Generative Scene Graph Networks

Fei Deng[1], Zhuo Zhi[2], Donghun Lee[3], Sungjin Ahn[1]

[1]Rutgers University, [2]University of California, San Diego, [3]ETRI

Introduction

- Goal:
  - Unsupervised scene graph discovery
- Motivation:
  - Model part-whole relationships
  - Discover modular primitives
  - Help systematic generalization
  - Improve data efficiency

PartNet objects (Mo et al., 2019)

Hierarchical Scene Representations

Previous Work

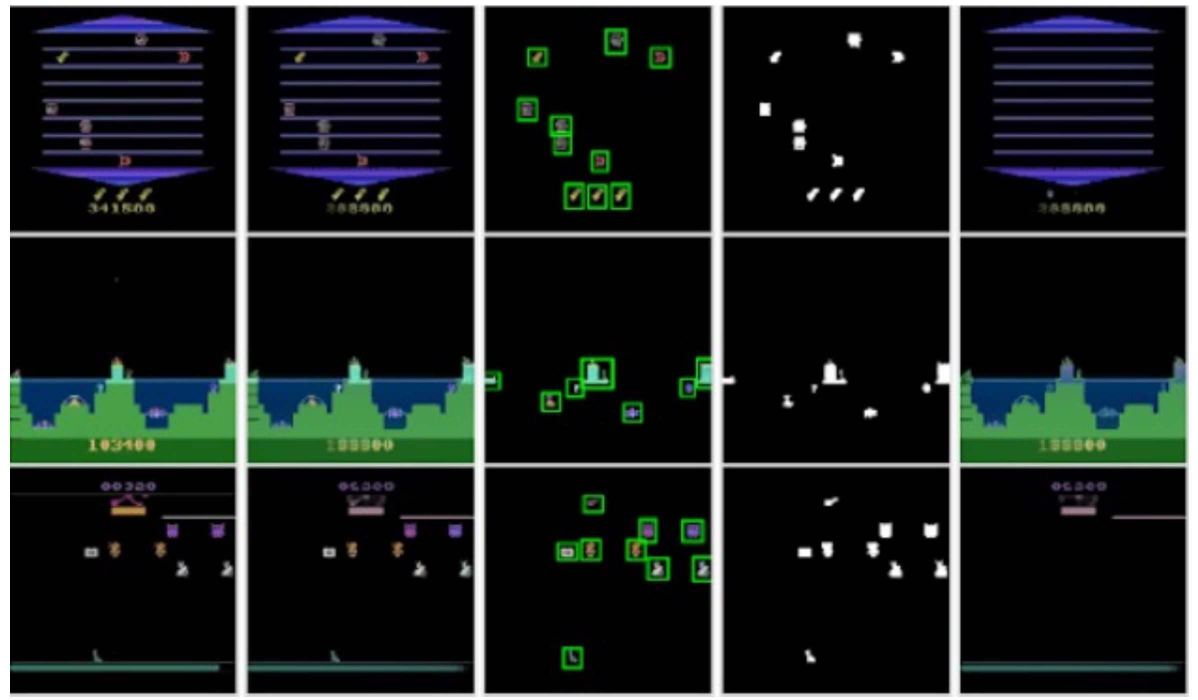
- Need supervision
  - 3D supervision
  - Part-level supervision
- Assume single object
- Inference OR generation

This Work

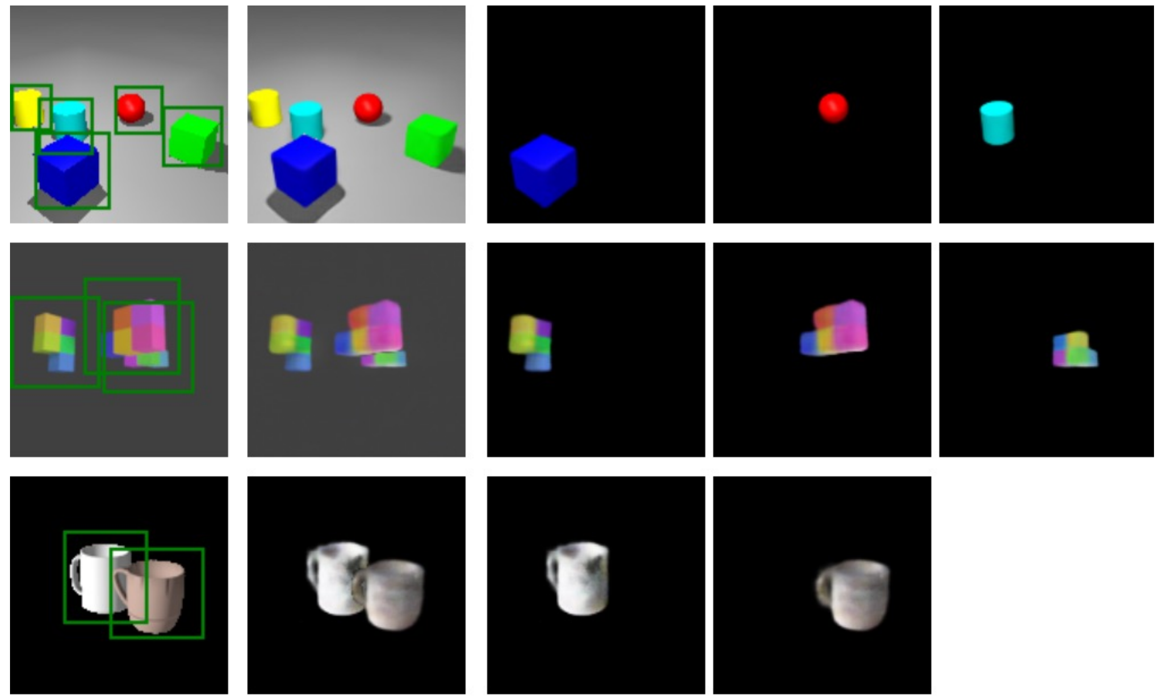
- Fully unsupervised
  - 2D image input
  - No part labels
- Multi-object scenes
- Inference AND generation

