COSMOS: Neurosymbolic Grounding for Compositional World Modeling

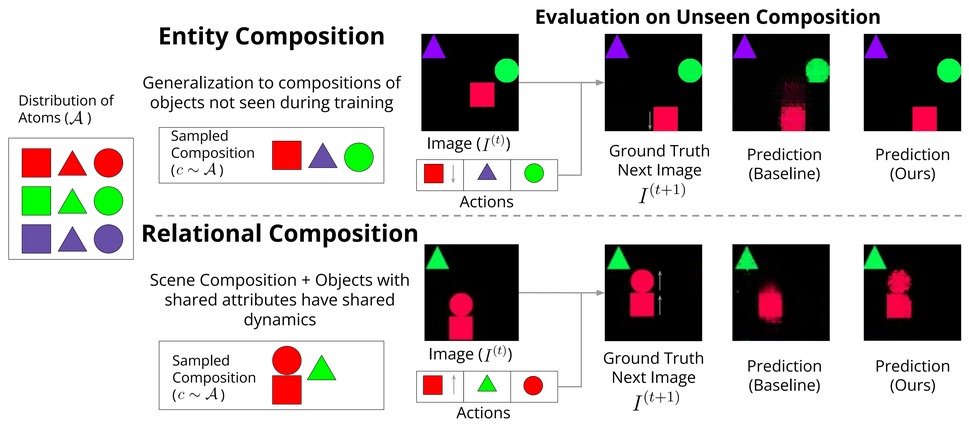
Atharva Sehgal[1], Arya Grayeli[1], Jennifer J Sun[2], Swarat Chaudhuri[1]
[1]UT Austin [2]Caltech

https://bit.ly/cosmos-wm
Website / Arxiv / Code

Problem

Learning object-oriented world models that compositionally generalize

Types of Compositions



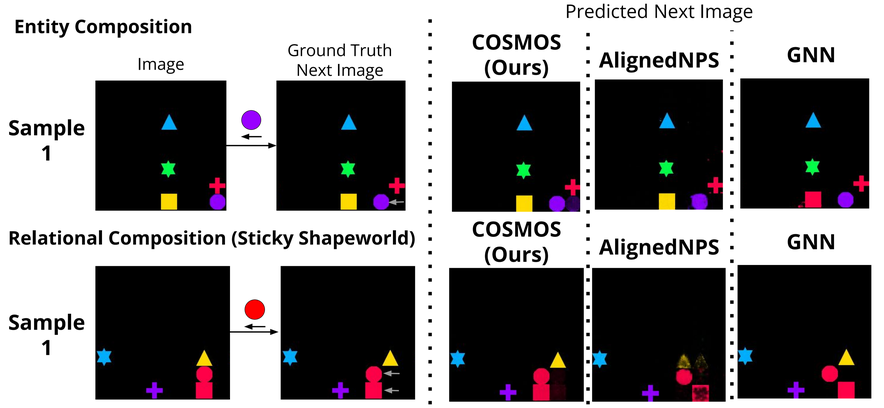
Results

Takeaways:

- Relational Composition is more challenging than Entity Composition.
- Small performance degradation from 3 objects to 5 objects.
- Cosmos achieves best next state reconstruction (MSE).
- Cosmos auto-encoder reconstruction (AE-MSE) is consistently good.
- Eq-MRR is misleading
  - Accurate MRR requires consistently good AE-MSE

