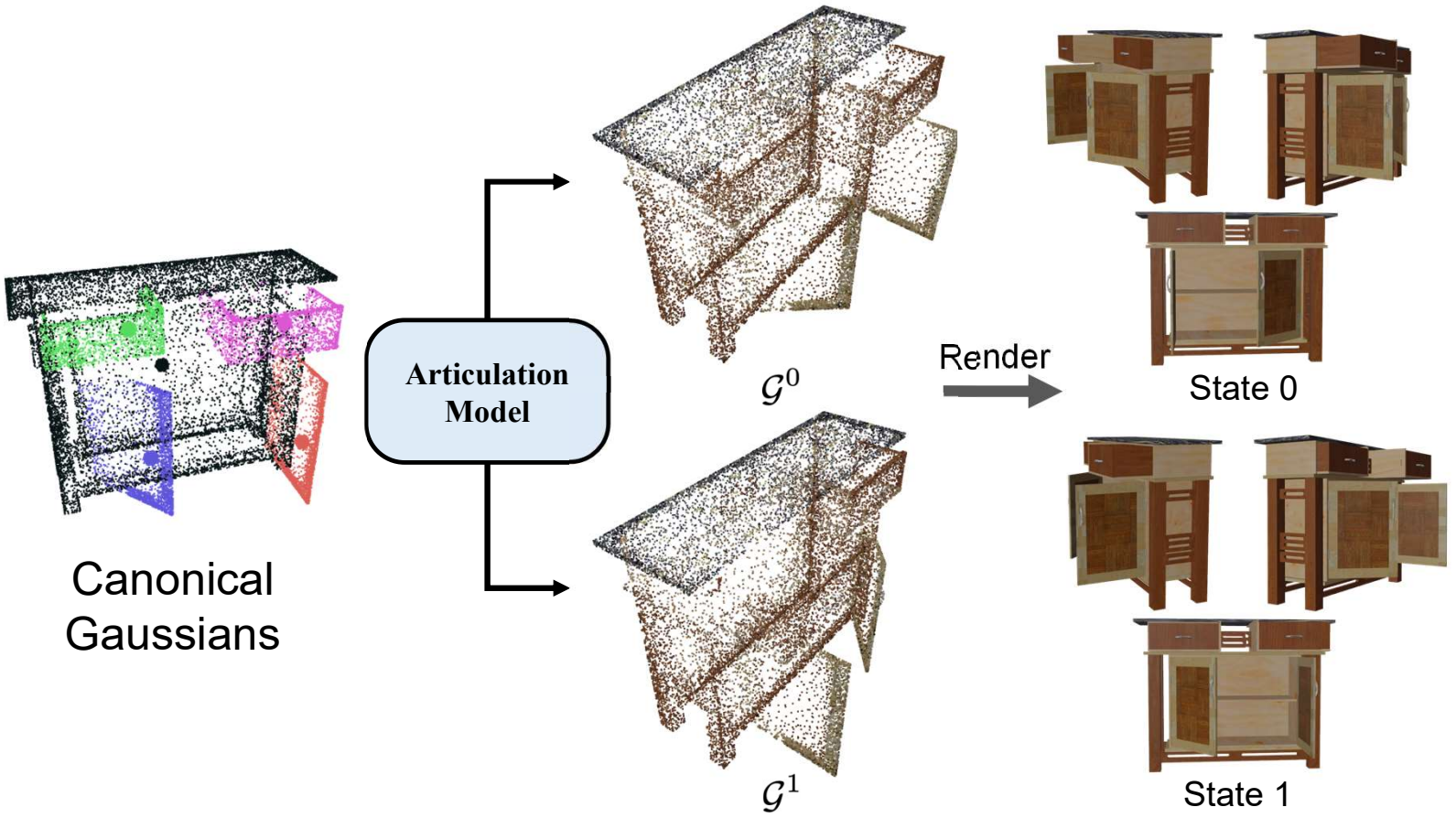
Articulated object reconstruction







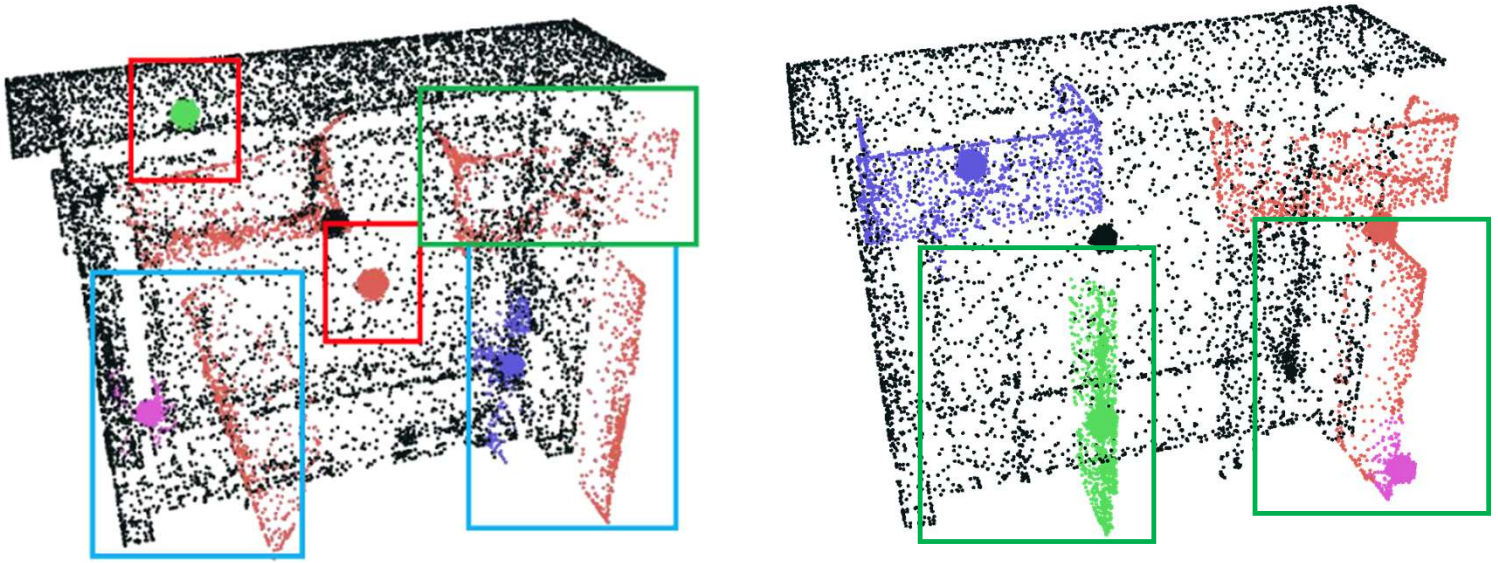
Problem formulation



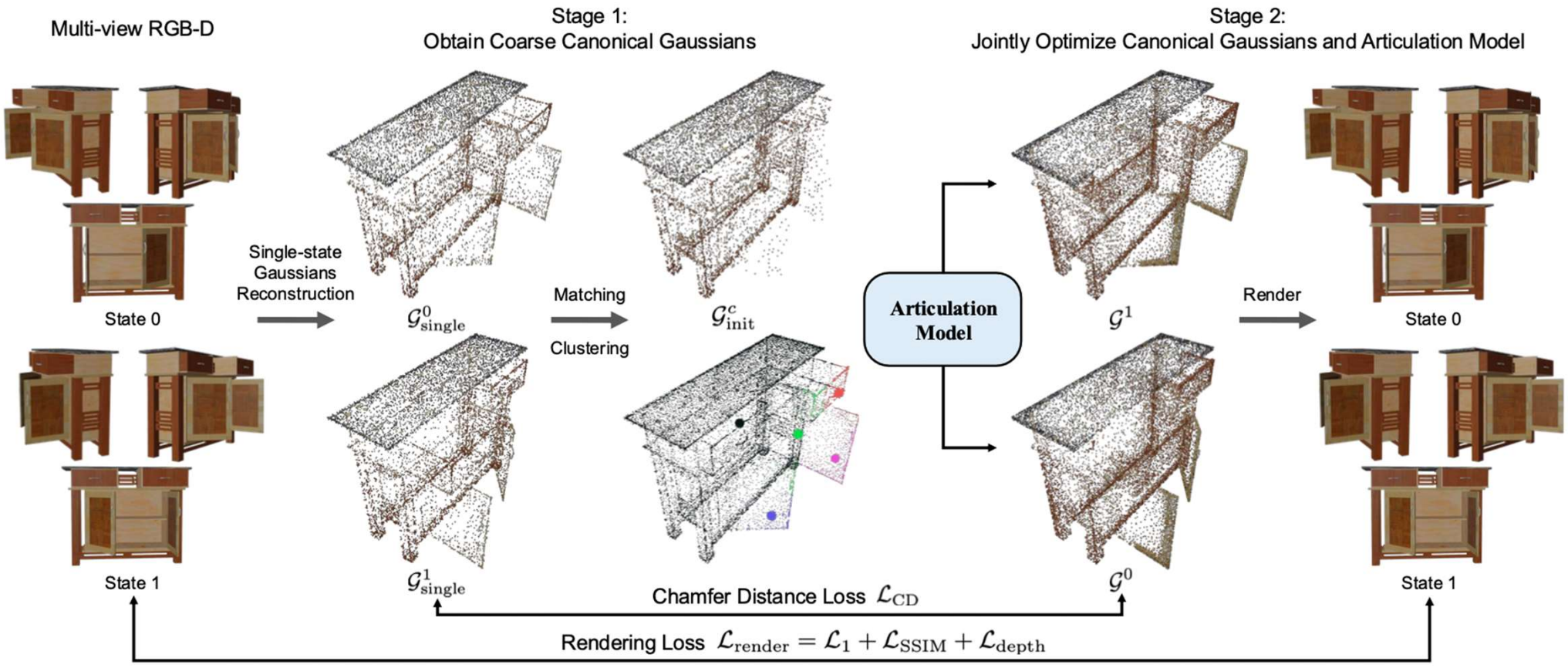
Key challenges

Simultaneous optimization of many correlating variables via rendering

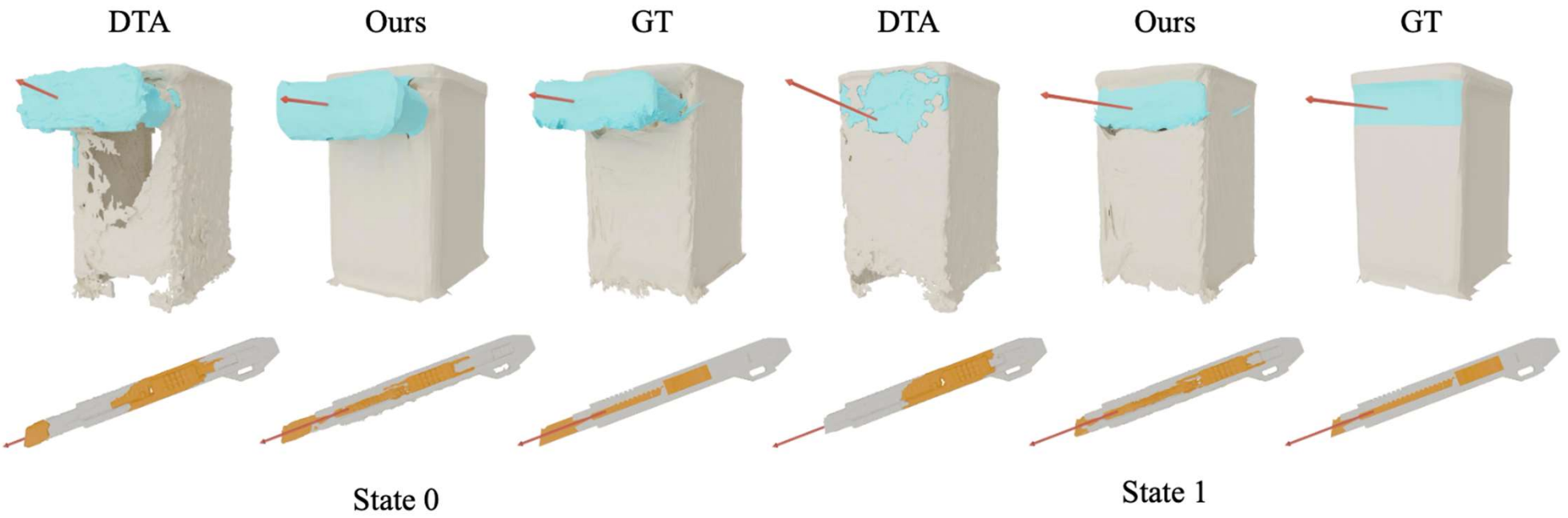
- Canonical Gaussians (base geometry)
- Object part identification (part movement identification)
- Dynamics modeling over Gaussians (articulation parameters)



ArtGS


Multi-view RGB-D
Stage 1:
Obtain Coarse Canonical Gaussians
Stage 2:
Jointly Optimize Canonical Gaussians and Articulation Model
State 0
Single-state
Gaussians
Reconstruction
$\mathcal{G}^0_{\text{single}}$
Matching
Clustering
$\mathcal{G}^c_{\text{init}}$
Articulation
Model
$\mathcal{G}^1$
Render
State 0
State 1
$\mathcal{G}^1_{\text{single}}$
$\mathcal{G}^0$
State 1
Chamfer Distance Loss $\mathcal{L}_{\text{CD}}$
Rendering Loss $\mathcal{L}_{\text{render}} = \mathcal{L}_1 + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{depth}}$