Object-Centric Representations

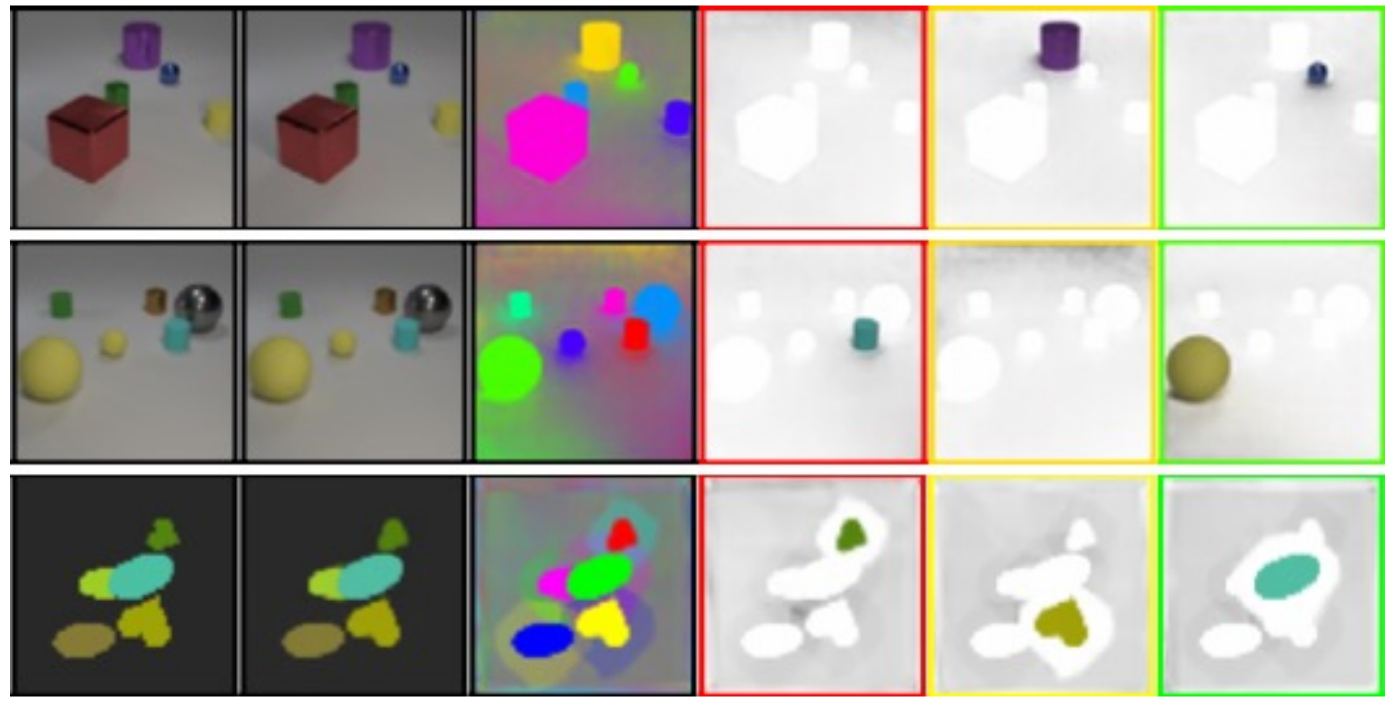
- Unsupervised object-level decomposition



SPACE (Lin et al., 2020)



ROOTS (Chen et al., 2020)

