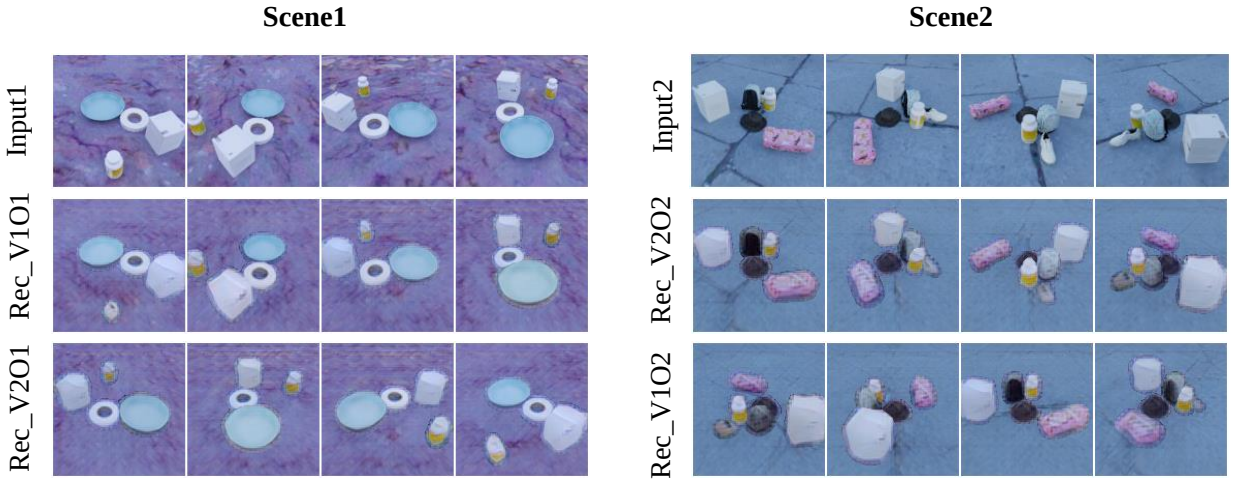

Figure 1. **The visualization of disentanglement learning of OIGO on GSO dataset.** 'Input1' and 'Input2' indicate two scene images observed from four random viewpoints. 'Rec_V1O1' and 'Rec_V2O2' denote reconstruction observed from four random viewpoints of two scenes. 'Rec_V2O1' and 'Rec_V1O2' represent, respectively, an image reconstructed using the object representation of 'Scene1' with the viewpoint representation of 'Scene2' and an image reconstructed using the object representation of 'Scene2' with the viewpoint representation of 'Scene1'.

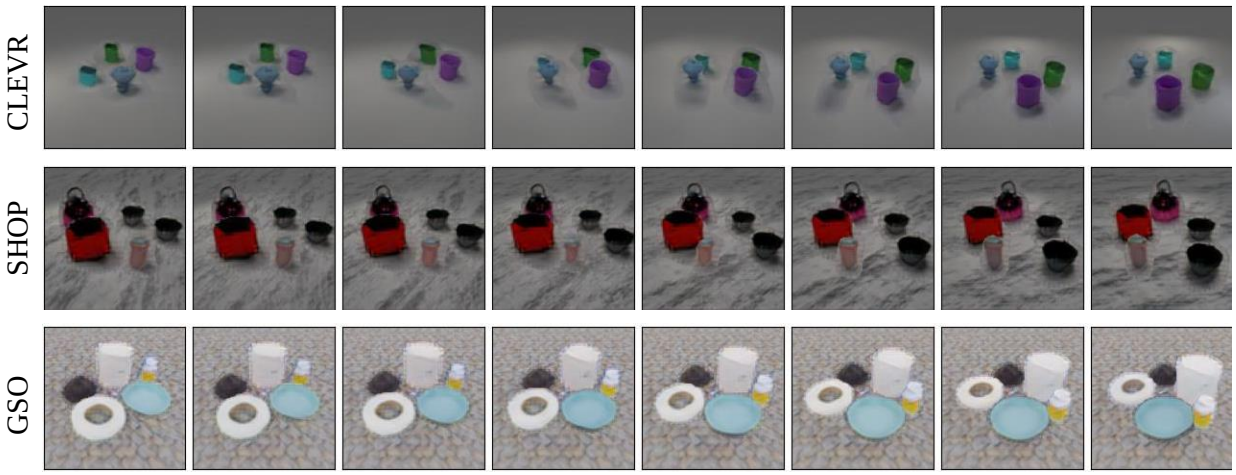

Figure 2. **The visualization of viewpoint interpolation of OGIO on CLEVR, SHOP, and GSO datasets.** Each row represents the interpolation visualization of one dataset. Interpolation between viewpoint representations of the first column image and the last column image of each row.

