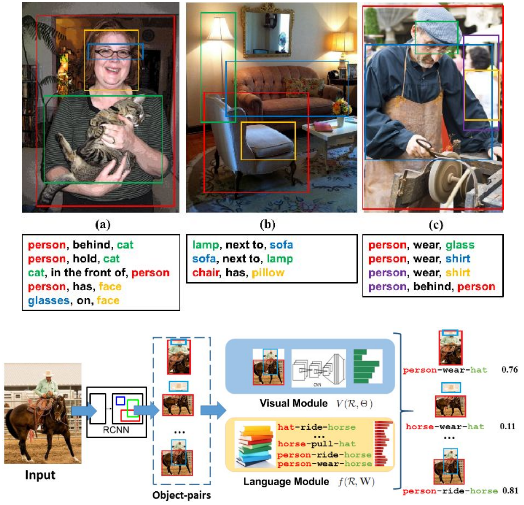
RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning

Xiaojian Ma[1], Weili Nie[2], Zhiding Yu[2], Huaizu Jiang[3], Chaowei Xiao[2,4], Yuke Zhu[2,5], Song-Chun Zhu[1], Anima Anandkumar[2,6]

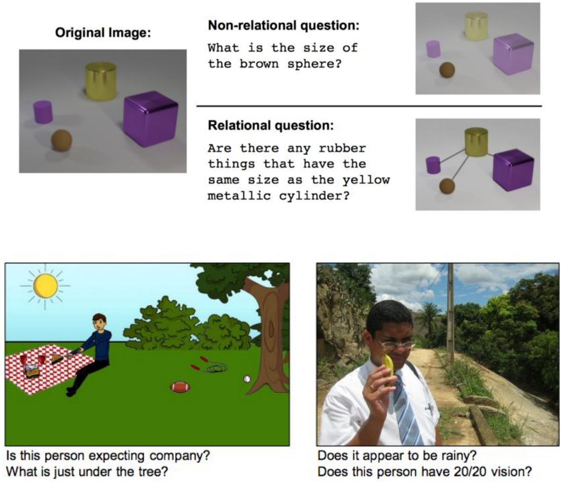
[1]UCLA [2]NVIDIA [3]Northeastern University [4]ASU [5]UT Austin [6]Caltech

Motivation

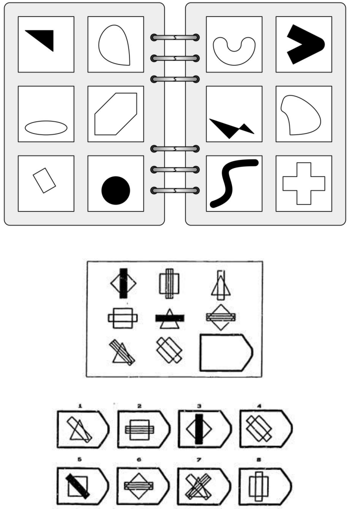
Visual Relational Reasoning: the niche of visual intelligence!



Visual Relationship Recognition



Visual Question Answering

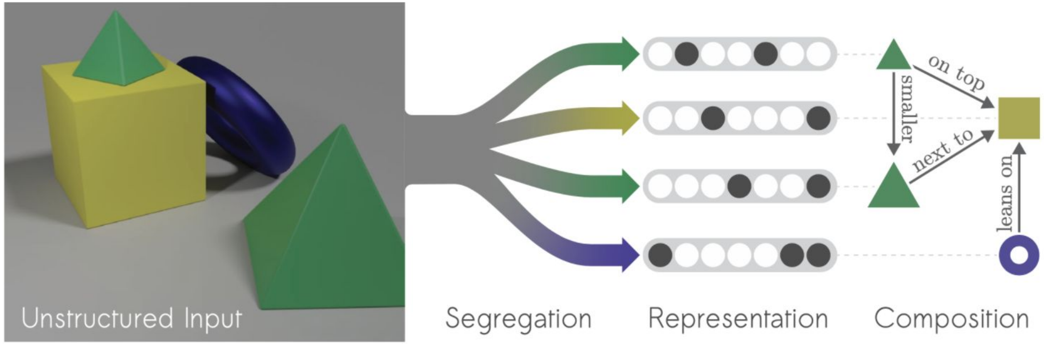


Abstract Visual Reasoning

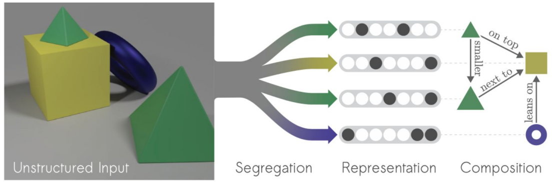
Motivation

What makes Visual Relational Reasoning so challenging?

