S-INF: Towards Realistic Indoor Scene Synthesis via Scene Implicit Neural Field

Zixi Liang, Guowei Xu, Haifeng Wu, Ye Huang, Wen Li*, Lixin Duan

Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China

AAAI-25 / IAAI-25 / EAAI-25 | FEBRUARY 25 - MARCH 4, 2025 | PHILADELPHIA, USA

Abstract — Learning-based methods have become increasingly popular in 3D indoor scene synthesis (ISS), showing superior performance over traditional optimization-based approaches. These learning-based methods typically model distributions on simple yet explicit scene representations using generative models. However, due to the oversimplified explicit representations that overlook detailed information and the lack of guidance from multimodal relationships within the scene, most learning-based methods struggle to generate indoor scenes with realistic object arrangements and styles. In this paper, we introduce a new method, Scene Implicit Neural Field (SINF), for indoor scene synthesis, aiming to learn meaningful representations of multimodal relationships, to enhance the realism of indoor scene synthesis. S-INF assumes that the scene layout is often related to the object-detailed information. It disentangles the multimodal relationships into scene layout relationships and detailed object relationships, fusing them later through implicit neural fields (INFs). By learning specialized scene layout relationships and projecting them into S-INF, we achieve a realistic generation of scene layout. Additionally, S-INF captures dense and detailed object relationships through differentiable rendering, ensuring stylistic consistency across objects. Through extensive experiments on the benchmark 3D-FRONT dataset, we demonstrate that our method consistently achieves state-of-the-art performance under different types of ISS.

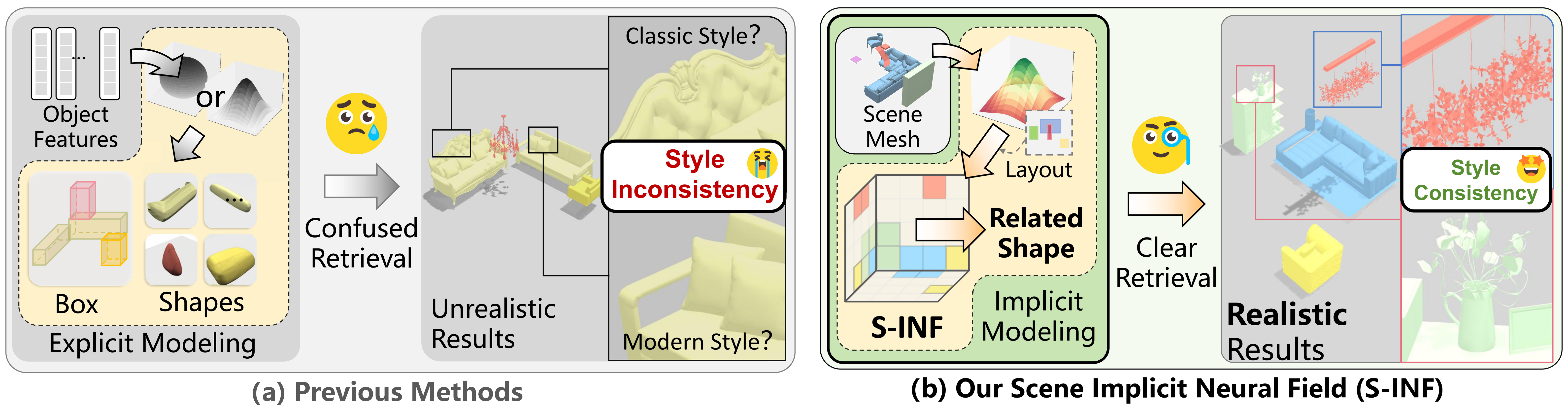

Figure 1: We enhance the implicit modeling process from the S-INF with scene layout relationships and detailed object relationships, to achieve more realistic generations.

