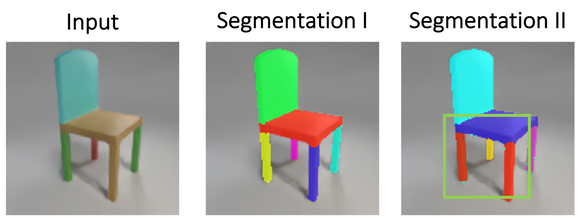
Language-Mediated, Object-Centric Representation Learning

Ruocheng Wang* Stanford University
Jiayuan Mao* MIT CSAIL
Samuel J. Gershman† Harvard University
Jiajun Wu† Stanford University
*,† indicates equal contribution

Project Page: lang-orl.github.io

Motivation



- **Language** provides cues about how a scene should be decomposed into individual objects.

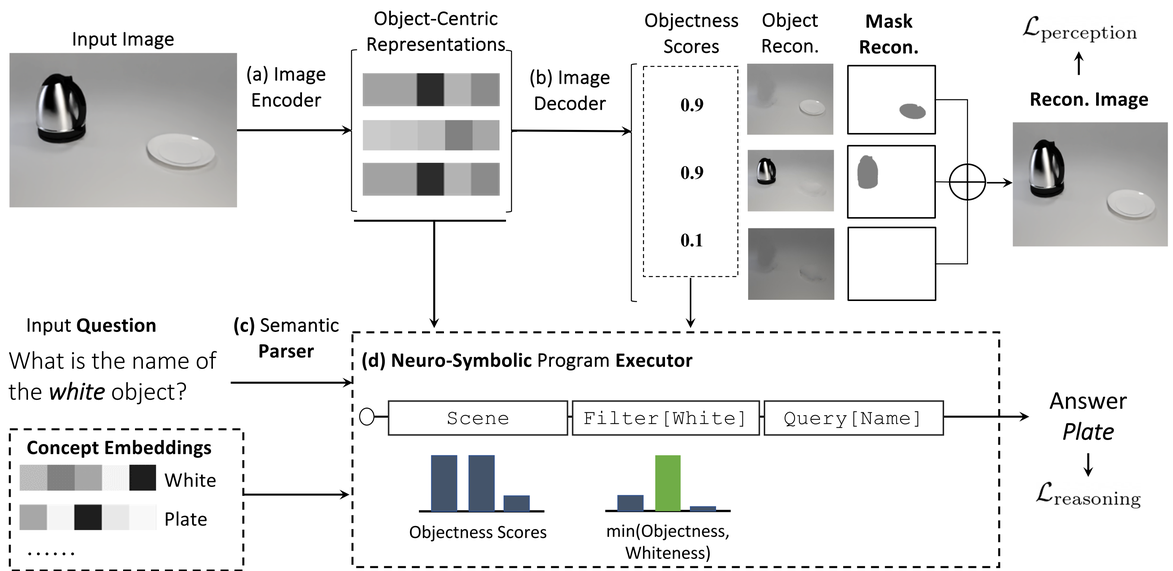
Contributions

- We present Language-mediated, Object-centric Representation Learning (LORL), a paradigm for learning disentangled, object- centric scene representations from vision and language.
- LORL consistently improves the performance unsupervised object discovery methods.
- Representations learned by LORL are useful for various downstream tasks like instance retrieval and referring expression comprehension.

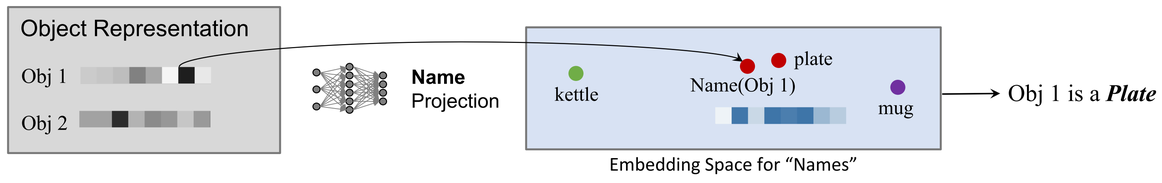
Object-Centric Representation Learning (ORL)

- Given an image, ORL aims to decompose a scene into a series of object profiles $\{(z_1, m_1), \ldots, (z_K, m_K)\}$, where z_i is the object feature, and m_i is the object mask.
- Previous ORL models only leverage **visual** cues from the image to discover individual objects.
- We study how **language** can interact with ORL.

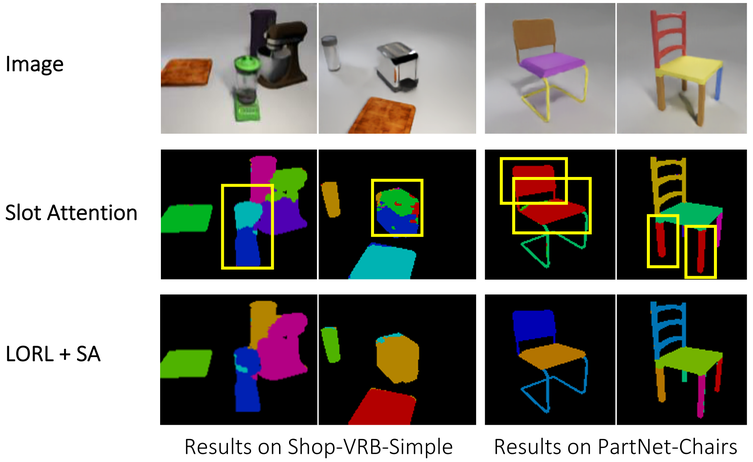
Language-Mediated, Object-Centric Representation Learning (LORL)



Neuro-Symbolic Program Executor for Query[Name]



Object Discovery Measured By Segmentation



Results on Shop-VRB-Simple | Results on PartNet-Chairs

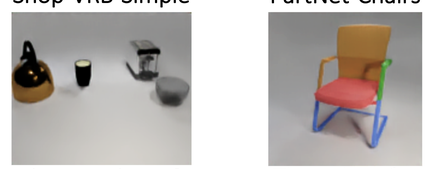
	ARI↑	GT Split↓	Pred Split↓
SA	83.51	15.68	13.19
LORL + SA	**89.23**	**9.95**	**10.18**

Results on Shop-VRB-Simple

	ARI↑	GT Split↓	Pred Split↓
SA	87.32	12.54	22.99
LORL + SA	**95.81**	**3.39**	**2.92**

Results on PartNet-Chairs

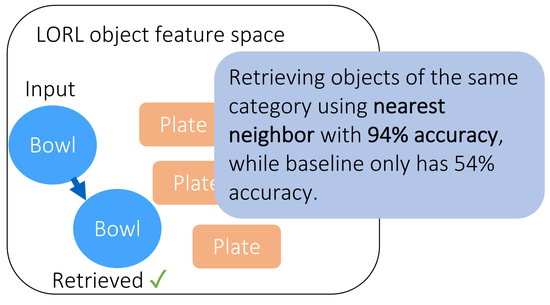
Datasets Examples



Segmentation Metrics

ARI ↑: Standard metric following prior work, treats segmentation as clustering of pixels.
GT Split↓ : Measures the ratio of objects that are covered by more than one prediction mask.
Pred Split↓: Measures the ratio of prediction masks that cover more than one object.

Instance Retrieval using Object Rep.



Referring Expression Comprehension

Given an expression referring to a set of objects in the scene, such as "the white plates", the model should return the corresponding object masks.

	Supervision	Recall@0.5
LORL+SA	No further training	84.4
IEP-Ref	Fully Supervised	90.1

Results on Shop-VRB-Simple