NEURAL
INFORMATION
PROCESSING
SYSTEMS

LiveScene:
Language Embedding Interactive Radiance Fields for Physical Scene Rendering and Control

Neurips 2024

Delin Qu*, Qizhi Chen *, Pingrui Zhang,

Xianqiang Gao, Bin Zhao, Zhigang Wang, Dong Wang†, Xuelong Li

delinqu.cs@gmail.com | https://livescenes.github.io

復旦大學 FUDAN UNIVERSITY

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

浙江大學 ZHEJIANG UNIVERSITY

IPEC

Paper Video Code Data

Making Sesame-Ginger Asian Salad
Cooking Noodles
Cooking an Omelet
Cooking Dumplings
Cooking Pasta
Cooking Brownies
Cooking Tomato & Eggs
Making Milk Tea
Cooking Scrambled Eggs
Cooking Sushi Rolls
Making Chai Tea
Making Greek Salad

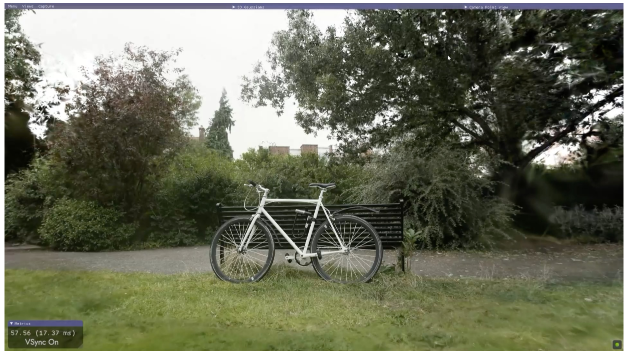
Virtual Reality

Content Creation

Embodied Intelligence

Multimodal Understanding

Challenges in Capturing Real-World Interactions



3d Static Scene[1][2]

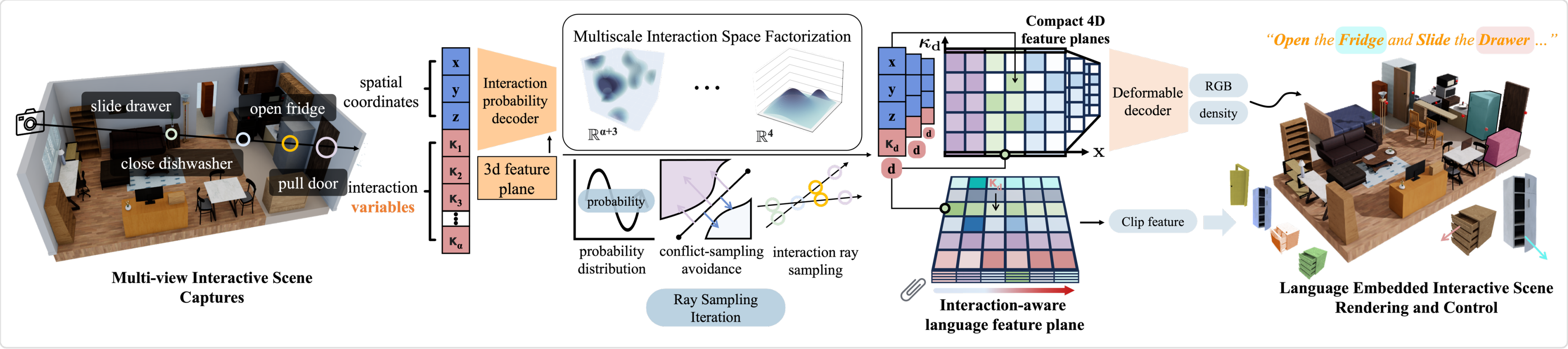
4D Deformable Scene[3][4]

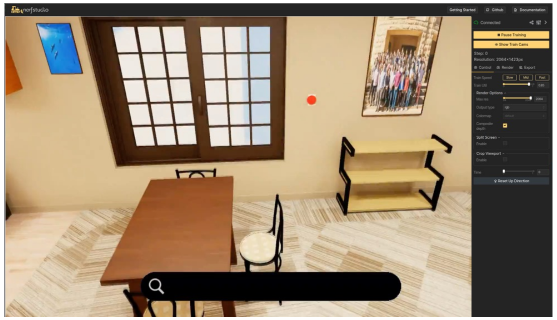
Object Level Control[5][6]