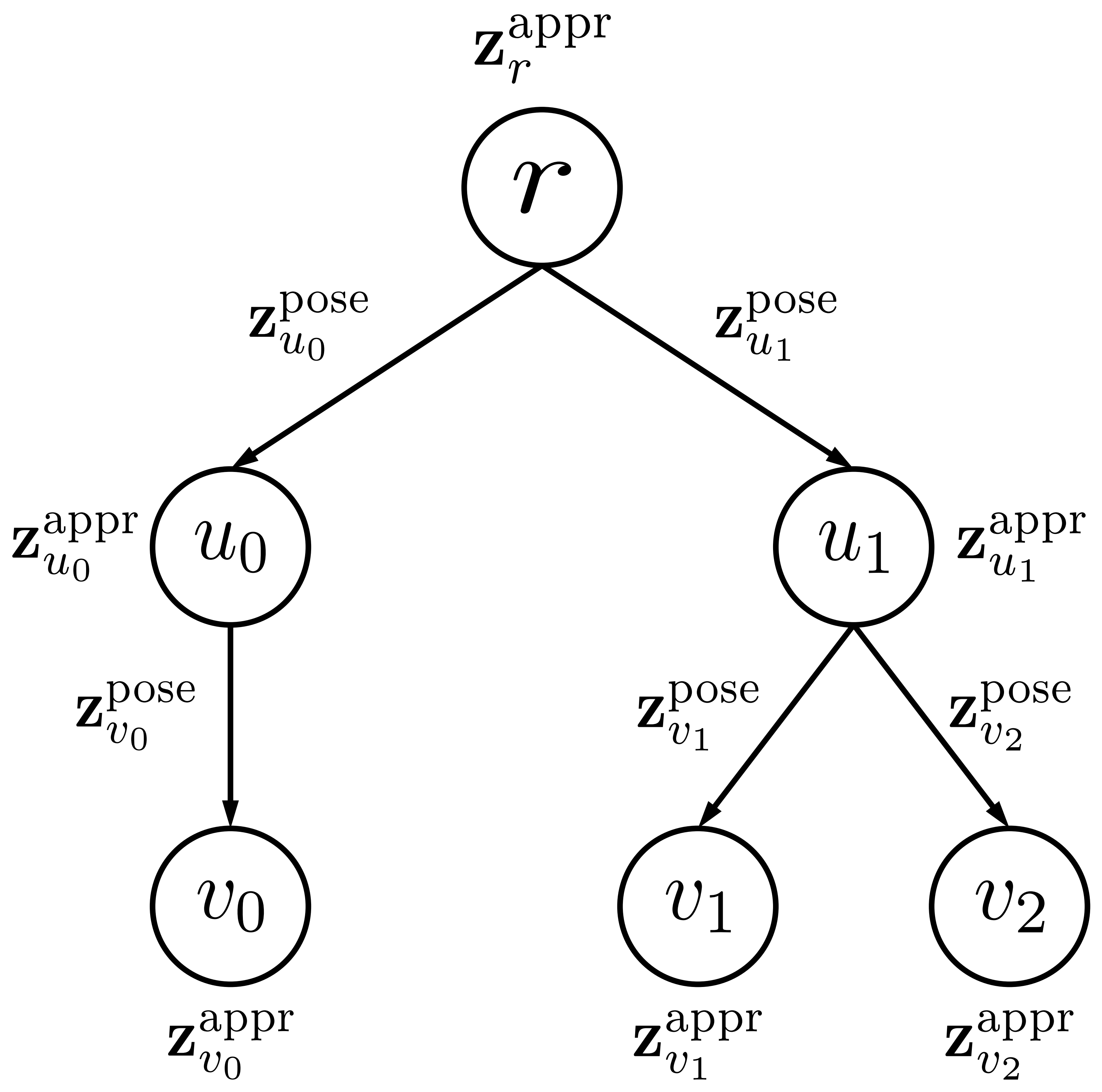


Slot Attention (Locatello et al., 2020)

GSGN: Probabilistic Scene Graph

- Nodes: entity appearance
- Edges: relative pose for composition
- Prior factorization:

$$p(\mathbf{z}_{\text{fg}}) = \underbrace{p(\mathbf{z}_r^{\text{appr}})}_{\textit{root}} \prod \underbrace{p(\mathbf{z}_v^{\text{pose}} \,|\, \mathbf{z}_{pa(v)}^{\text{appr}})\, p(\mathbf{z}_v^{\text{appr}} \,|\, \mathbf{z}_{pa(v)}^{\text{appr}})}_{\textit{parent} \to \textit{child}}$$