-> How do our humans perform visual reasoning?

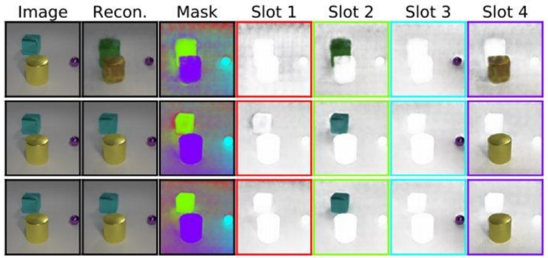


On the Binding Problem in Artificial Neural Networks, *arXiv*, 2020

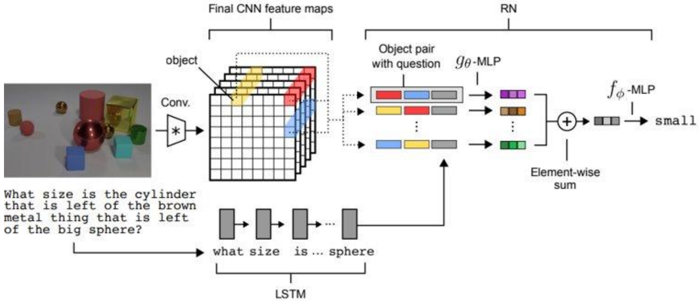


Motivation

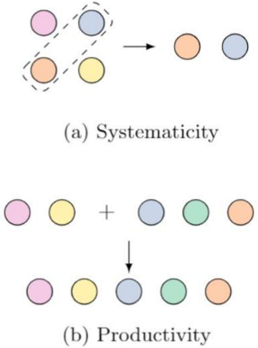
- What makes Visual Relational Reasoning so challenging?



Object-centric (disentangled) representations



Relational inductive bias

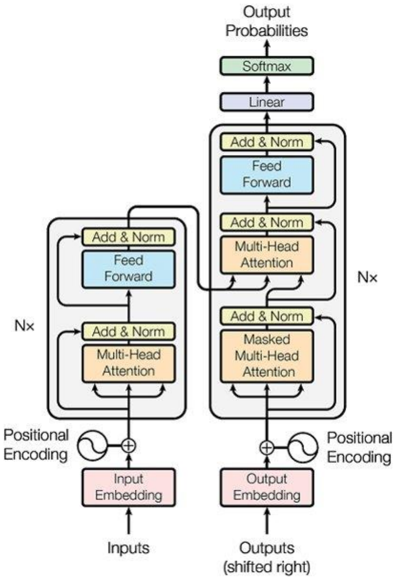


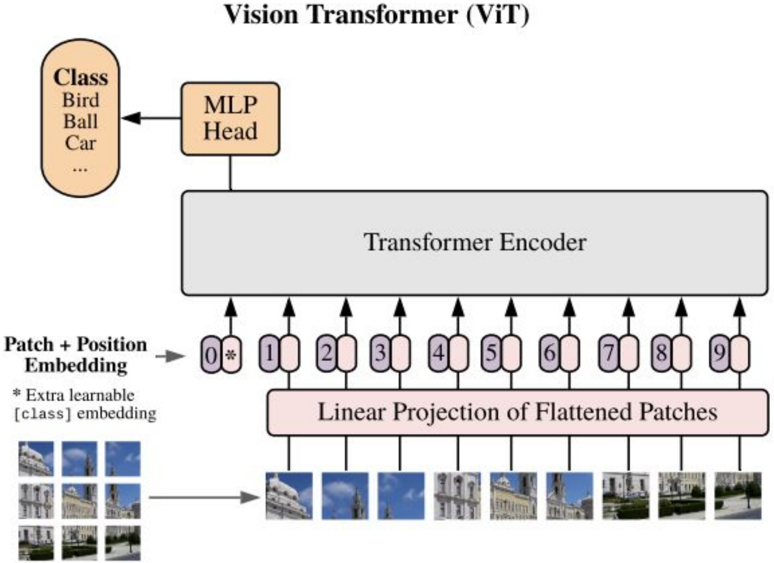
Systematic generalization

Object-Centric Learning with Slot Attention, in NeurIPS, 2020
A simple neural network module for relational reasoning, in *NeurIPS*, 2017
Compositionality decomposed: how do neural networks generalise?, in *JAIR*, 2020