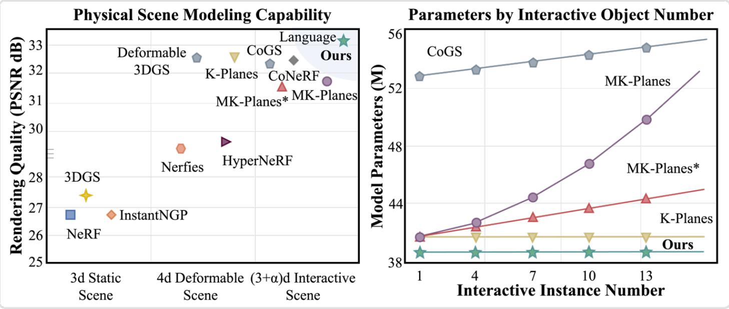
[1] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.
[2] Kerbl, Bernhard, et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." ACM Trans. Graph. 42.4 (2023): 139-1.
[3] Fridovich-Keil, Sara, et al. "K-planes: Explicit radiance fields in space, time, and appearance." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[4] Yang, Ziyi, et al. "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[5] Kania, Kacper, et al. "Conerf: Controllable neural radiance fields." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[6] Yu, Heng, et al. "Cogs: Controllable gaussian splatting." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Complexity of Modeling High-Dimensional Interactive Scenes
Significantly Increasing Computational Time and Memory Cost
Lack of Comprehensive Scene-Level Datasets

The Overview of LiveScene


slide drawer
open fridge
close dishwasher
pull door
spatial coordinates
interaction variables
Multi-view Interactive Scene Captures
x
y
z
K1
K2
K3
Kα
Interaction probability decoder
3d feature plane
Multiscale Interaction Space Factorization
$\mathbb{R}^{\alpha+3}$
$\mathbb{R}^4$
probability
probability distribution
conflict-sampling avoidance
interaction ray sampling
Ray Sampling Iteration
κ_d
Compact 4D feature planes
Kd
d
x
Deformable decoder
RGB
density
Clip feature
Interaction-aware language feature plane
"Open the Fridge and Slide the Drawer ..."
Language Embedded Interactive Scene Rendering and Control



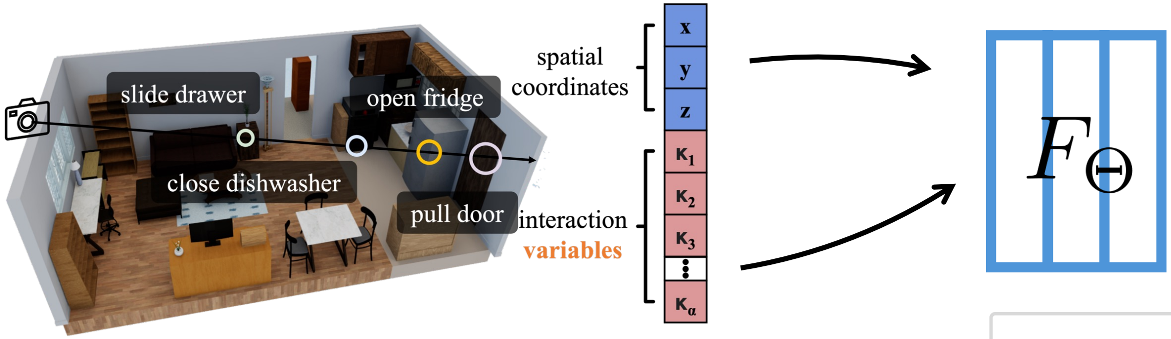


▶ the first scene-level language-embedded interactive radiance field

▶ enabling manipulation of multiple articulated objects and language-based interaction

Interactive Space

Assuming a non-rigidly interactive scene with α control variables κ = [κ$_1$, κ$_2$, ..., κ$_α$] corresponding to α objects, we delineate its representation by a high-dimensional function:



Multi-view Interactive Scene Captures

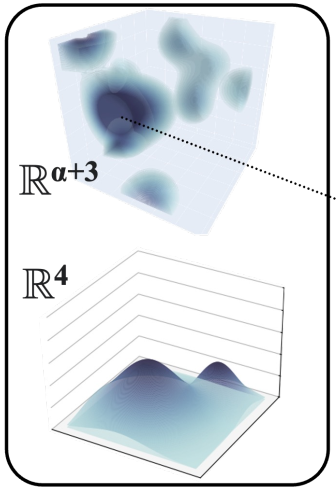
$$\mathbf{y} = \rho(\mathbf{x}, \boldsymbol{\kappa}, \mathcal{H}; \boldsymbol{\theta}),$$

jointly model with spatial and interaction variables is complex and computational

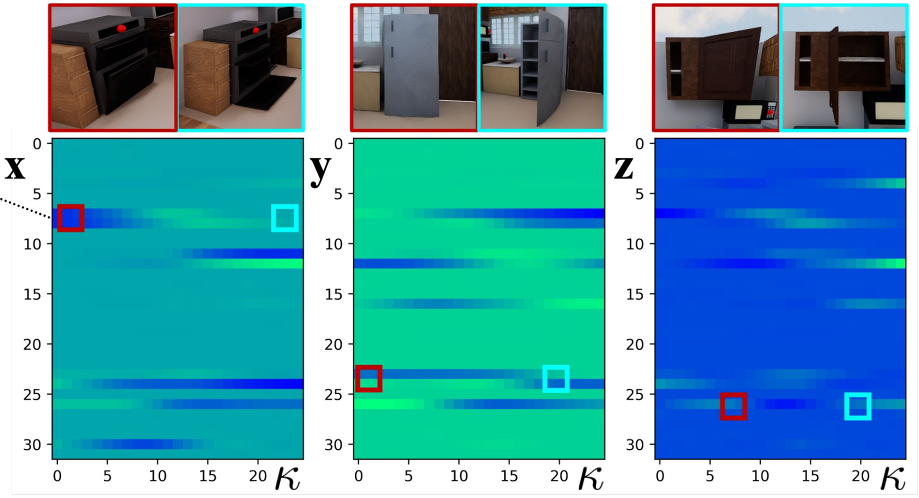
Ray Samples $\mathbf{p} = [\mathbf{x} \mid \boldsymbol{\kappa}] \in \mathbb{R}^{(3+\boldsymbol{\alpha})}$

Multi-scale Interaction Space Factorization

Interaction features are distributed in (3 + α)-D interactive space and aggregate into cluster centers, which can be projected into a compact 4-dimensional space $\mathbb{R}^4$