Results

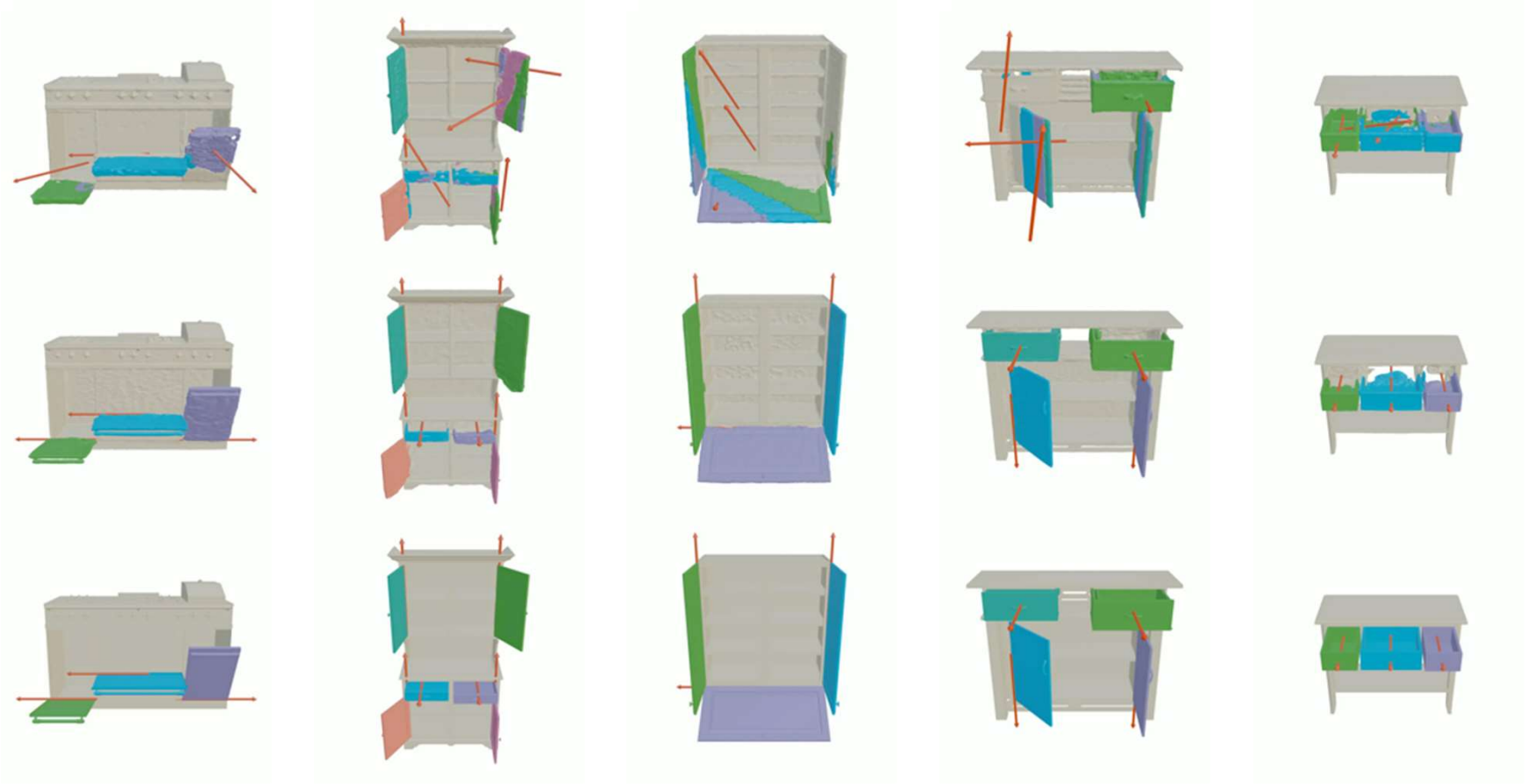


Results

DTA

Ours

GT



ArtGS: Building Interactable Replicas of Complex Articulated Objects via Gaussian Splatting