Background

Transformers and Vision transformers





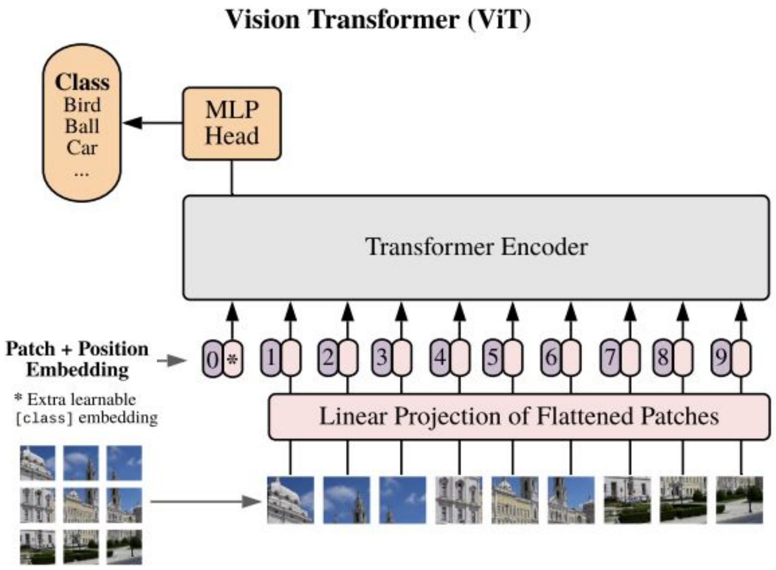
Transformer: explicitly capture the **pairwise relations** among input entities.

Vision transformers: **image patches (object candidates)** as input entities.

Background

Given the appealing nature of Vision transformers (ViTs) on **object-centric learning** and **relational inductive bias**, we choose to start with this model and see if we can make it better.

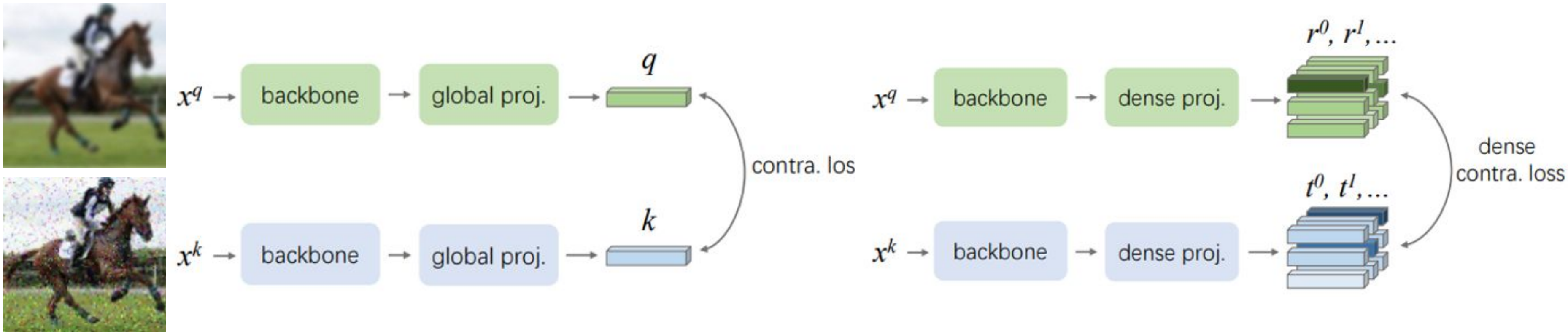
We propose to use **self-supervised contrastive learning** to achieve this goal.



Vision transformers: **image patches (object candidates)** as input entities.