Interaction Space Factorization

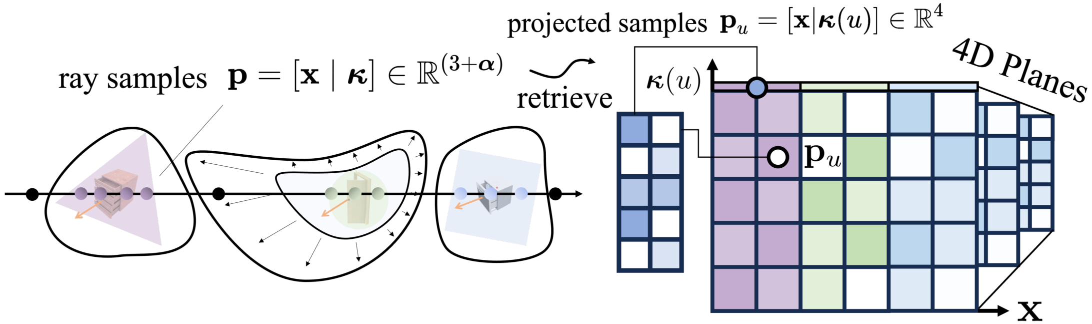


a) Feature visualization in compact 4d planes

b) Interaction Scene with 10 objects

Multi-scale Interactive Ray Sampling

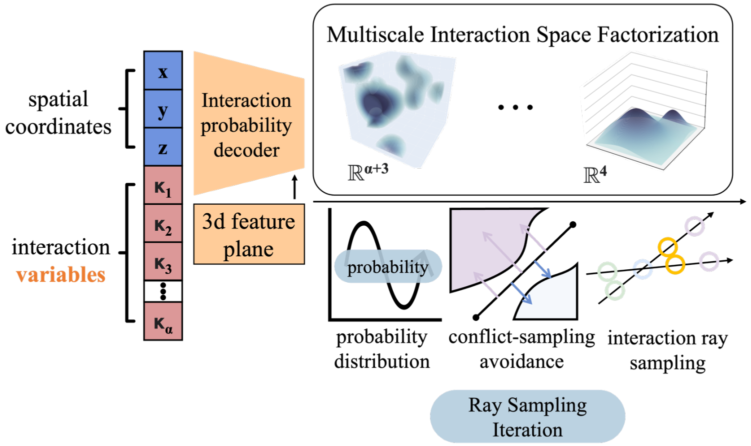
For a given intersection point p, the deformable features can be retrieved from the corresponding local 4D deformable field by maximizing sampling probability P



ray samples $\mathbf{p} = [\mathbf{x} \mid \boldsymbol{\kappa}] \in \mathbb{R}^{(3+\boldsymbol{\alpha})}$
projected samples $\mathbf{p}_u = [\mathbf{x}|\boldsymbol{\kappa}(u)] \in \mathbb{R}^4$
retrieve
$\boldsymbol{\kappa}(u)$
$\mathbf{p}_u$
4D Planes
$\mathbf{x}$

a) Local deformable fields

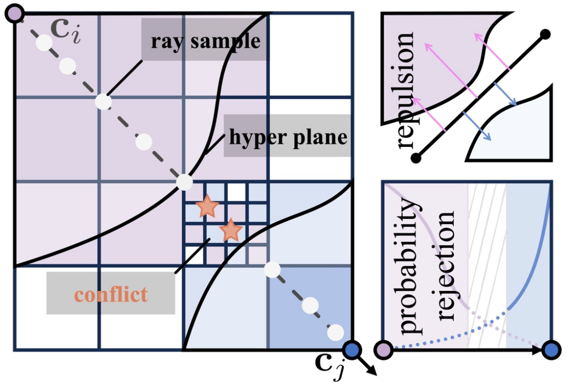
b) Compact multiscale interaction feature storage



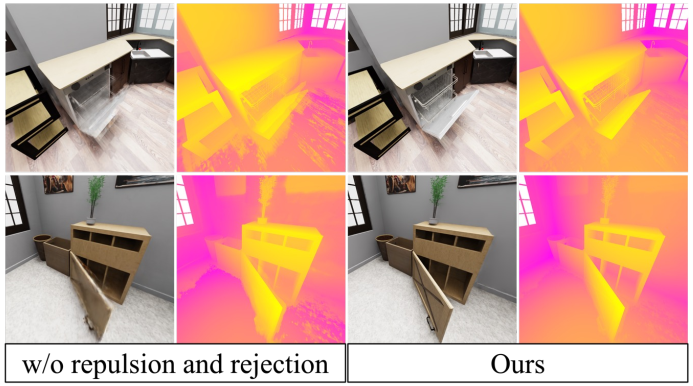
$$\mathbf{p}_u = [\mathbf{x}|\boldsymbol{\kappa}(u)]\,, \quad u = \arg\max_i\{\mathbf{P}_i\}, \quad \mathbf{P} = \Theta(\boldsymbol{\kappa}, \boldsymbol{\theta}),$$