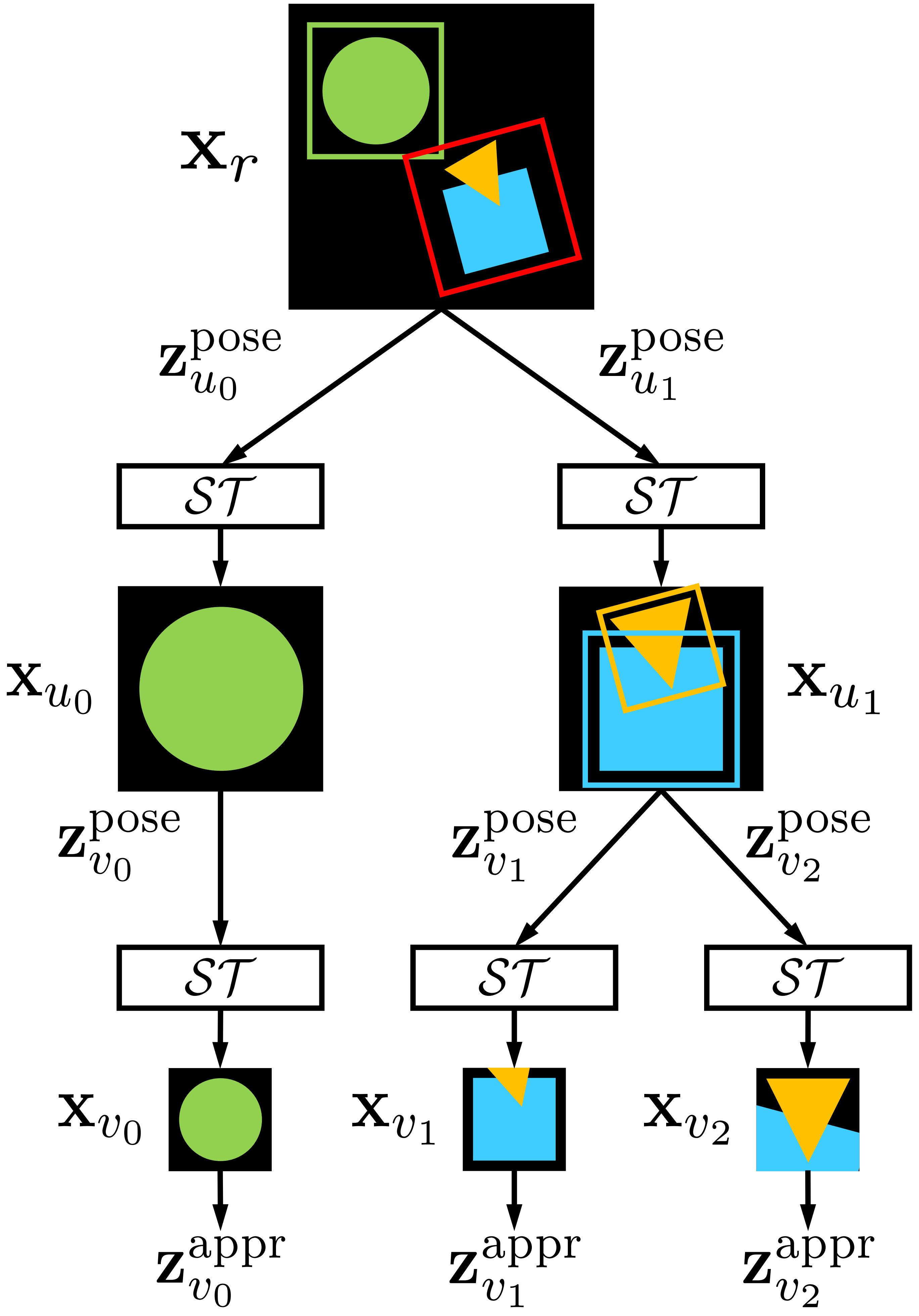
GSGN: Top-Down Inference

- First: scene → objects
- Then: object → parts

$$q(\mathbf{z}_{\text{fg}} \,|\, \mathbf{x}) = \underbrace{q(\mathbf{z}_r^{\text{appr}} \,|\, \mathbf{x})}_{\textit{root}} \prod \underbrace{q(\mathbf{z}_v^{\text{pose}}, \mathbf{z}_v^{\text{appr}} \,|\, \mathbf{z}_{pa(v)}^{\text{appr}}, \mathbf{x})}_{\textit{parent} \rightarrow \textit{child}}$$

- Use prior for guidance:

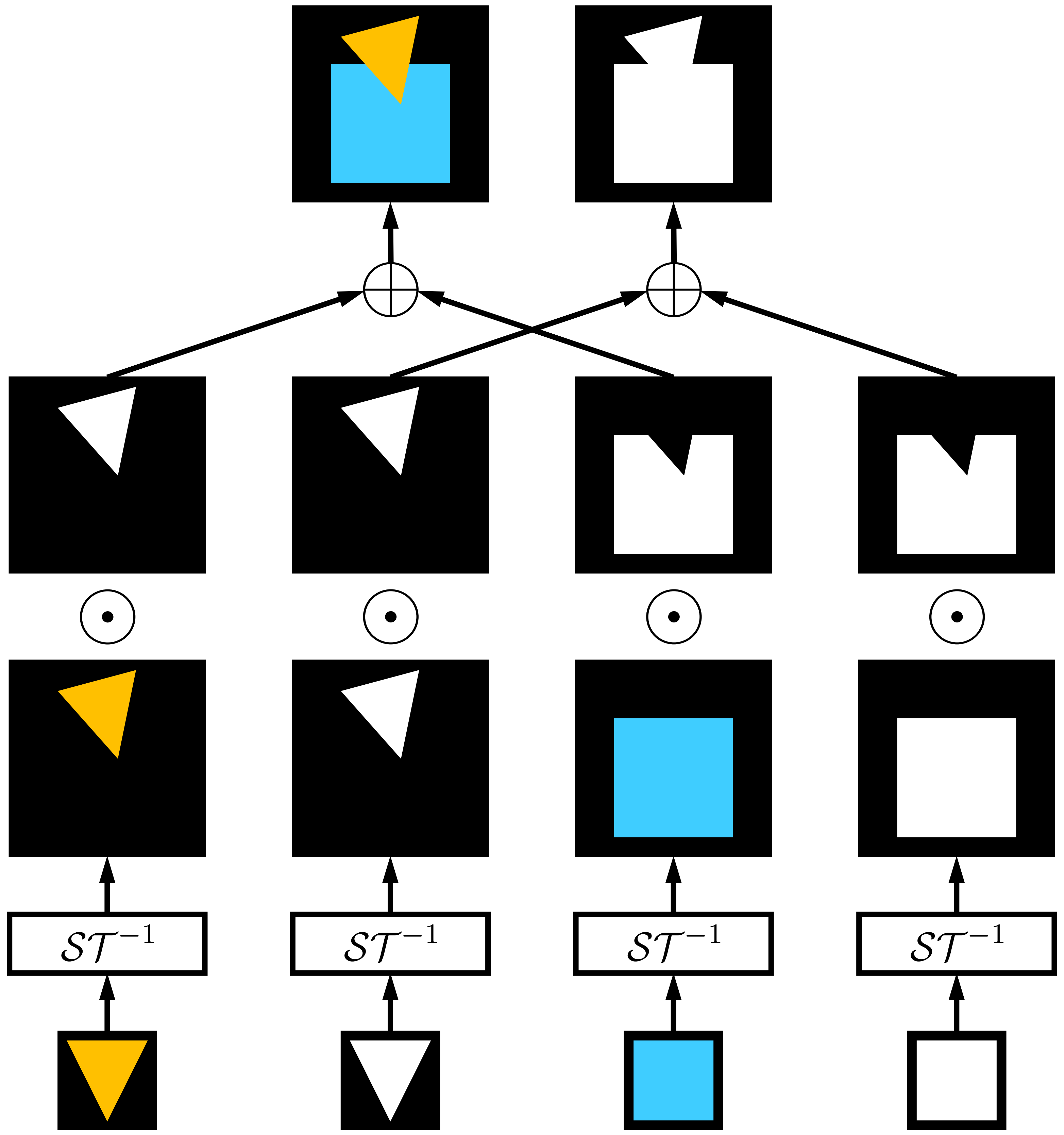
$$q(\mathbf{z}_v^{\text{appr}} \,|\, \mathbf{z}_{pa(v)}^{\text{appr}}, \mathbf{x}) \propto \underbrace{p(\mathbf{z}_v^{\text{appr}} \,|\, \mathbf{z}_{pa(v)}^{\text{appr}})}_{\textit{prior}} \underbrace{q_{\text{SPACE}}(\mathbf{z}_v^{\text{appr}} \,|\, \mathbf{x}_v)}_{\textit{SPACE}}$$