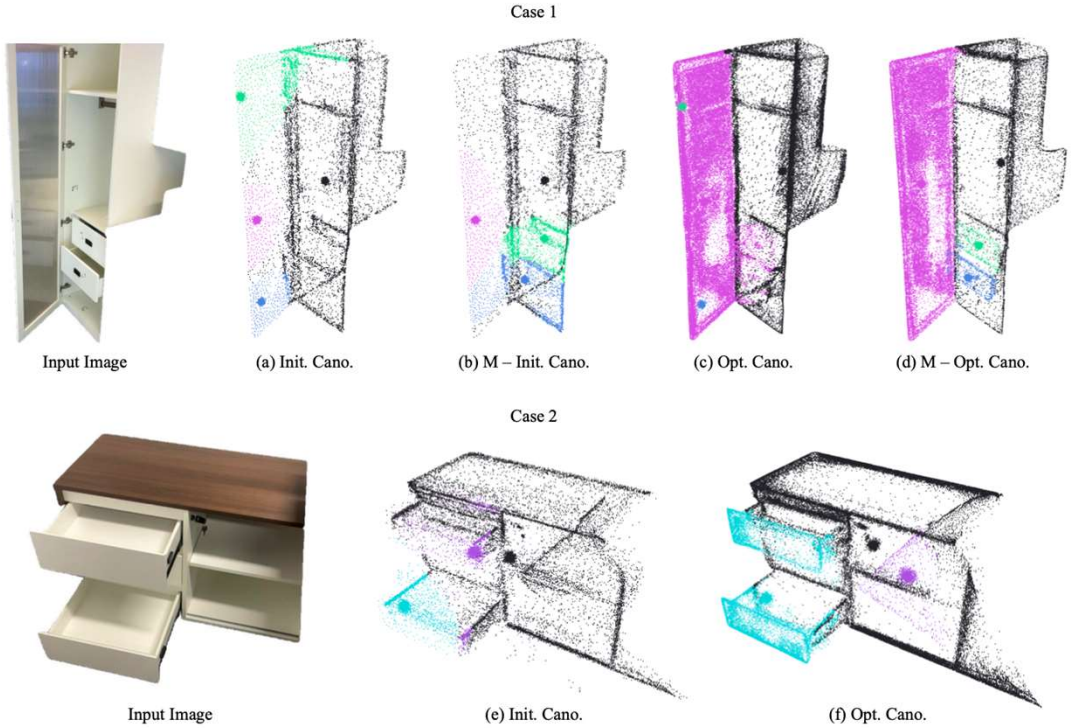
Discussion

Problem Setting

- The two-state setting causes confusion
- Initialization is key to success
- Requires high-quality recording of objects

Future?

- From static captures to videos
- Leveraging pre-trained models (e.g. SAM)
- Feed-forward reconstruction without per-object optimization



Physical Plausible Scene Reconstruction

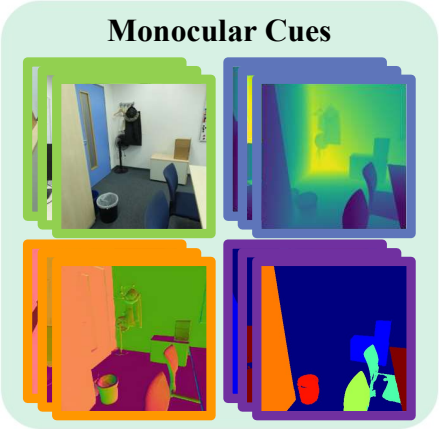
Decompositional Neural Scene Reconstruction with Generative Diffusion Prior