Feature Repulsion and Probability Rejection

a repulsion loss for ray pairs ($\mathbf{r}_i$, $\mathbf{r}_j$), and amplify the feature differences between distinct deformable regions, promoting the separation of deformable field



a) Local Feature Conflicts



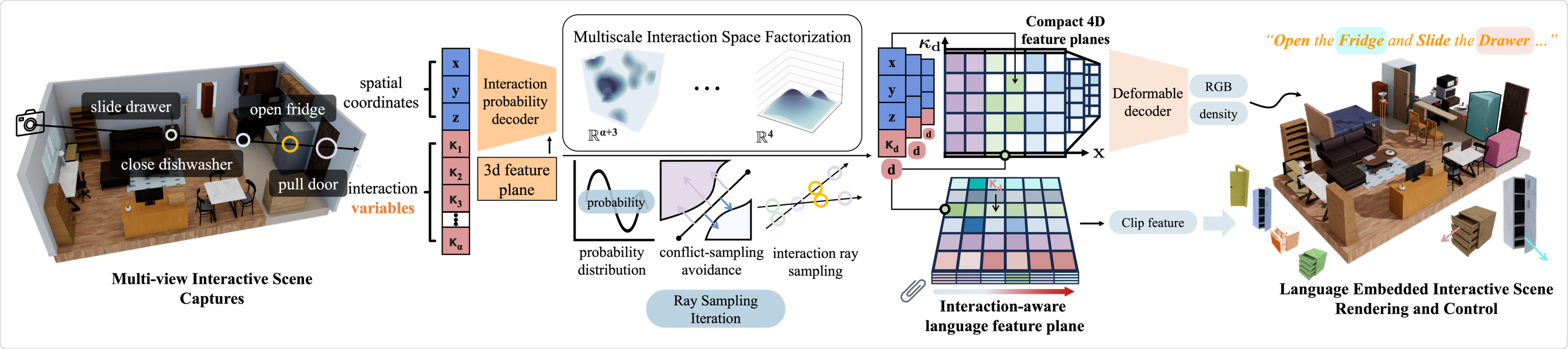
b) Rendering Quality Comparison

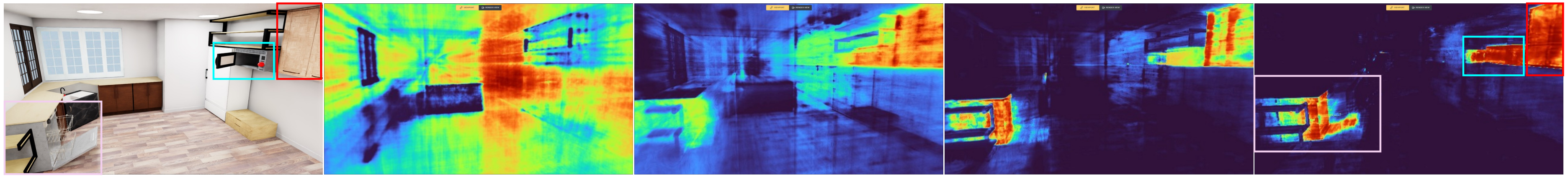
$$u = \begin{cases} \arg\max_i\{\mathbf{P}_i\}, & \text{if} \quad \mathbf{P}_i \geq s \\ -1, & \text{otherwise} \end{cases}$$

Probability Rejection

$$\mathcal{L}_{\text{repuls}} = \mathbf{ELU}(K - \|(\mathbf{M}_i \odot \mathbf{M}_j)(\mathcal{F}_i - \mathcal{F}_j)\|),$$