GSGN: Compositional Decoder

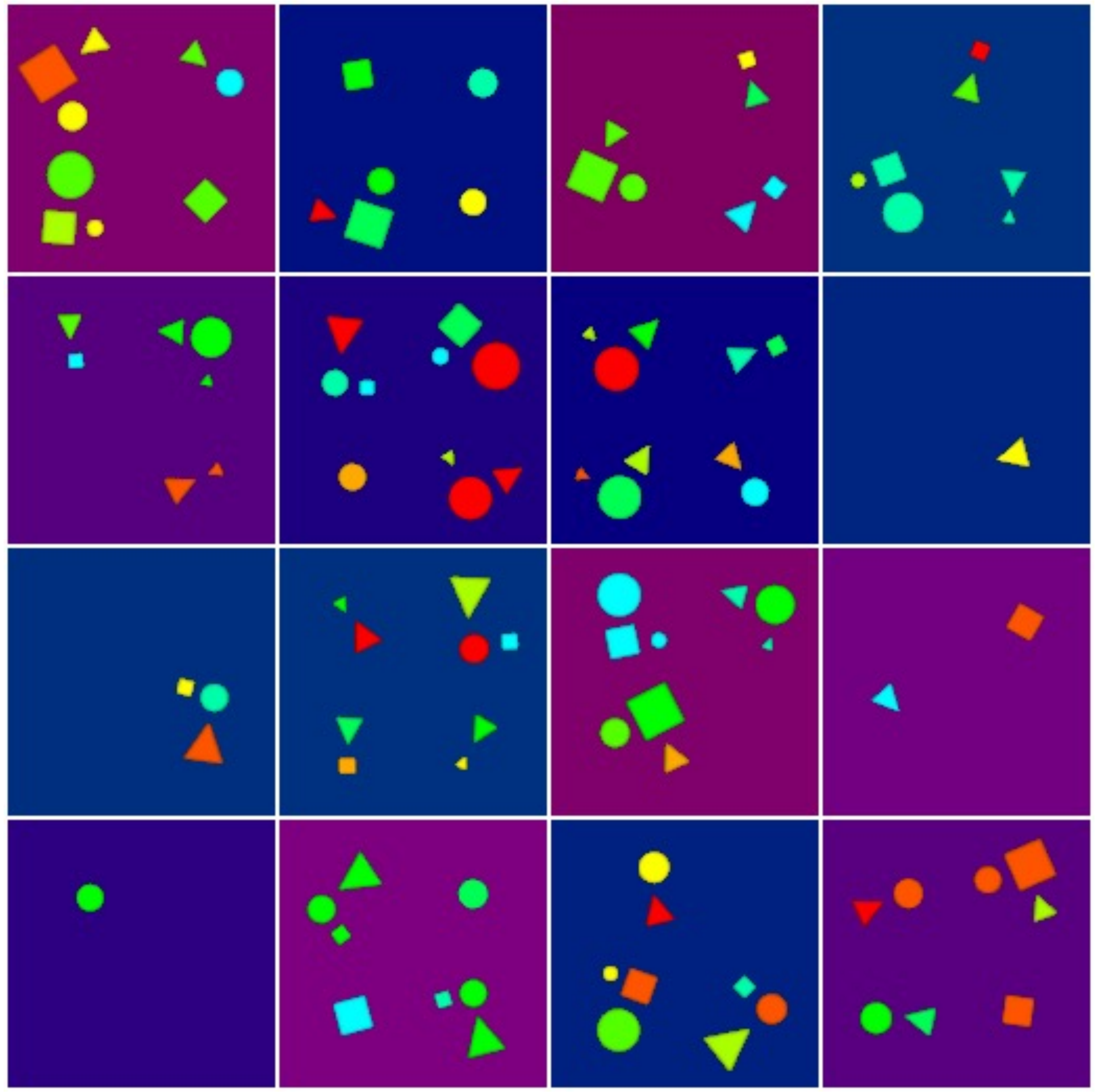
- Recursive composition
  - Coordinate transform
  - Alpha compositing

$$\hat{\mathbf{x}}_u = \sum \boldsymbol{\alpha}_v \odot \mathcal{ST}^{-1}(\hat{\mathbf{x}}_v,\ \mathbf{z}_v^{\text{pose}})$$