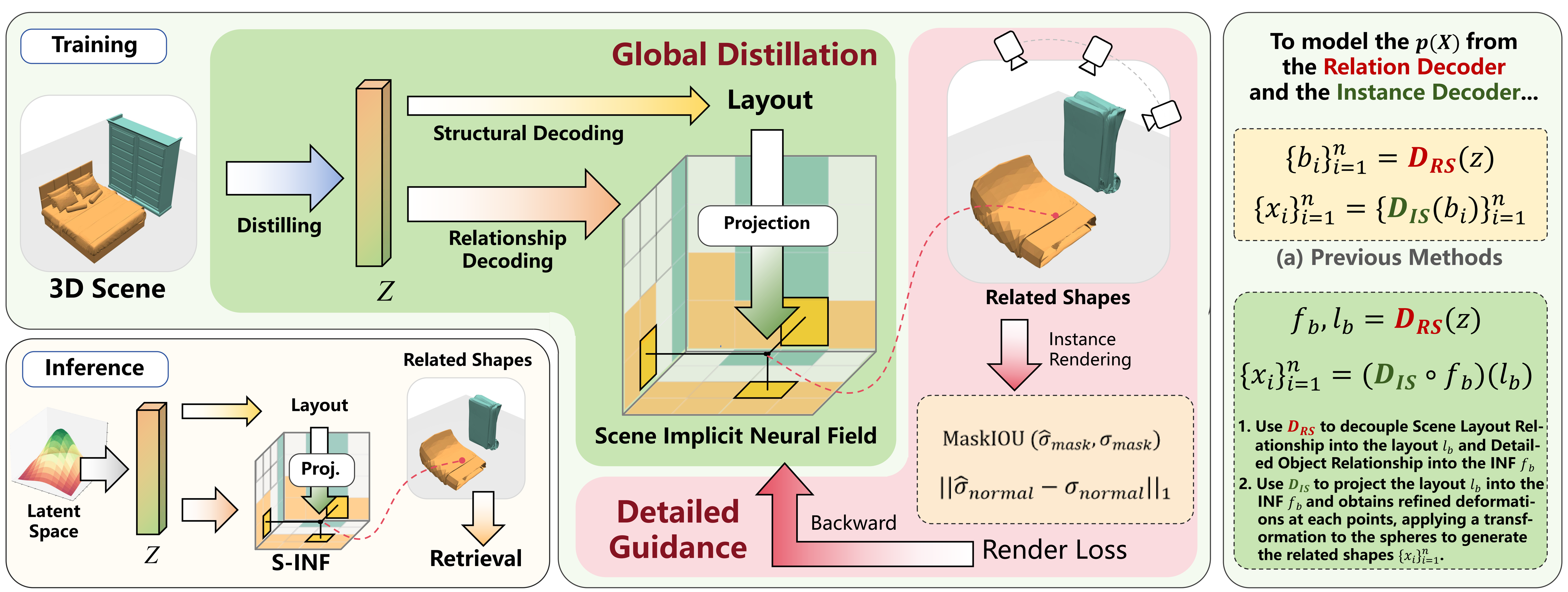

Figure 2: Our approach focuses on developing the S-INF to enable efficient capture of multimodal relationships and generate realistic and reliable 3D indoor scenes.