The Overview of LiveScene





Interactive Scene

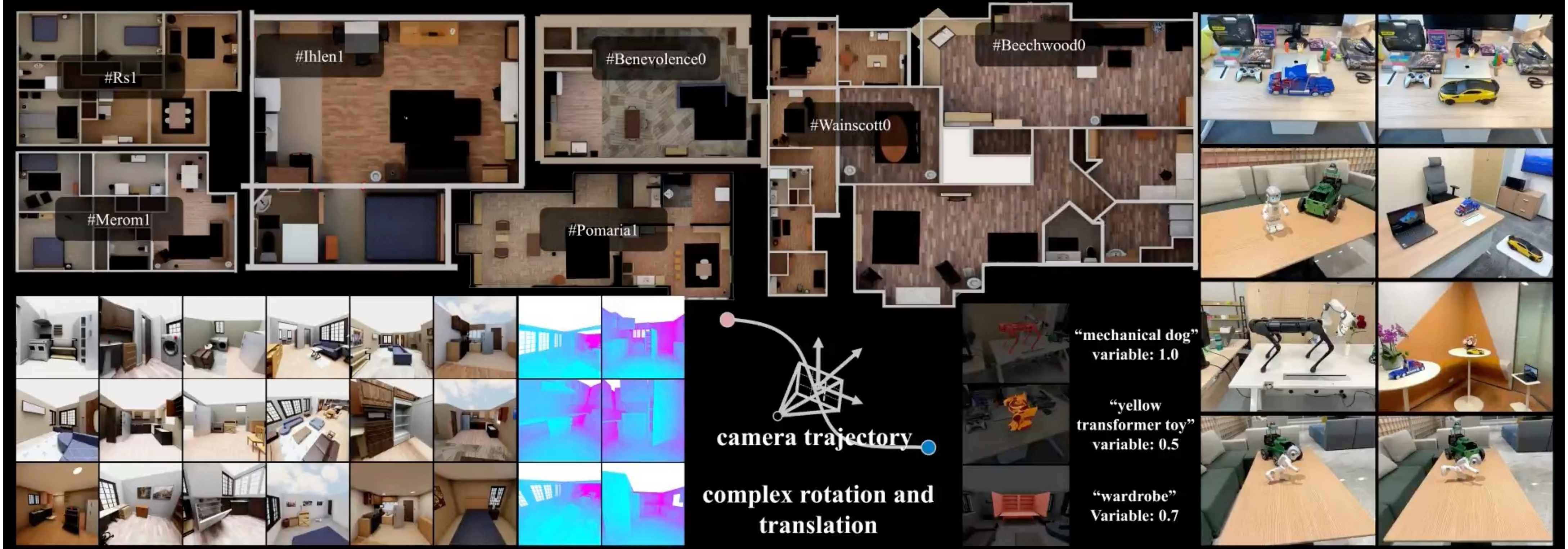
step 100

step 250

step 500

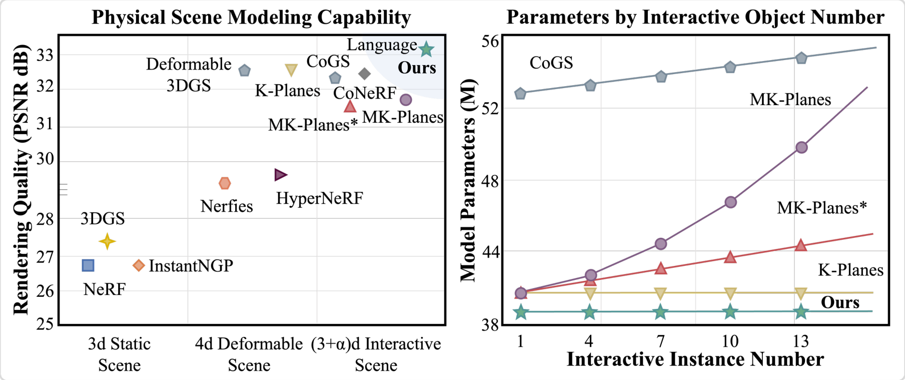
step 1000

OmniSim Behavior Synthetic and InterReal Dataset



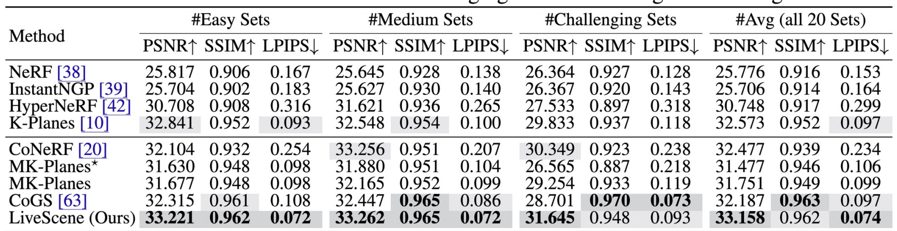
#28 Interactive Subsets with 2 Millions Sample including RGB, Depth, Segmentation, Camera Poses, Interaction Variables, and Object Captions Modalities