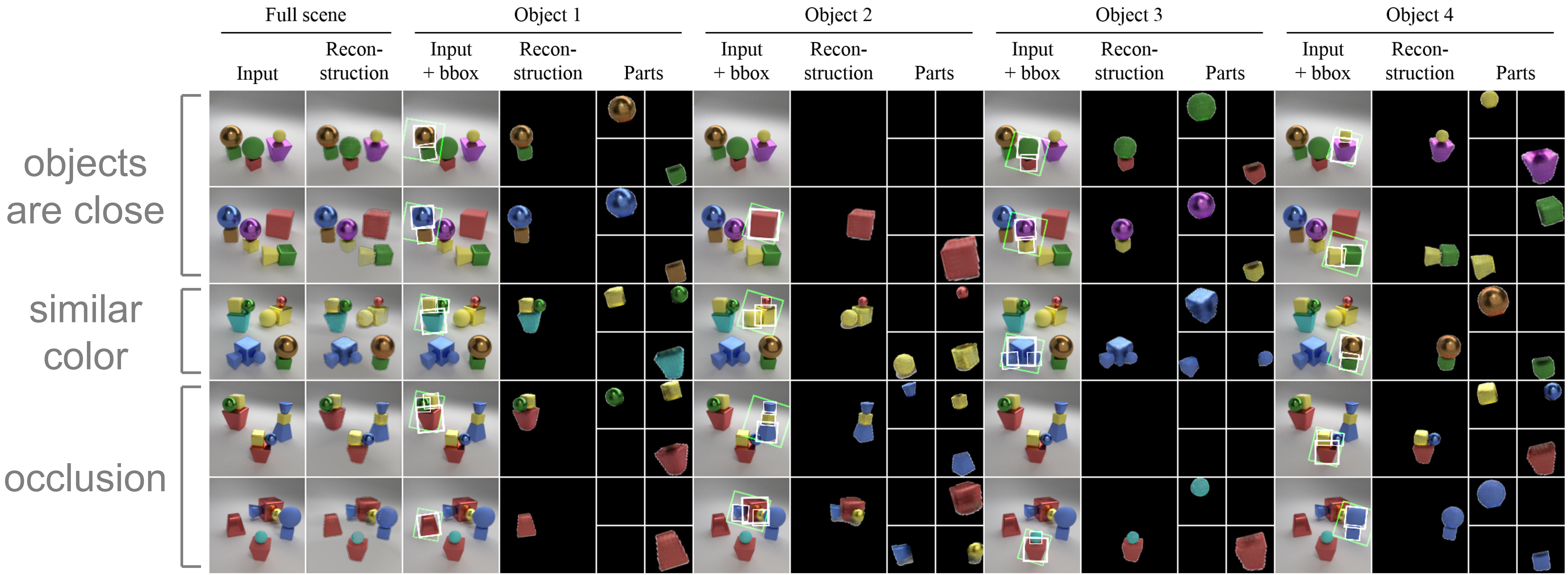
Datasets

2D Shapes



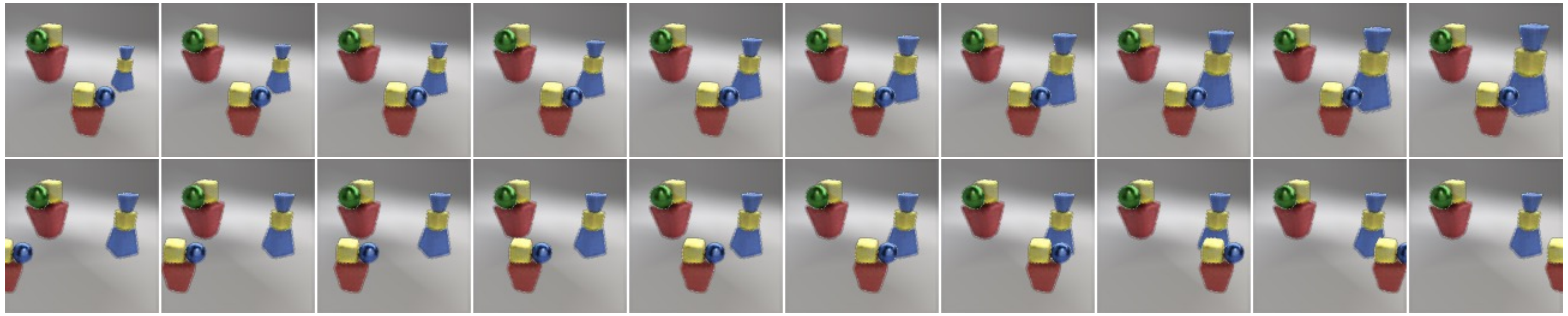
Compositional CLEVR

Scene Graph Inference

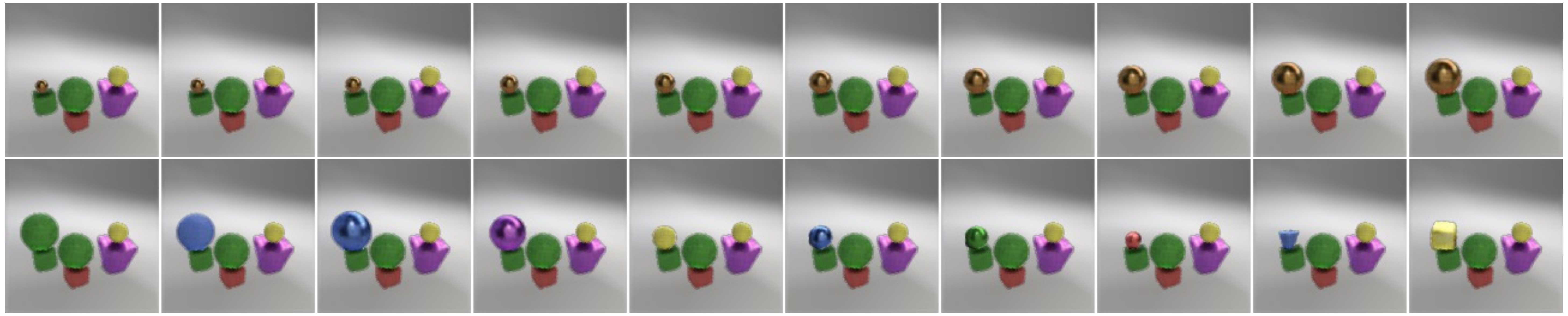


Scene Graph Manipulation

- Object-level manipulation

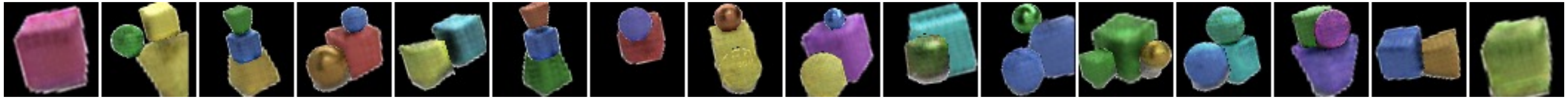


- Part-level manipulation

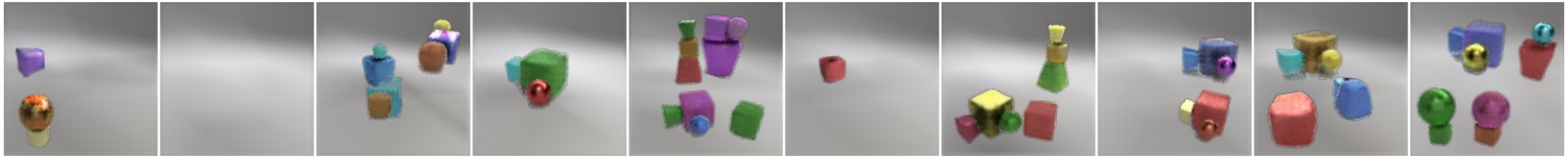


Generation from Prior