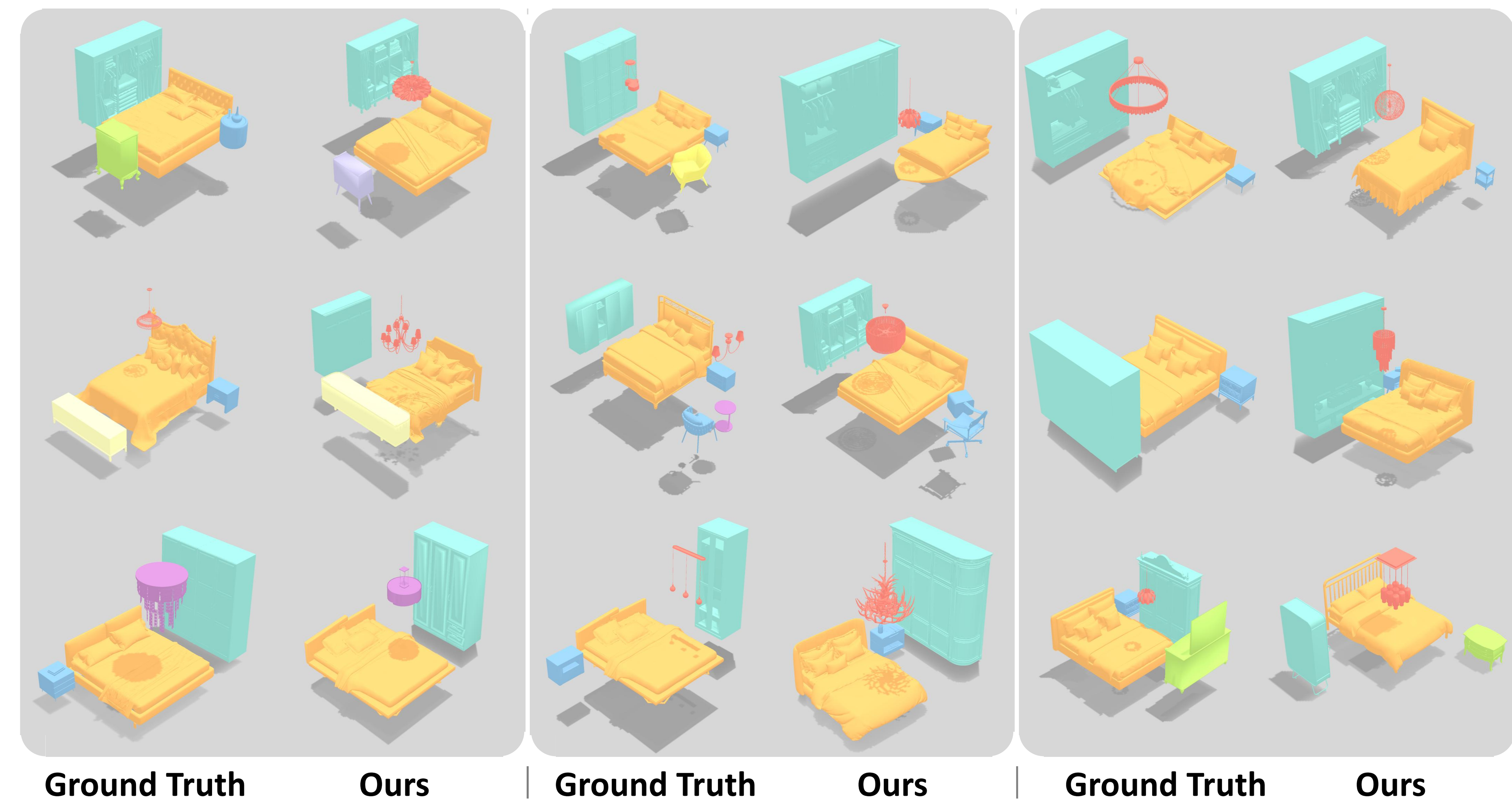

Figure 3: Comparisons between ground truth (cols 1, 3, and 5) and our generated 3D indoor scenes (cols 2, 4, and 6).

		Bedroom				Living Room				Dining Room			
Met.	Re?	FID↓	KL↓	SCA~	Div↑	FID↓	KL↓	SCA~	Div↑	FID↓	KL↓	SCA~	Div↑
NN	×	161.39	0.2521	0.8832	0.7284	188.25	0.1455	0.8565	0.7354	203.34	0.4252	0.9163	0.7346
Tr	×	156.42	0.2082	0.8742	0.7271	156.48	0.1339	0.8583	0.7414	162.82	0.2528	0.8821	0.7456
SP	×	164.70	0.2165	0.8596	0.7338	172.31	0.1357	0.8205	0.7390	168.85	0.2640	0.8812	0.7441
Ours	×	**131.47**	**0.1985**	**0.8393**	**0.7379**	**155.47**	**0.1334**	**0.8034**	**0.7431**	**157.93**	**0.2342**	**0.8476**	**0.7465**
SG	√	119.68	0.2403	0.8128	0.7083	179.32	0.2190	0.8669	0.7260	99.97	0.2866	0.7062	0.7146
AT	√	118.38	0.2069	0.7632	0.7161	174.13	0.3024	0.8871	0.7320	131.16	0.2397	0.7802	0.7063
NN	√	114.63	0.2521	0.7105	0.7072	81.34	0.1455	0.6455	0.7188	84.46	0.4252	0.6437	0.7174
Tr	√	117.02	0.2082	0.7865	0.7161	85.46	0.1339	0.6582	0.7106	97.55	0.2528	0.7020	0.7094
SP	√	110.74	0.2165	0.7994	0.7223	84.36	0.1357	0.6708	0.7269	131.20	0.2640	0.8792	0.7036
ES	√	107.27	-	0.6994	0.7259	109.30	-	0.6792	0.7275	119.30	-	0.7500	0.7208
DS	√	84.40	0.2060	0.6319	0.7118	82.00	0.1771	0.6433	0.7319	86.28	0.2464	0.6533	0.7206
IS	√	89.68	-	0.6202	0.7231	80.15	-	0.6435	0.7233	86.55	-	0.6706	0.7285
Ours	√	**79.52**	**0.1985**	**0.6128**	**0.7319**	**78.98**	**0.1334**	**0.6371**	**0.7322**	**79.95**	**0.2342**	**0.6236**	**0.7297**

Table 1: We conducted a quantitative comparison of our method with state-of-the-art approaches on the 3D-FRONT dataset, where our method consistently demonstrated superior performance. Note that since EchoScene (ES) and InstructScene (IS) are class-conditional, their category-KL divergence is excluded, "Re" is retreival and "~" in SCA indicates that 0.5 is optimal.