Experiment results



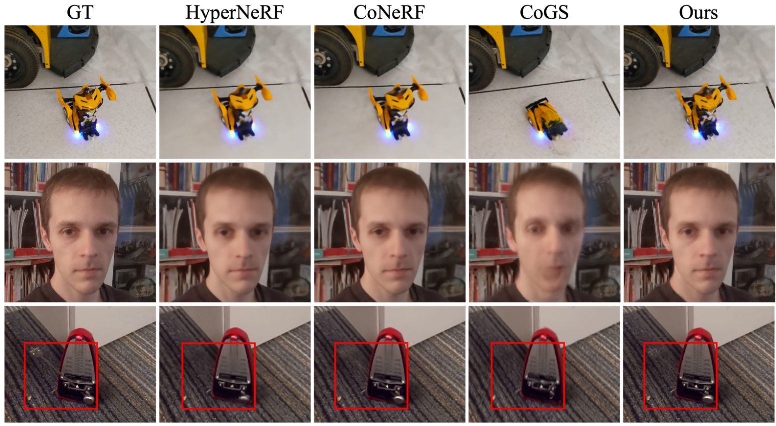
Modeling Capability and Efficiency

Table 2: **Quantitative results on OmniSim Dataset**. LiveScene outperforms prior works on most metrics and achieves the best PSNR on the #challenging subset with a significant margin.

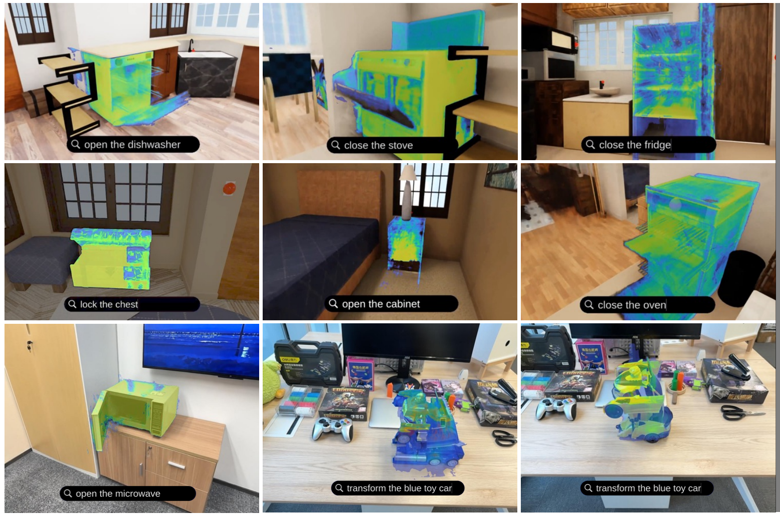


Method	#Easy Sets			#Medium Sets			#Challenging Sets			#Avg (all 20 Sets)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF [38]	25.817	0.906	0.167	25.645	0.928	0.138	26.364	0.927	0.128	25.776	0.916	0.153
InstantNGP [39]	25.704	0.902	0.183	25.627	0.930	0.140	26.367	0.920	0.143	25.706	0.914	0.164
HyperNeRF [42]	30.708	0.908	0.316	31.621	0.936	0.265	27.533	0.897	0.318	30.748	0.917	0.299
K-Planes [10]	32.841	0.952	0.093	32.548	0.954	0.100	29.833	0.937	0.118	32.573	0.952	0.097
CoNeRF [20]	32.104	0.932	0.254	33.256	0.951	0.207	30.349	0.923	0.238	32.477	0.939	0.234
MK-Planes*	31.630	0.948	0.098	31.880	0.951	0.104	26.565	0.887	0.218	31.477	0.946	0.106
MK-Planes	31.677	0.948	0.098	32.165	0.952	0.099	29.254	0.933	0.119	31.751	0.949	0.099
CoGS [63]	32.315	0.961	0.108	32.447	**0.965**	0.086	28.701	**0.970**	**0.073**	32.187	**0.963**	0.097
LiveScene (Ours)	**33.221**	**0.962**	**0.072**	**33.262**	**0.965**	**0.072**	**31.645**	0.948	0.093	**33.158**	0.962	**0.074**