Table 1. Performance comparison of OIGO, OCLOC, SIMONe, STEVE, and SlotDiffusion for mBO metric

MODEL	CLEVR	SHOP	GSO
OCLOC	0.026±6E-3	0.065±8E-4	0.184±3E-4
SIMONE	0.254±5E-4	0.276±6E-5	0.063±5E-6
SIMONE (S)	0.494±4E-5	0.501±2E-5	0.084±4E-6
STEVE	**0.781±3E-3**	**0.734±4E-4**	0.614±5E-4
SLOTDIFFUSION	0.201±4E-3	0.18±4E-2	0.462±5E-4
OIGO	0.475±4E-4	0.558±3E-3	**0.773±2E-4**

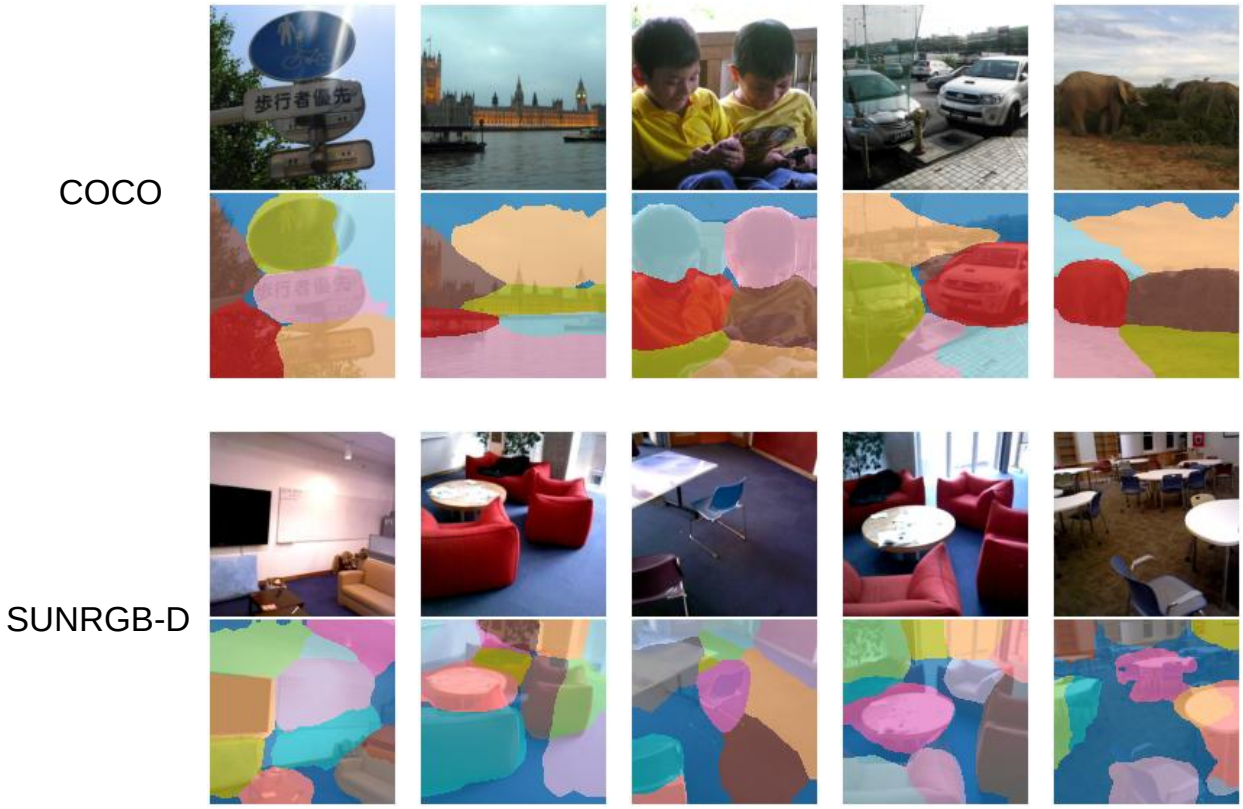

Figure 3. **The scene segmentation visualization results of OIGO on two complex real scene datasets, COCO and SUNRGB-D.**