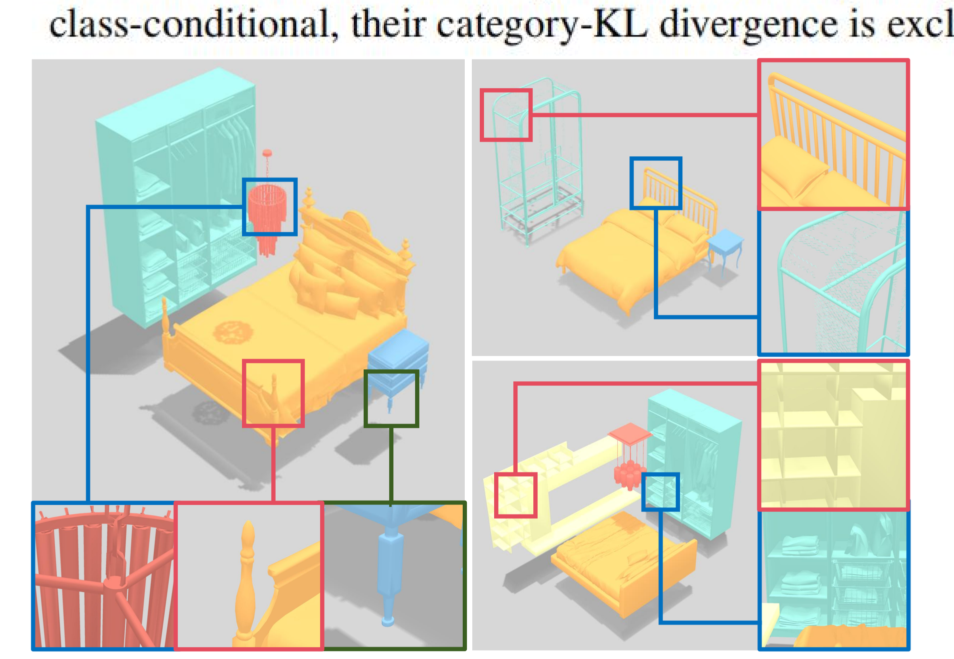

Figure 4: Realistic ISS from multimodal relationships

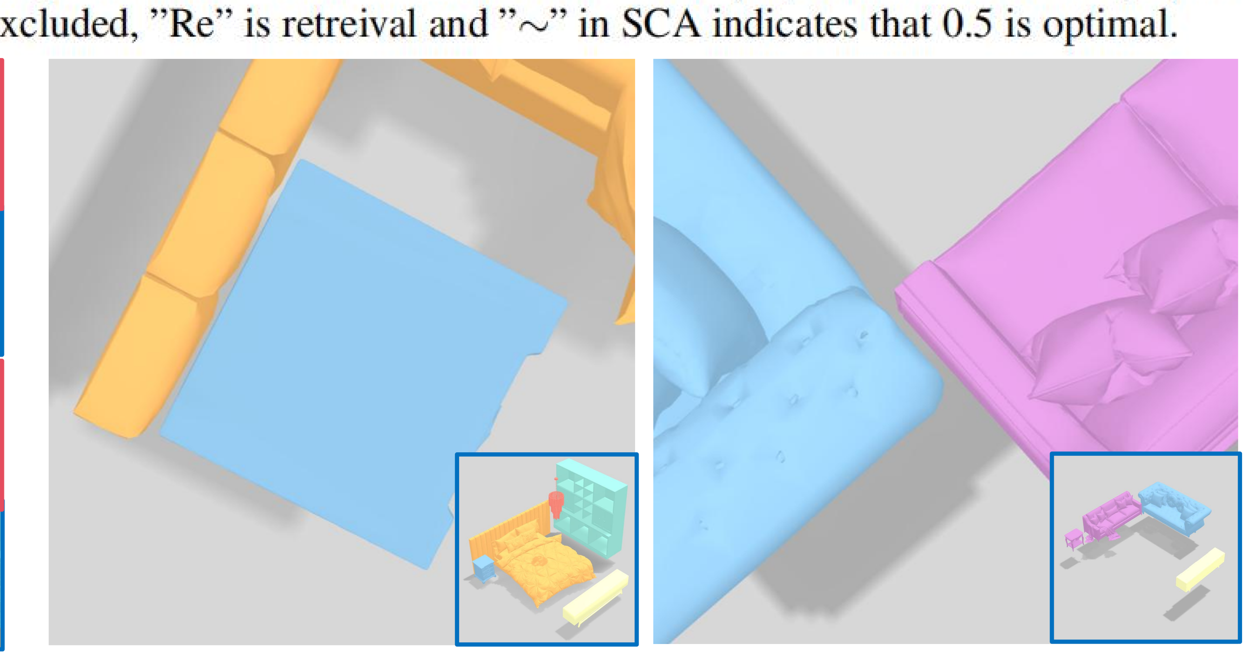

Figure 5: Style consistancy ISS from detailed object relationships.