Background

Contrastive learning (CL) tasks for ViT



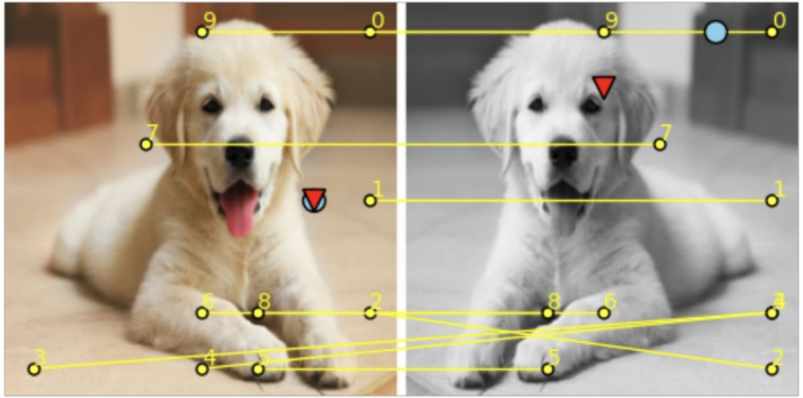
Global CL: Contrasting the **global features** of input images.

Local CL: Contrasting the **(spatially) local image features** of the input images.

Dense Contrastive Learning for Self-Supervised Visual Pre-Training, in *CVPR*, 2021

Background

How can vision transformer benefit from contrastive learning?



Global CL: Contrasting the **global descriptions** of input images.
-> boosting **relational meaning and reasoning** via instance contrasting

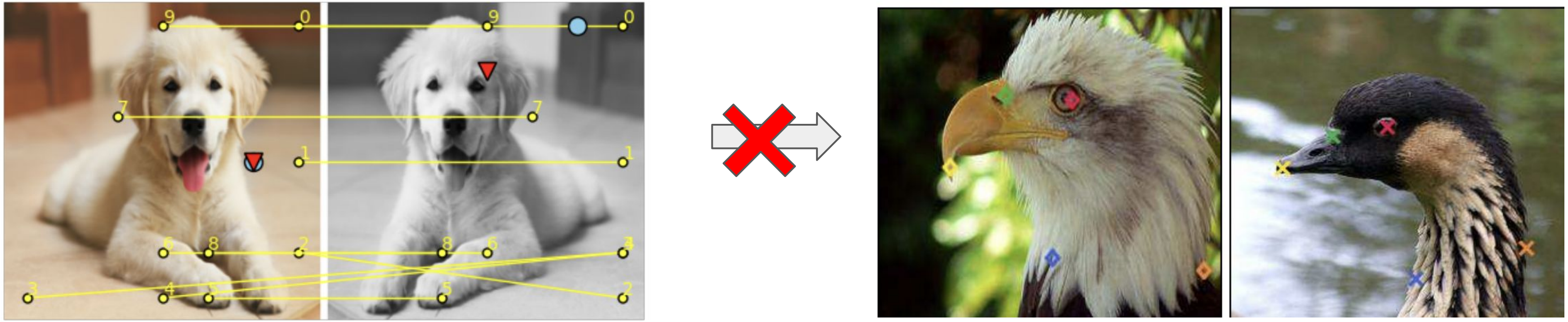
Local CL: Contrasting the **(spatially) local descriptions** of the input images.
-> boosting **object-centric representation** via unsupervised **correspondence learning**

Background

An issue: the original global and local CL are *concepts/semantics-free* -- image are treated as **isolated** samples. Therefore, these CL methods will promote:

-representations that **fail** to capture the **semantic similarity** of different objects

-relational deduction that **fails** to exploit these **semantics** for more efficient / lifted reasoning.