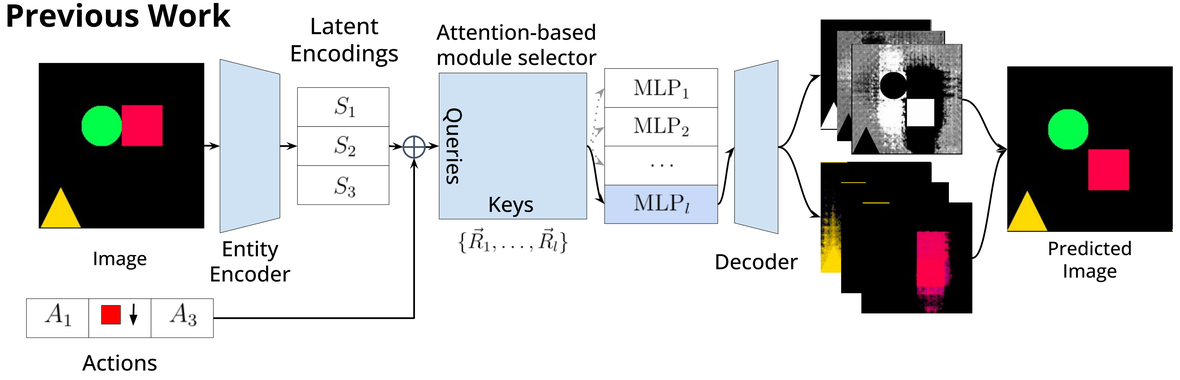
		3 objects			5 objects		
Dataset	**Model**	**MSE ↓**	**AE-MSE ↓**	**Eq.MRR ↑**	**MSE ↓**	**AE-MSE ↓**	**Eq.MRR ↑**
RC (Sticky)	COSMOS	**4.23E-03**	**4.90E-04**	**1.20E-01**	**4.15E-03**	**1.68E-03**	**3.67E-01**
	ALIGNEDNPS	1.14E-02	7.72E-03	8.01E-02	6.07E-03	2.47E-03	3.62E-01
	GNN	7.94E-03	5.11E-03	6.03E-04	6.21E-03	2.73E-03	5.30E-04
RC (Team)	COSMOS	**4.60E-03**	**4.33E-04**	1.04E-01	**5.53E-03**	1.86E-03	2.86E-01
	ALIGNEDNPS	1.24E-02	8.36E-03	**1.75E-01**	9.64E-03	3.12E-03	**2.93E-01**
	GNN	8.92E-03	3.82E-03	7.16E-04	7.01E-03	**1.62E-03**	5.46E-04
EC	COSMOS	**7.66E-04**	**6.34E-05**	2.99E-01	**4.08E-04**	**2.92E-06**	3.03E-01
	ALIGNEDNPS	3.51E-03	2.69E-03	2.97E-01	2.45E-03	1.22E-03	3.19E-01
	GNN	9.89E-03	1.03E-02	**5.50E-01**	1.20E-02	1.28E-02	**5.25E-01**

Table 1: Evaluation results on the 2D block pushing domain for entity composition (EC) and relational composition (RC) averaged across three seeds. We report next-state reconstruction error (MSE), autoencoder reconstruction error (AE-MSE), and the equivariant mean reciprocal rank (Eq.MRR) for three transition models: our model (COSMOS), an improved version of Goyal et al. (2021) (ALIGNEDNPS), and a reimplementation of Zhao et al. (2022) (GNN). Our model (COSMOS) achieves best next-state reconstructions for all datasets.



Algorithm

Previous Work



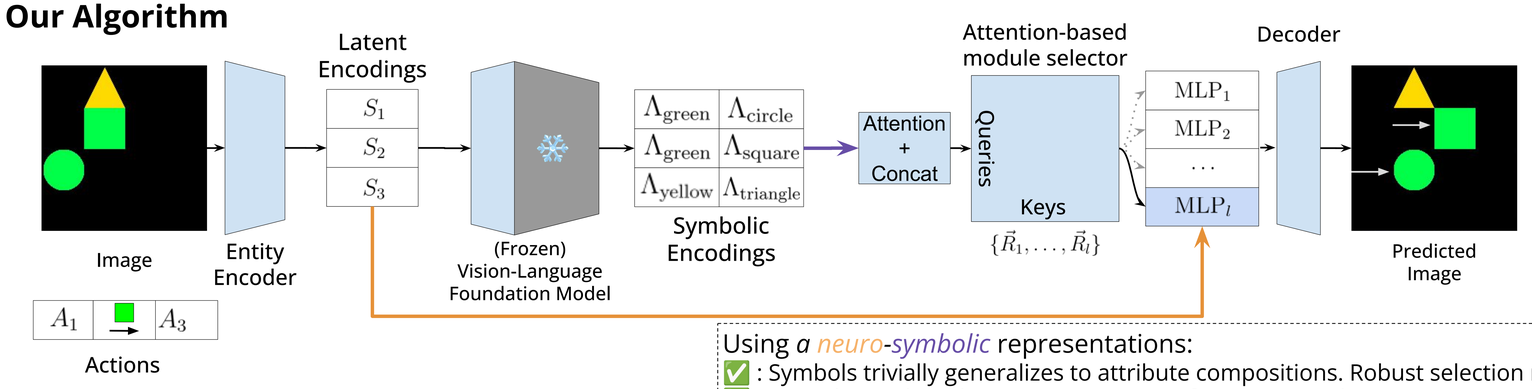
Works on Entity Composition

- Decompose into object-centric encodings (*slots*)
- Model interactions between slots using *modules*

Fails on Relational Composition

- Decomposing into slots doesn't capture shared attributes => cannot model shared dynamics.

Our Algorithm



Using *a neuro-symbolic* representations:
✅ : Symbols trivially generalizes to attribute compositions. Robust selection module!
✅ : Neural encodings learn very expressive features. Good image reconstructions!

Takeaways

- **Explicit symbolic knowledge** helps with compositionality
- **Extend**, rather than replace, deep representations
- **Foundation models** over language (and code) give symbols for free.