COAT: Measuring Object Compositionality in Emergent Representations

Sirui Xie[1], Ari Morcos[2], Song-Chun Zhu[1], Ramakrishna Vedantam[2]

[1]*UCLA*, [2]*FAIR*

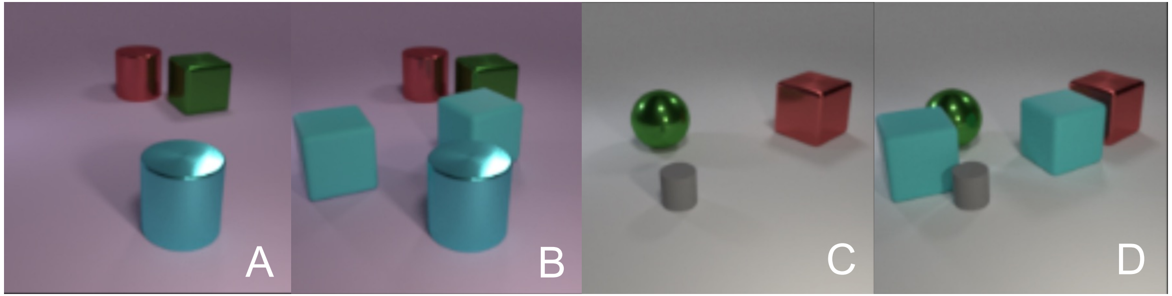
FACEBOOK AI

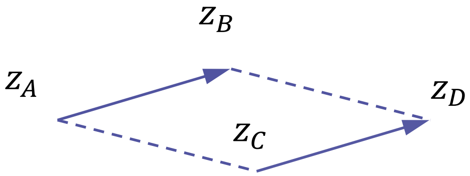
Motivation

- Object-centric representation is desirable for reasoning and planning
- Object mask-based metrics are built upon pixel space, and can only evaluate generative models with “slot” structures
- We measure object compositionality in the representation space

Key Idea

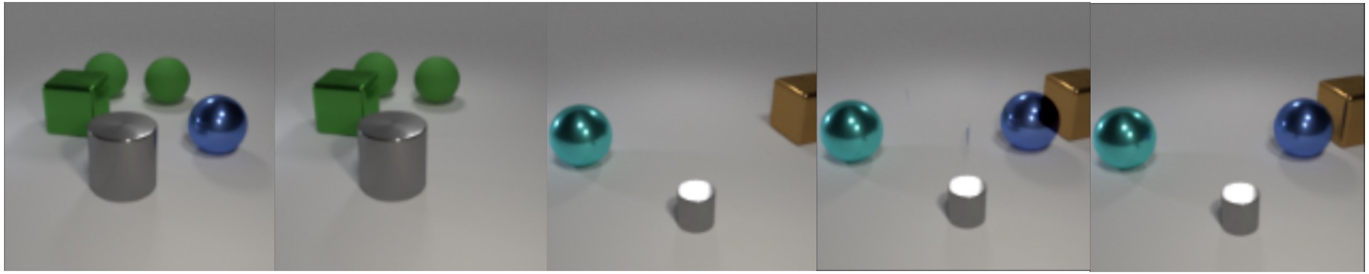
To evaluate if a generic representation exhibits certain **geometric properties**.





Shortcuts

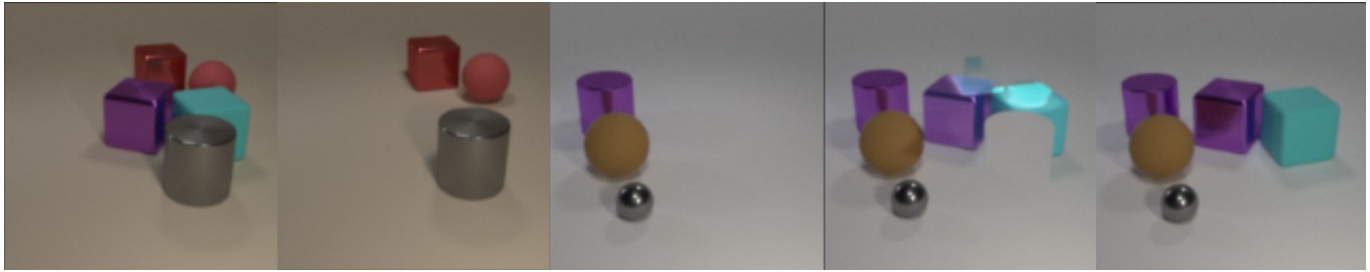
Pixel representation $q(\mathbf{x})=\mathbf{x}$ can be **trivially compositional** in weakly occluded scenes.



B $-$ **A** $+$ **C** $=$ **D'** $\approx$ **D**

Shortcuts

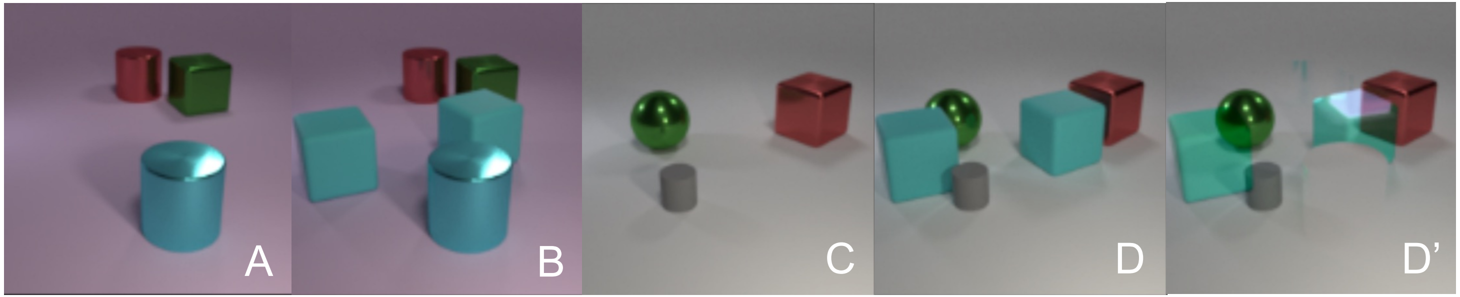
Object compositionality is only non-trivial when transformations induce strong occlusion.

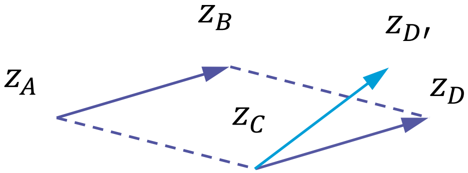


$$B \quad - \quad A \quad + \quad C \quad = \quad D' \quad \neq \quad D$$

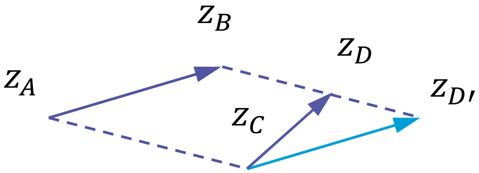
Hard negatives

Null hypothesis: the parallelogram holds for A, B, C and D'.





Compositional representation



Non-compositional representation

Compositional Object Algebra Test (COAT)

Parallelogram test

- $\mathcal{L}(\mathbf{z}_A, \mathbf{z}_B, \mathbf{z}_C, \mathbf{z}_D) = ||\mathbf{z}_B - \mathbf{z}_A + \mathbf{z}_C - \mathbf{z}_D||$

Shortcut detection

- $P\big(\mathcal{L}(\mathbf{z}_A, \mathbf{z}_B, \mathbf{z}_C, \mathbf{z}_D) < \mathcal{L}(\mathbf{z}_A, \mathbf{z}_B, \mathbf{z}_C, \mathbf{z}_{D'})\big) > 0.5$

Normalization and correction for chance

- $\mathrm{COAT} = 1 - \frac{\mathcal{L}(\mathbf{z}_A, \mathbf{z}_B, \mathbf{z}_C, \mathbf{z}_D)}{\mathrm{E}_{\hat{D}}[\mathcal{L}(\mathbf{z}_A, \mathbf{z}_B, \mathbf{z}_C, \mathbf{z}_{\hat{D}})]}$

Experiments

Table 1. Models, their inductive biases, their training paradigms, their training sets, and their performance on ARI and COAT. "HN" is the Hard Negative Test; models need to pass all hard negative tests to obtain a COAT score, otherwise it is indicated with "-". Since representations directly obtained from slot attention do not perform well on the COAT metric, we also tried some post-processing: * indicates duplication removal, †indicates removing "invisible slots" with zero mask weights. (w/o occlusion) and (w/ occlusion) indicate non-occluded (Figure 2a) and strongly occluded (Figure 2b) test cases. Statistics are Mean and SEM summarized over 5 random seeds.

	Slot Structure $\mathbf{z} = [\mathbf{z}_0, \cdots, \mathbf{z}_K]$	Factorized Prior $p(\mathbf{z}) = \Pi_{k=1}^{K} p(\mathbf{z}_k)$	Training Paradigm	Train Set	ARI (%)	HN L_2	L_2-COAT (%)	HN acos	*acos*-COAT (%)
Pixel baseline (w/ occlusion)	N/A	N/A	N/A	N/A	N/A	N/A	75.47	N/A	36.28
Pixel baseline (w/o occlusion)	N/A	N/A	N/A	N/A	N/A	N/A	97.18	N/A	73.17
Auto-encoder(w/ occlusion)	No	No	Generative	IID	N/A	Fail	-	Fail	-
β-TC-VAE (w/ occlusion)	No	Yes	Generative	IID	N/A	Fail	-	Fail	-
Slot attention (w/ occlusion)	Yes	No	Generative	IID	95.53 ± 1.84	Pass	48.55 ± 14.11	Pass	21.53 ± 10.73
Slot attention* (w/ occlusion)	Yes	No	Generative	IID	95.53 ± 1.84	Pass	60.70 ± 15.55	Pass	31.18 ± 8.01
Slot attention*†(w/ occlusion)	Yes	No	Generative	IID	95.53 ± 1.84	Pass	77.07 ± 0.72	Pass	43.12 ± 0.78
Slot attention*†(w/o occlusion)	Yes	No	Generative	IID	95.53 ± 1.84	Pass	83.84 ± 6.23	Pass	47.45 ± 4.34
Slot attention*†(w/ occlusion)	Yes	No	Generative	CORR	69.12 ± 9.34	Pass	64.82 ± 9.20	Pass	31.95 ± 7.21
IODINE (w/ occlusion)	Yes	Yes	Generative	IID	92.21 ± 0.15	Pass	47.52 ± 0.29	Pass	16.33 ± 0.33
IODINE (w/ occlusion)	Yes	Yes	Generative	CORR	40.08 ± 8.90	Fail	-	Pass	9.16 ± 1.08
MoCo v2 ConvNet (w/ occlusion)	No	N/A	Discriminative	IID	N/A	Fail	-	Pass	14.05 ± 1.25

Thank you!