- Object generation



- Scene generation

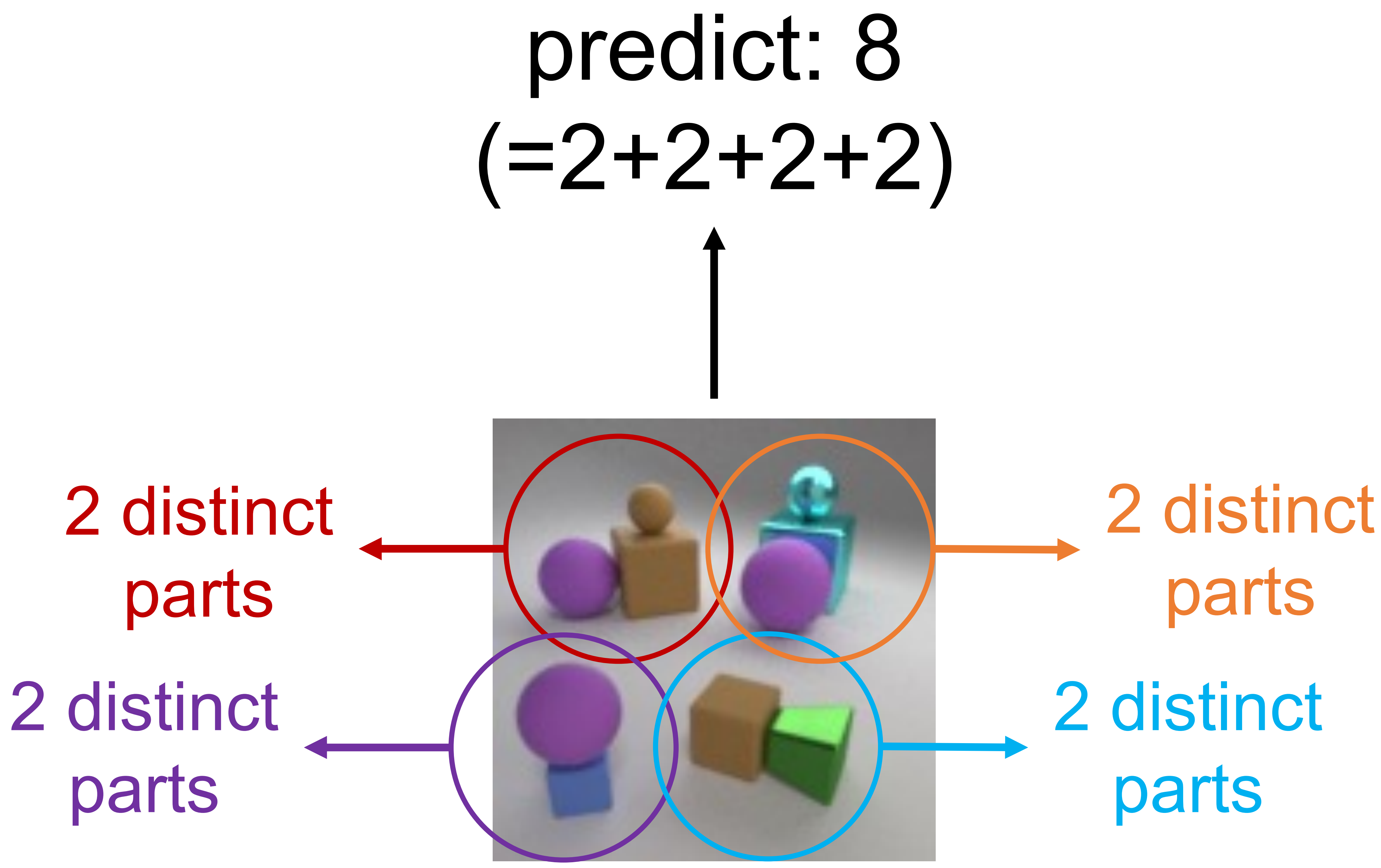


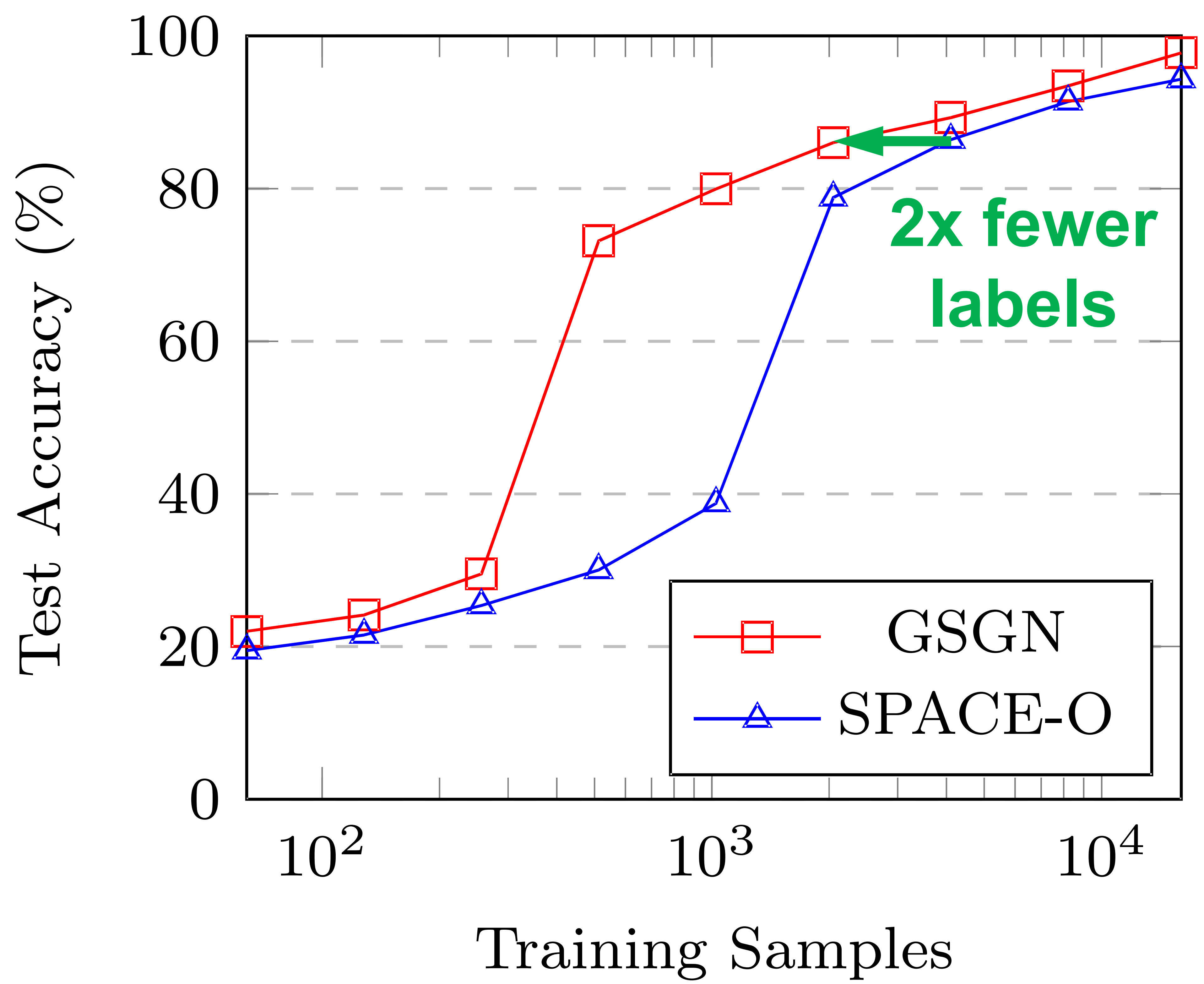
Robustness to Occlusion

Severe occlusion ↔ **Slight occlusion**

Min Visible Pixels Per Part	<100		100~200		>200	
Metric	Part Count Accuracy	Part Recall	Part Count Accuracy	Part Recall	Part Count Accuracy	Part Recall
SPACE-P	12.24%	86.03%	85.66%	97.95%	96.11%	99.48%
GSGN	95.92%	98.93%	98.33%	99.77%	98.76%	99.86%
GSGN-9	89.80%	97.35%	96.92%	97.85%	98.12%	97.62%
GSGN-No-Share	85.71%	96.34%	96.13%	99.15%	97.56%	99.46%

Data Efficiency in Downstream Tasks





Conclusion

- Unsupervised scene graph discovery from multi-object scenes

- Scene graph inference under severe occlusion
- Out-of-distribution generation
- Better data efficiency in downstream tasks