Render Quality Comparision

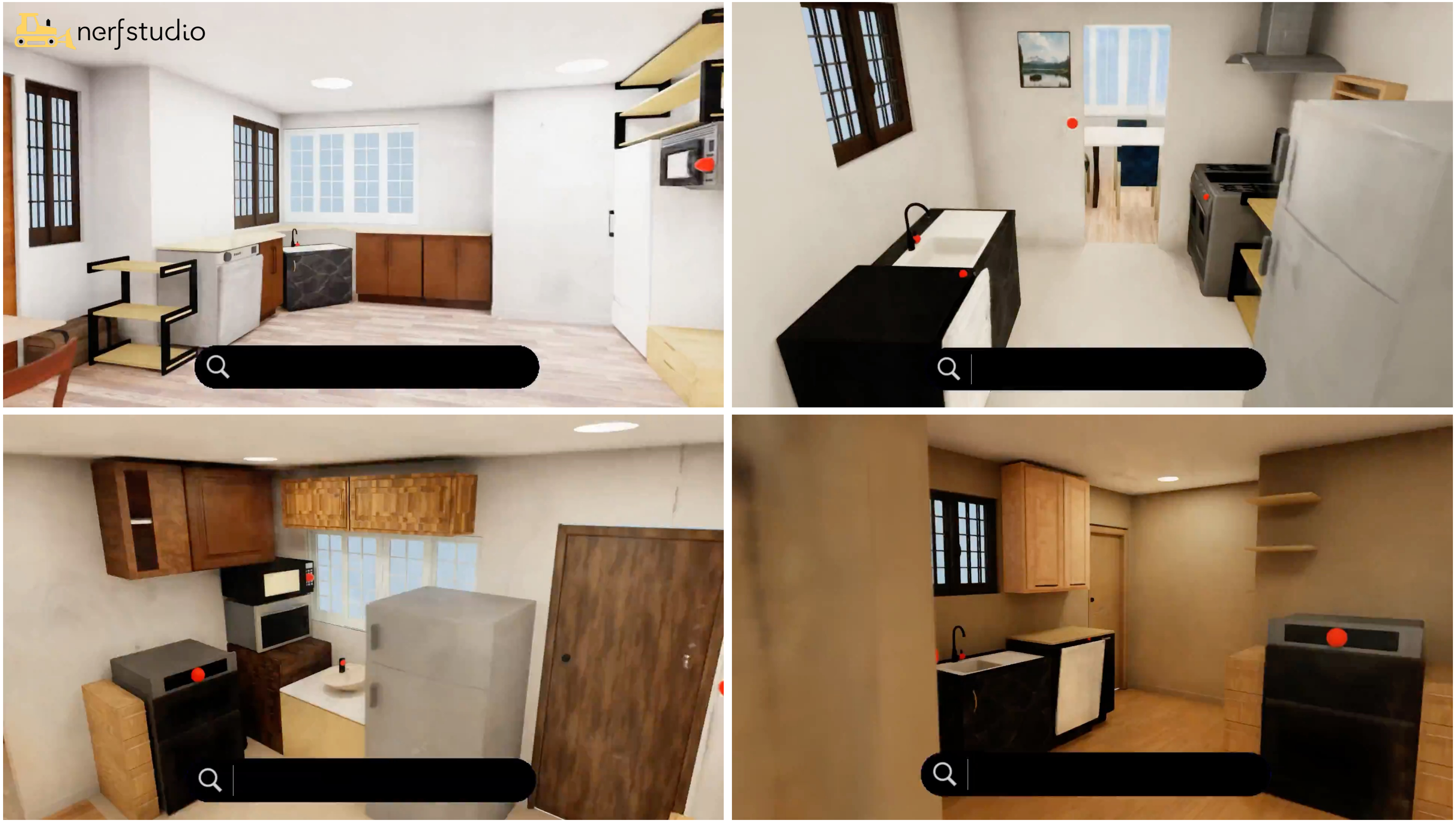


Rendering visualization



Language Control

Control with Language Instruction


nerfstudio

For more extensive evaluation and dataset download please check the paper and project website

https://livescenes.github.io