Table 2. Ablation study of OIGO on viewpoint-attribute disentanglement. 'w/o VO' idicates the model does not disentangle viewpoint and object attributes. 'w/o TS' denotes the model simultaneously infers viewpoint and object attribute representations via the slot attention module, the same as OCLOC. 'w/o FR' denotes the model uses the Spatial Broadcast Decoder in OCLOC instead of the proposed *Patch decoder*. The ACC of OIGO, 'w/o VO', 'w/o TS' and 'w/o FR' is shown in the table.

MODEL	CLEVR	SHOP	GSO
'W/O TS'	0.736±5E-3	0.765±4E-4	0.464±5E-3
'W/O VO'	0.435±6E-4	0.576±8E-3	0.635±5E-4
'W/O TS'	0.713±3E-4	0.765±2E-3	0.801±4E-4
OIGO	**0.788±6E-4**	**0.826±3E-4**	**0.864±1E-4**

Table 3. The training cost of OIGO and the comparison methods.

DATASETS	METHODS	NUMBER OF MODEL PARAMETERS	TIMES	GPU USAGE	BATCH SIZE
CLEVR	OIGO	862M	40H	2	16
	GOCL	1715M	10H	1	32
	LSD	9M	39H	1	64
	STEVE	17M	8H	1	32
SHOP	OIGO	862M	46H	2	8
	GOCL	171M	10H	1	32
	LSD	9M	39H	1	64
	STEVE	17M	8H	1	32
GSO	OIGO	862M	50H	2	16
	GOCL	171M	13H	1	32
	LSD	348M	44H	1	64
	STEVE	16M	10H	1	32

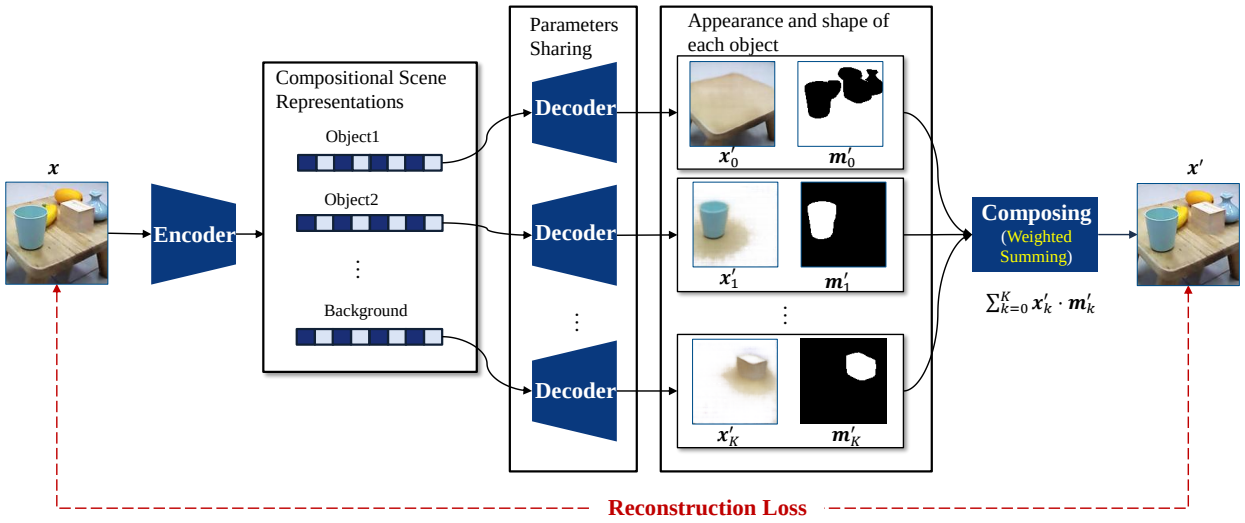

Figure 4. The framework of Unsupervised Object-Centric Learning based on Reconstruction Loss.

Table 4. Quantitative comparison of object attribute learning (i.e., scene segmentation) between OIGO and GOCL on CLEVR, SHOP, and GSO datasets.