RelViT

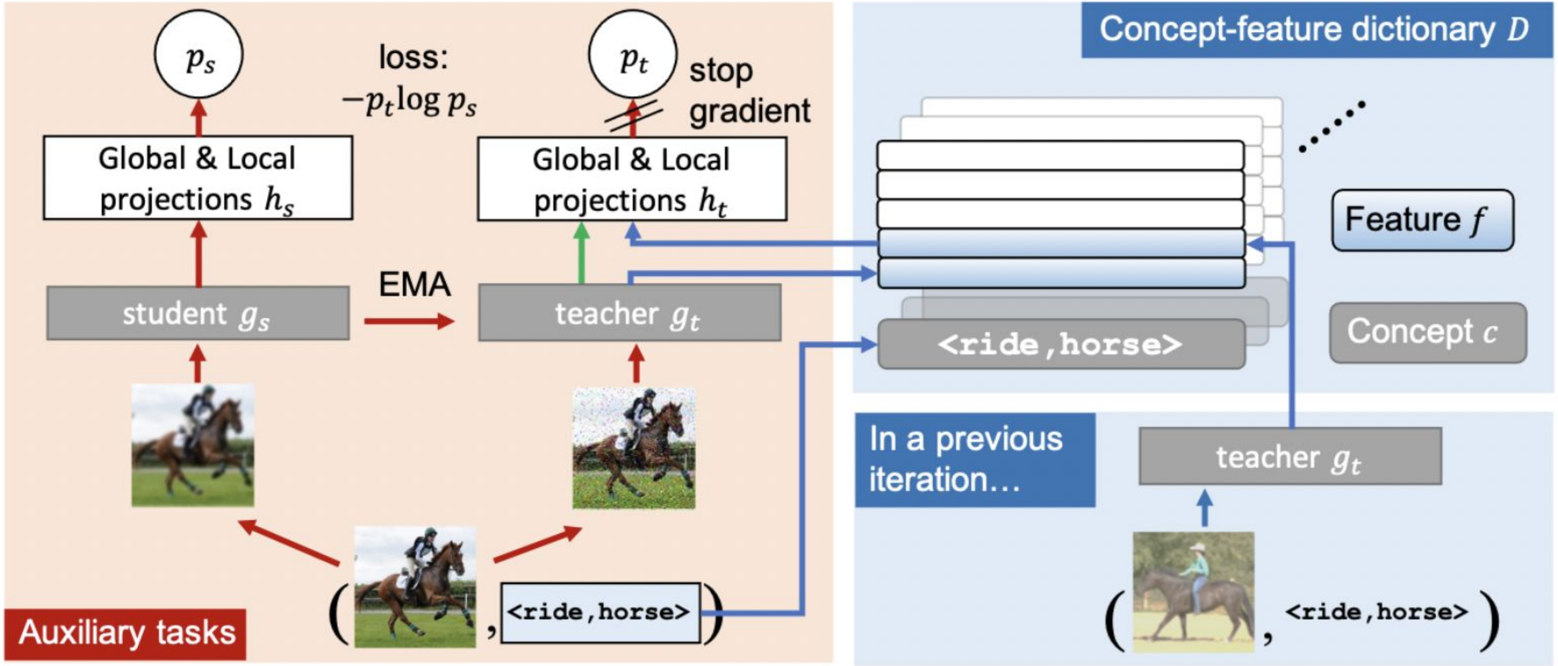


Figure 1: An overview of our method. Red+Green: the learning pipeline of DINO (Caron et al., 2021) and EsViT (Li et al., 2021); Red+Blue: our pipeline.

RelViT

- Concept-guided Vision Transformer

```
// Pseudo code
Input: image x, concept c
x1, x2 = aug1(x), aug2(x)
f1, f2 = backbone(x1), backbone(x2)
fn2 = dequeue_n_enqueue(f2, c)
loss_global = DINO_GLOBAL(f1[0], fn2[0]) // f*[0] is [CLS]
loss_local = DINO_LOCAL(f1[1:], fn2[1:])
(loss_local + loss_global).backward()
```

Experiments

We evaluate RelViT on two datasets:

-**HICO**: Human-object-interaction recognition

Formula: `I => <object, interaction>`

-**GQA**: Relational visual question answering

Formula: `<Q, I> => answer category`

A1. Is the tray on top of the table black or light brown? light brown
A2. Are the napkin and the cup the same color? yes