CVPR 2025

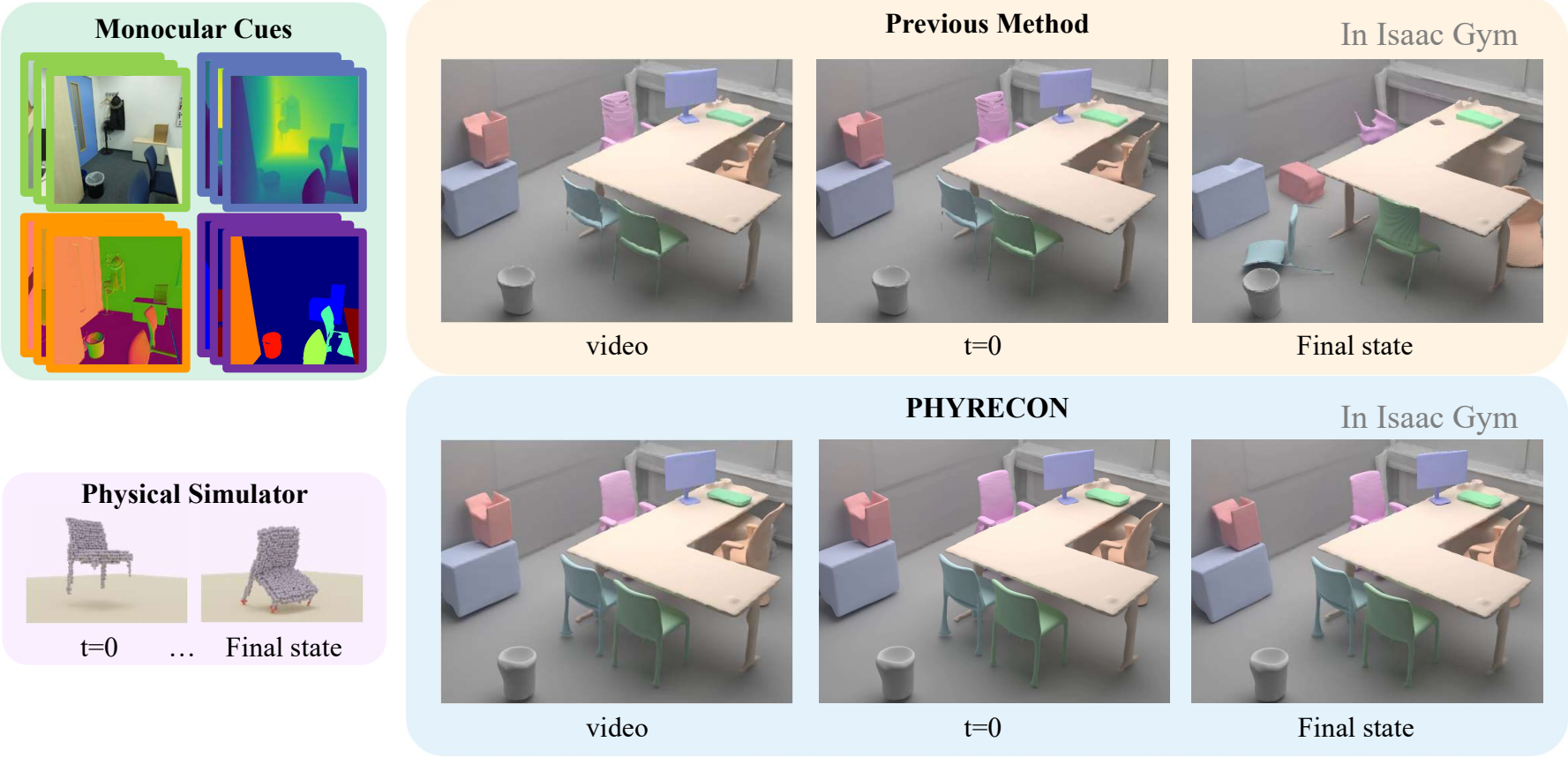
PhyRecon: Physically Plausible Neural Scene Reconstruction

NeurIPS 2024

Physically plausible scene reconstruction

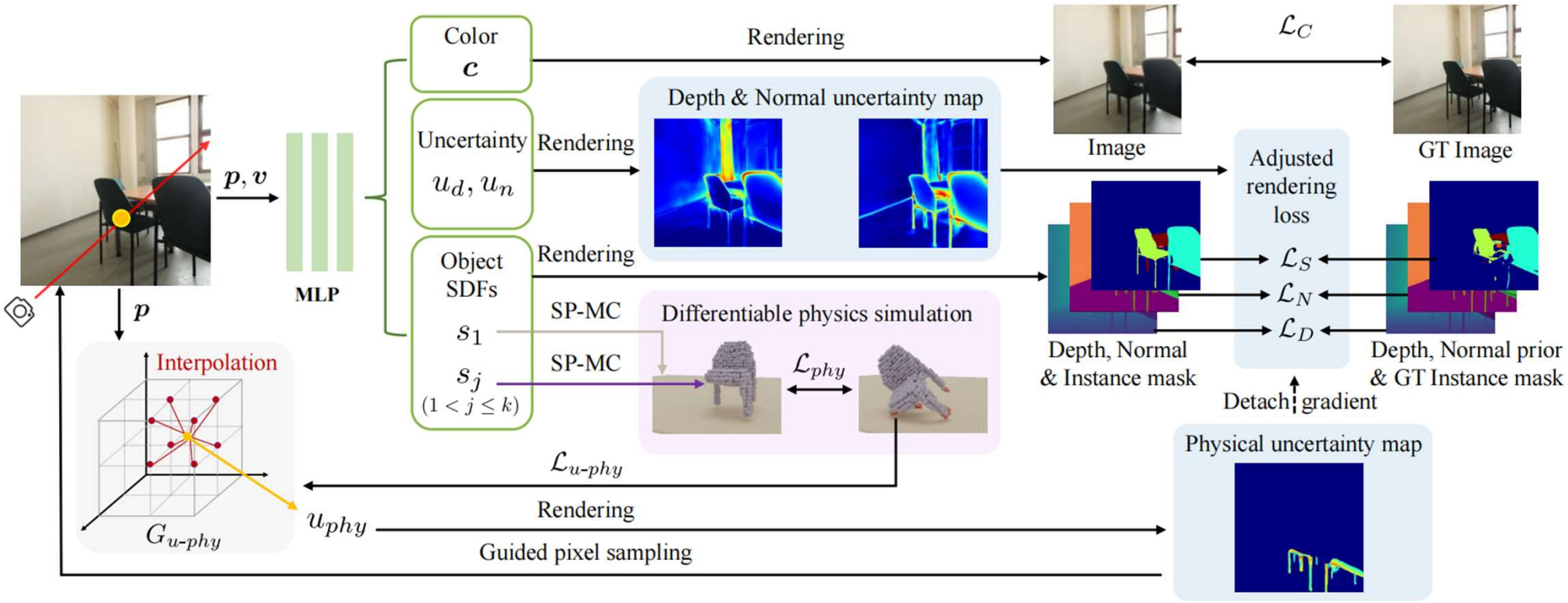


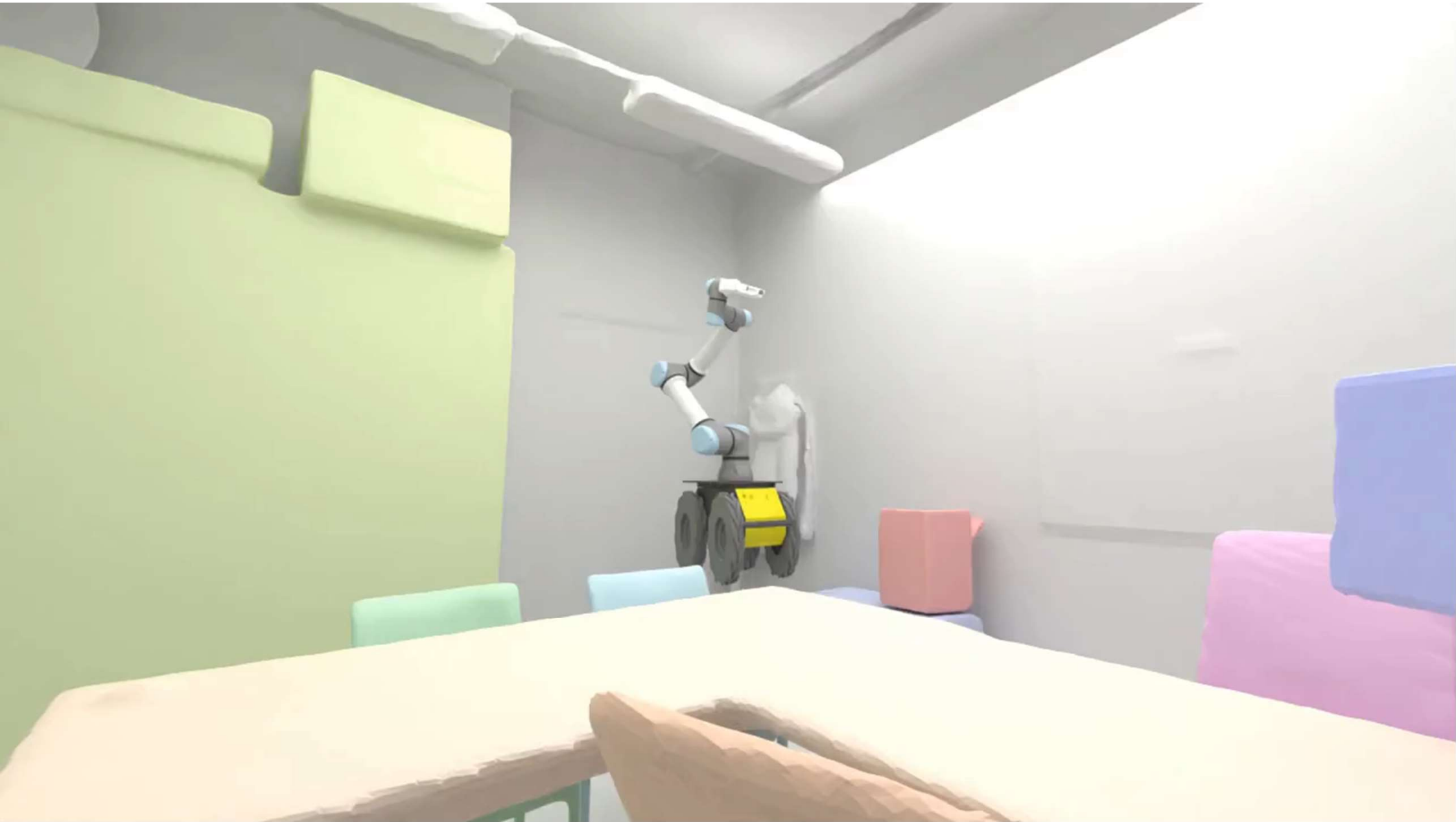
Physically plausible scene reconstruction



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

PhyRecon


Color
c
Uncertainty
u_d, u_n
Object
SDFs
s_1
s_j
$(1 < j \leq k)$
Rendering
Rendering
Rendering
SP-MC
SP-MC
Depth & Normal uncertainty map
Differentiable physics simulation
$\mathcal{L}_{phy}$
Image
$\mathcal{L}_C$
GT Image
Adjusted
rendering
loss
$\mathcal{L}_S$
$\mathcal{L}_N$
$\mathcal{L}_D$
Depth, Normal
& Instance mask
Depth, Normal prior
& GT Instance mask
Detach gradient
p, v
p
MLP
Interpolation
$G_{u\text{-}phy}$
u_{phy}
$\mathcal{L}_{u\text{-}phy}$
Rendering
Guided pixel sampling
Physical uncertainty map



Enough?

- In regions scarcely observed in the input image, objects tend to grow protrusions under the influence of physical loss, maintaining stability but distorting the shape.

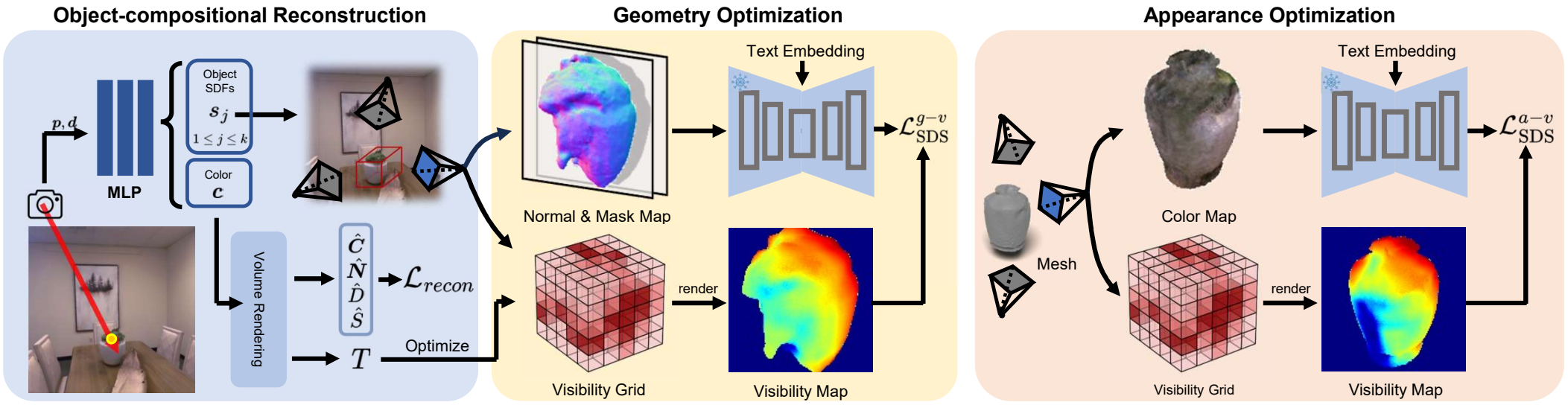
Image

Image View

New View

Leveraging diffusion prior

- **Object-compositional Reconstruction**: Optimize the SDF for each object in the scene.
- **Geometry Optimization**: Incorporate a text-guided geometry prior.
- **Appearance Optimization**: Incorporate a text-guided appearance prior.



Ni et al., Decompositional Neural Scene Reconstruction with Generative Diffusion Prior, CVPR 2025

DP-Recon for game (Replica by 10-views)

Interaction with Scenes


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.
15x
Active Perception
COME-Robot, ICRA 2025

Overall

From the Real2Sim perspective

- Asset substitution with physical optimization can give pretty good static scenes
- Reconstruction of scenes and interactable objects are starting to work
- EAI tasks like vision-language navigation can already be tested on these scenes
- Need more efficient and high-quality scene/object reconstructions for manipulation

…

More to come from BIGAI

https://physcene.github.io/

https://meta-scenes.github.io

https://dp-recon.github.io

https://articulate-gs.github.io

Thank you!