DATA SET	MODEL	ARI-A↑	ARI-O↑	MIOU↑	MBO↑
CLEVR	GOCL	0.303±4E-4	0.046±2E-3	0.025±3E-3	0.065±5E-3
	OIGO	**0.605±2E-4**	**0.954±4E-3**	**0.568±3E-3**	**0.475±4E-4**
SHOP	GOCL	0.295±4E-3	0.016±5E-3	0.032±3E-3	0.054±3E-3
	OIGO	**0.587±2E-4**	**0.945±5E-3**	**0.563±4E-3**	**0.558±3E-3**
GSO	GOCL	0.385±3E-3	0.011±5E-3	0.065±6E-3	0.105±8E-3
	OIGO	**0.721±1E-3**	**0.946±4E-3**	**0.650±3E-3**	**0.773±2E-4**

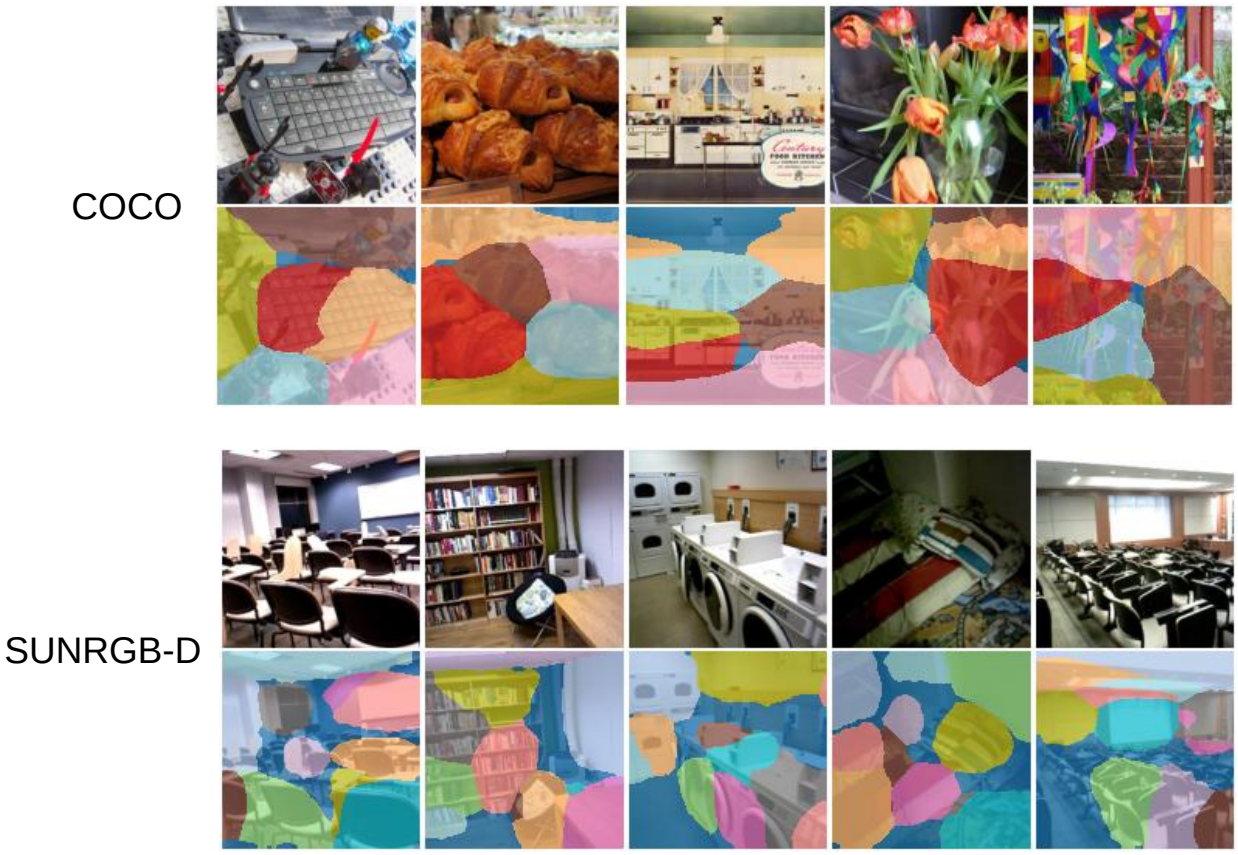

Figure 5. **The failure cases of OIGO on two complex real-world datasets, COCO and SUNRGB-D.**