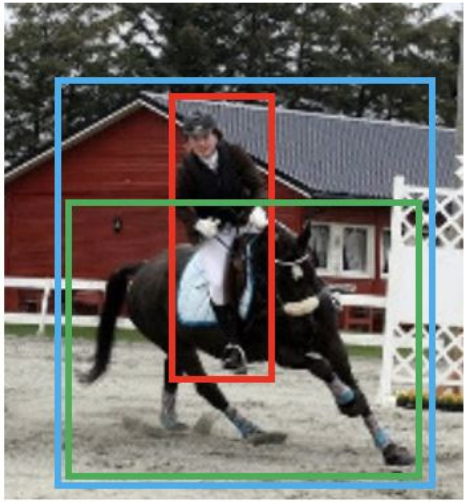
Experiments

Concept in HICO

-> HOI category (#concepts=600)

-> Interaction category (#concepts=117)

-> Object category (#concepts=80)

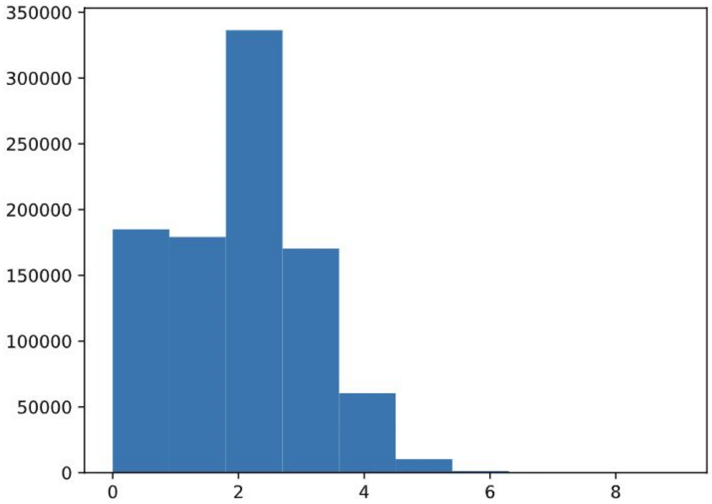


<Ride, Horse>

Experiments

Concept in GQA

We propose to parse the question into concept tokens.



#concept per question

Top 10 concepts

man 39540

animal 34279

furniture 29982

white 22728

woman 19550

vehicle 19487

black 16639

person 15906

shirt 15418

table 12999

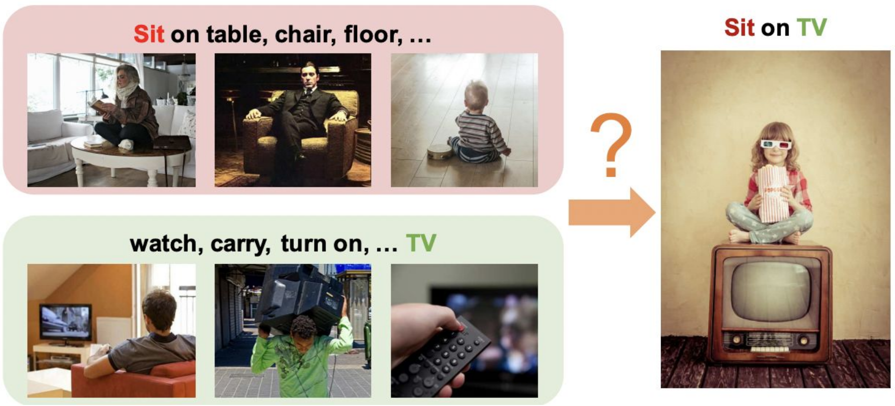
Experiments

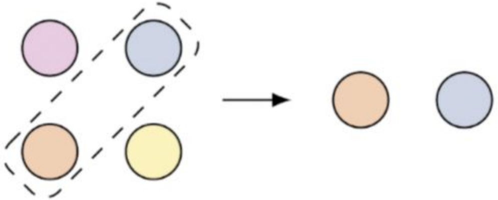
Systematic generalization test for HICO:

-We make several HOI category **unseen** during training, e.g **<TV, sit>** only appears in testing data.

-We ensure the training data includes all the objects and interactions (e.g. **TV** and **sit**).

-Testing **systematicity** of systematic generalization.





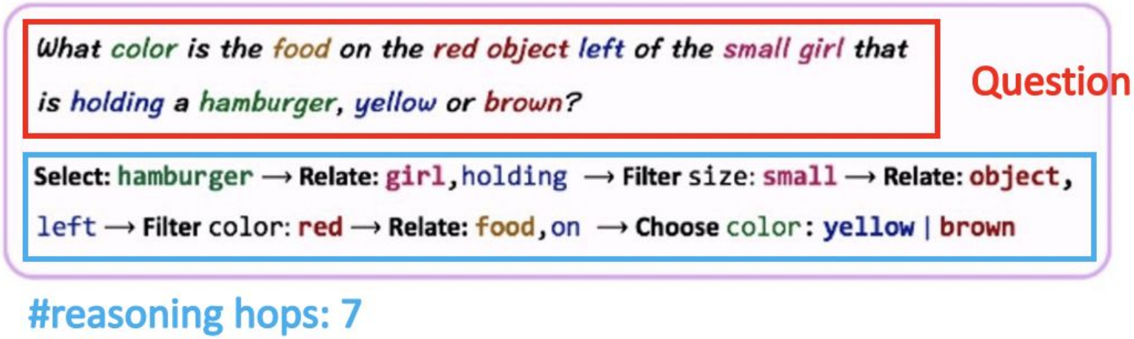
(a) Systematicity

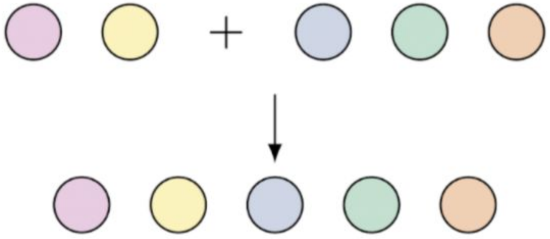
Experiments

Systematic generalization test for GQA:

-Each GQA question is also labelled with a reasoning program.

-We make training data only contain questions with shorter reasoning programs -- testing **productivity** of systematic generalization.





(b) Productivity

Experiments (HICO)

-We largely improve the current learners on both **standard test** and **systematic generalization test**, **without** any oracle object-centric representations (bboxes).

-Our model makes significant progress on **unseen categories**.

Method	Ext. superv.	Backbone	Orig.	Systematic-easy		Systematic-hard	
				Full cls.	Unseen cls.	Full cls.	Unseen cls.
Mallya & Lazebnik (2016)*		ResNet-101	33.8	-	-	-	-
Girdhar & Ramanan (2017)*	bbox	ResNet-101	34.6	-	-	-	-
Fang et al. (2018)*	pose	ResNet-101	39.9	-	-	-	-
Hou et al. (2020)†		ResNet-101	28.57	26.65	11.94	21.76	10.58
ViT-only		PVTv2-b2	35.48	31.06	11.14	19.03	18.85
EsViT (2021)		PVTv2-b2	38.23	35.15	11.53	22.55	21.84
RelViT (Ours)		PVTv2-b2	39.4	36.99	12.26	22.75	22.66
RelViT + EsViT (Ours)		PVTv2-b2	**40.12**	**37.21**	**12.51**	**23.06**	**22.89**

Experiments (GQA)

-We largely improve the current learners on both **standard test** and **systematic generalization test**, **without** any oracle object-centric representations (bboxes).

-This can be way impressive for VQA tasks -- object-detection play crucial role in almost all state-of-the-art VQA learners **but not with our method**.

Method	Bbox feat.*	Backbone	Orig.	Sys.
BottomUp (2018)	✓	ResNet-101	53.21	-
MAC (2018b)	✓	ResNet-101	54.06	-
MCAN-Small (2019)	✓	ResNet-101	58.35	36.21
MCAN-Small (2019)		ResNet-101	51.1	30.12
ViT-only		PVTv2-b2	56.62	31.39
EsViT (2021)		PVTv2-b2	56.95	31.76
RelViT (Ours)		PVTv2-b2	57.87	35.48

GQA overall accuracy	MCAN-Small (w/ bbox)	RelViT (PVTv2-b2)	RelViT (PVTv2-b3)	RelViT (Swin-base)
original	58.35	57.87	61.41	**65.54**
systematic	36.21	35.48	36.25	**37.51**

Takeaway

ViT is a promising architecture that offers **object-centric representations** and **relational inductive bias**.

Using **concept-guided contrastive learning** as an **auxiliary task** to further exploit the visual relational reasoning data could significantly boost the performance of ViTs on these tasks, especially on **systematic generalization**.

RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning