Abstract Visual Reasoning with Tangram Shapes


Anya Ji


Noriyuki Kojima


Noah Rush


Alane Suhr


Wai Keen Vong


Robert D. Hawkins


Yoav Artzi

lil.nlp.cornell.edu/kilogram

What does this look like?

cat

sleeping human

rabbit

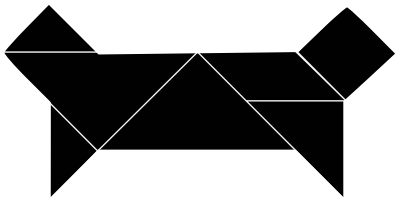
dog

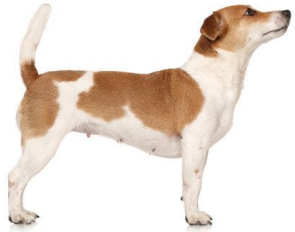
pig

This is a tangram puzzle

Tangrams are a Window into Abstraction

- **Tangrams** are abstract shapes built from 7 standard pieces

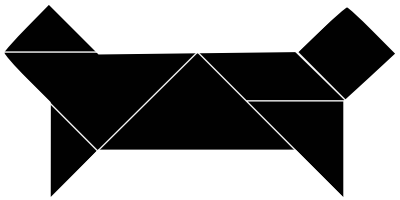


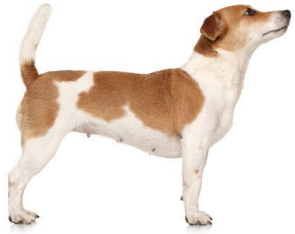


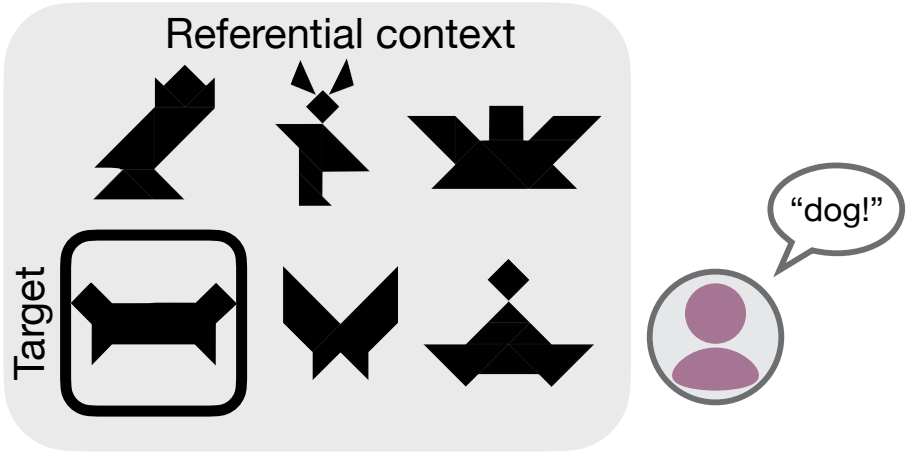
[Clark and Wilkes- Gibbs, 1986; Fox Tree, 1999; Hawkins et al., 2020]

Tangrams are a Window into Abstraction

- **Tangrams** are abstract shapes built from 7 standard pieces
- Often used in **reference games** to study abstraction and convention formation in humans
- Both important questions for NLP models and cognitive science
- But: research relies on a small set, limiting potential for generalization



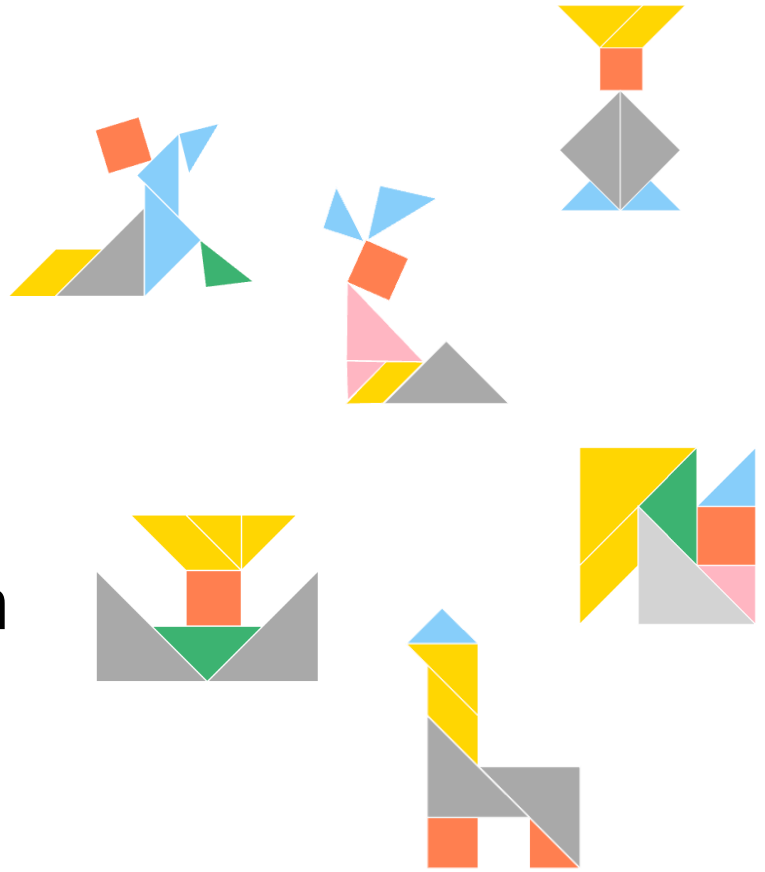




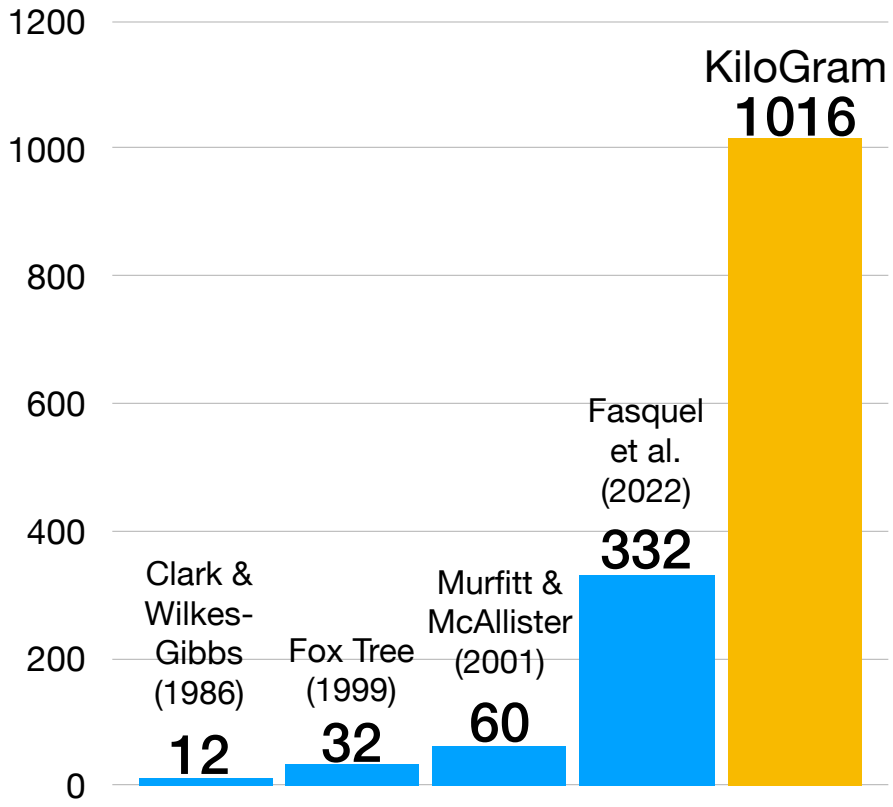
[Clark and Wilkes- Gibbs, 1986; Fox Tree, 1999; Hawkins et al., 2020]

Overview

- The KiloGram dataset
- Analyzing model generalization



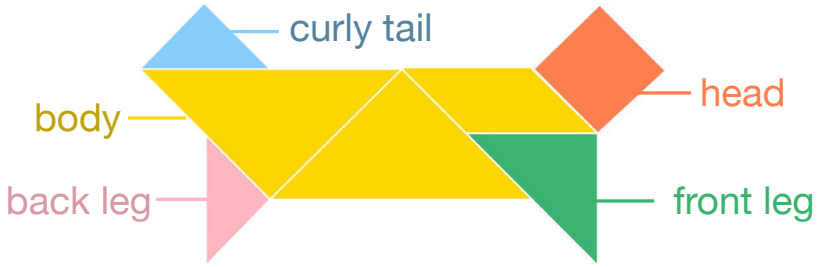
All Our Tangrams Belong to You



- Kilogram **significantly expands** the current resources
- **1016** tangrams
- **Vectorized representation** with standardized pieces

[Slocum. 2003. *The Tangram Book: The Story of the Chinese Puzzle with over 2000 Puzzles to Solve*.]

Language Annotations



dog

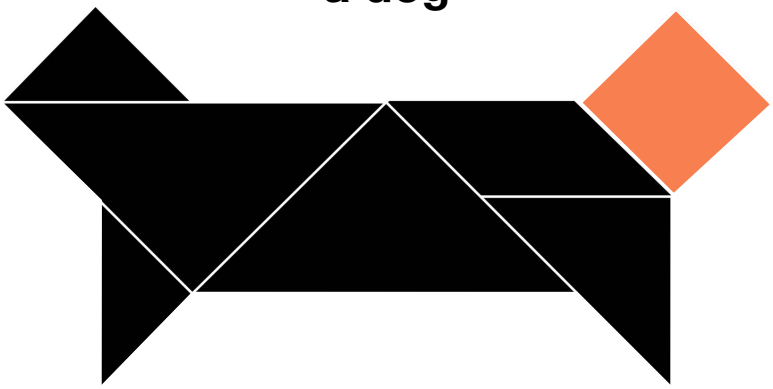
- Each tangram comes with language annotations
- Previous use of tangrams includes only **whole-shape descriptions**
- We also annotate:
  - **Part segmentation** along tangram pieces
  - Annotations for **part names**
- Allows us to explore the relationship between the whole shape and the parts in abstract reasoning

Annotation Task

This shape, as a whole, looks like **a dog** .

Annotation Task

a dog



This part looks like **head** .

Annotation Task

a dog

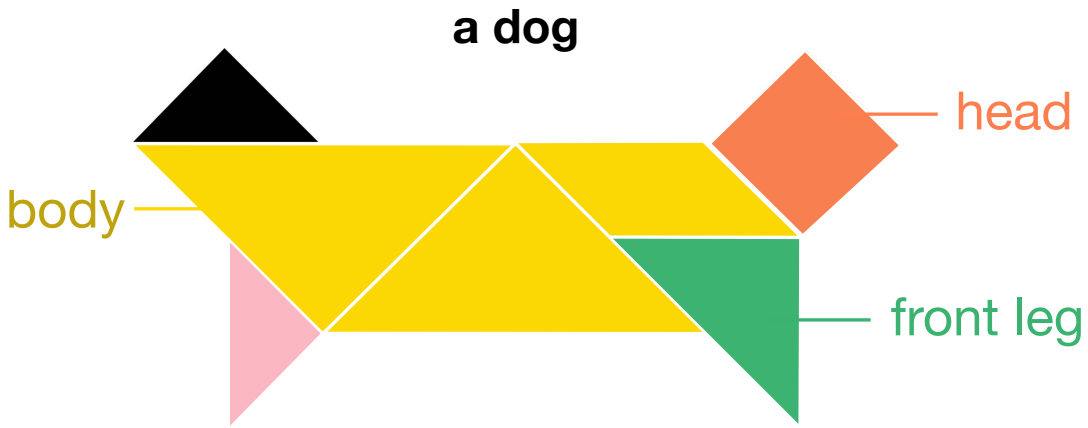
This part looks like **body** .

Annotation Task

a dog

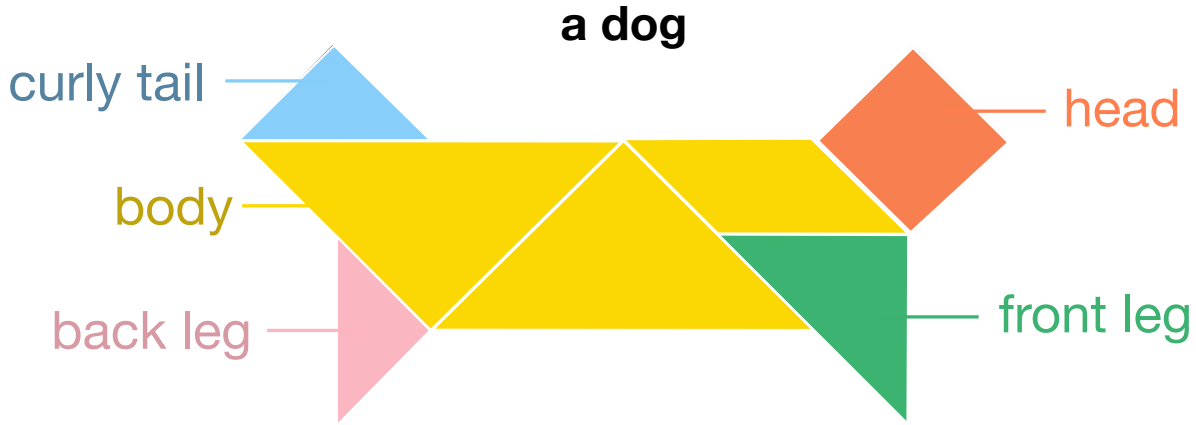
This part looks like **front leg** .

Annotation Task



This part looks like **back leg** .

Annotation Task



This part looks like **curly tail** .

Whole-shape name

a dog

Segmentation map

Submit

Part names

curly tail

body

back leg

head

front leg

What Did We Get?

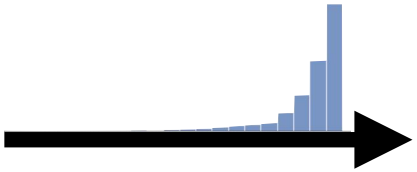
- 1,016 tangrams, 13,404 annotations
- 10 individual annotations for each tangram
- Densely annotated 74 tangrams: 50 annotations each
- Vocabulary size: 4,522

Diverse Data: Shape Naming

Low Diversity

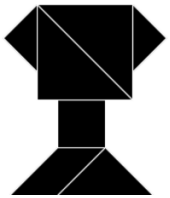
dog
dog
dog

fish
goldfish
a fish



Shape Naming Divergence

High Diversity

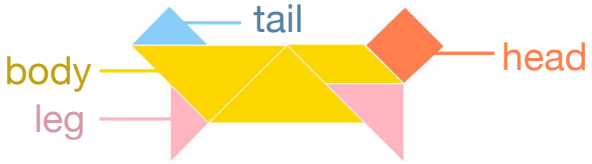


trophy
robot
princess leia from star wars

a rose
drill bit
street sign pole

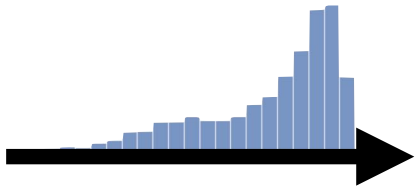
Diverse Data: Part Naming

Low Diversity



dog

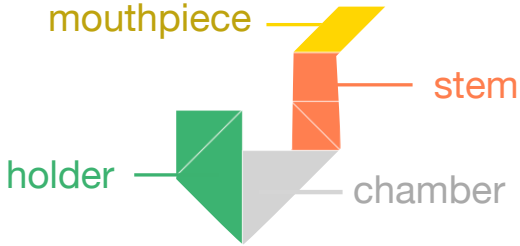
dog



Part Naming Divergence

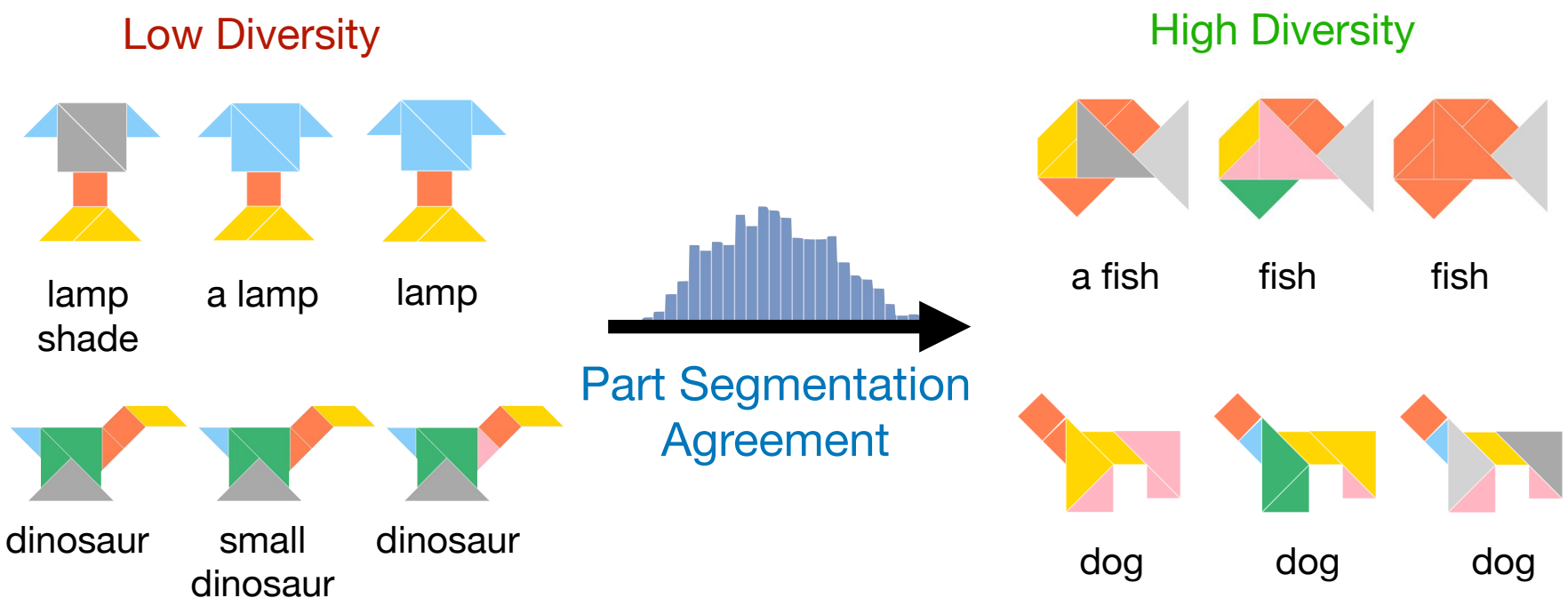
High Diversity

pipe



pipe

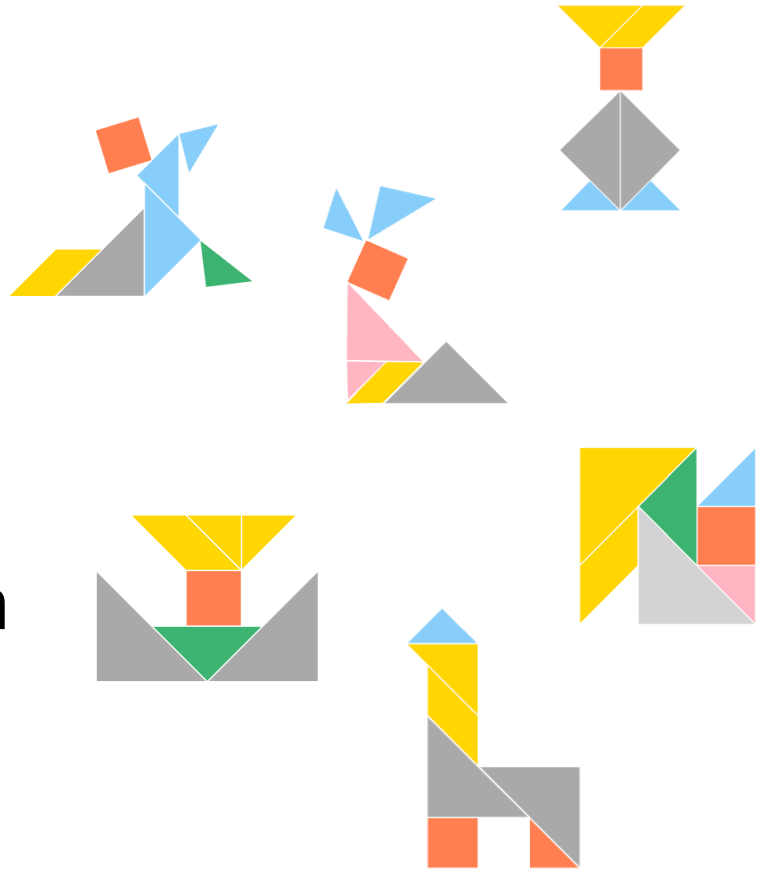
Diverse Data: Segmentation



Overview

- The KiloGram dataset

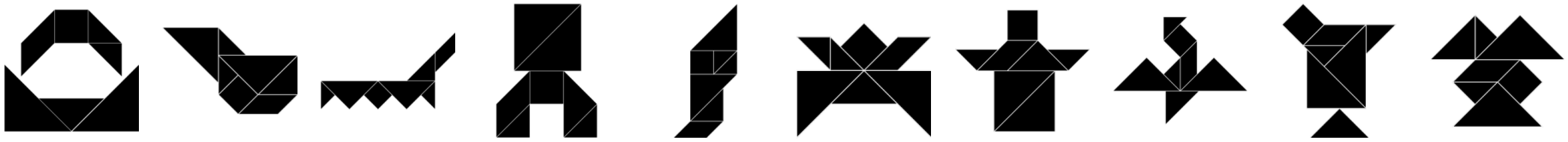
- Analyzing model generalization



Abstraction as Generalization

- Models that **generalize** should recognize known concepts in their **abstract** form
- **Reference games** with abstract stimuli test vision-language models for abstraction
- KiloGram allows doing this **at scale**

Reference Games



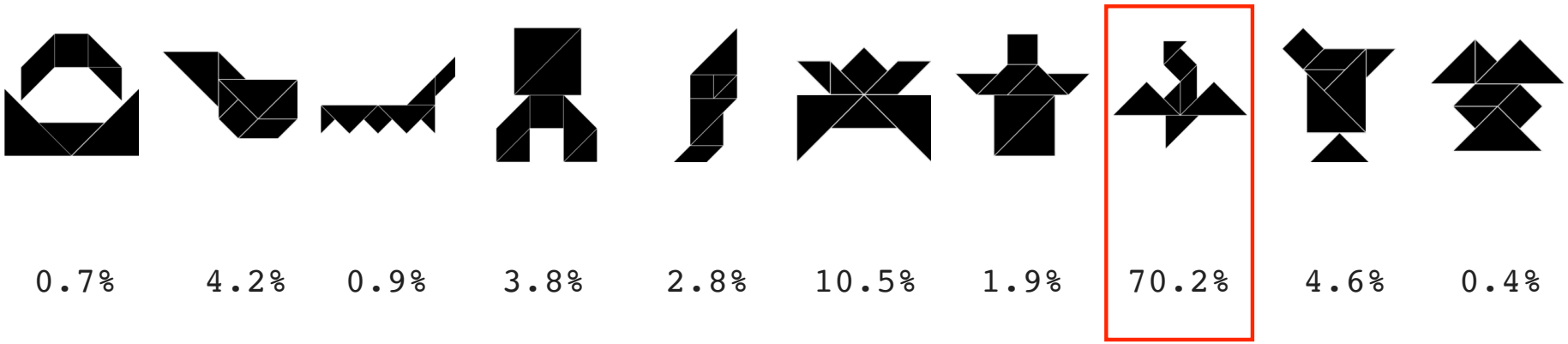
?

a flying goose

Experimental setups:

1 **WHOLE + BLACK**	2	3	4

Reference Games



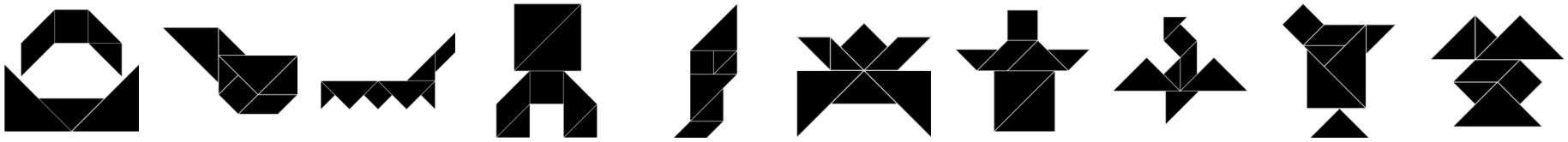
0.7% 4.2% 0.9% 3.8% 2.8% 10.5% 1.9% 70.2% 4.6% 0.4%

a flying goose

Experimental setups:

1	2	3	4
WHOLE + BLACK			

Reference Games



?

a flying goose with a head, wings, a neck, and a body

Experimental setups:

1 WHOLE + BLACK	**2 PARTS + BLACK**	3	4

Reference Games

?

a flying goose

Experimental setups:

1 WHOLE + BLACK	2 PARTS + BLACK	**3 WHOLE + COLOR**	4

Reference Games

?

a flying goose with a head, wings, a neck, and a body

Experimental setups:

1 WHOLE + BLACK	2 PARTS + BLACK	3 WHOLE + COLOR	**4 PARTS + COLOR**

Reference Games

?

a flying goose with a head, wings, a neck, and a body

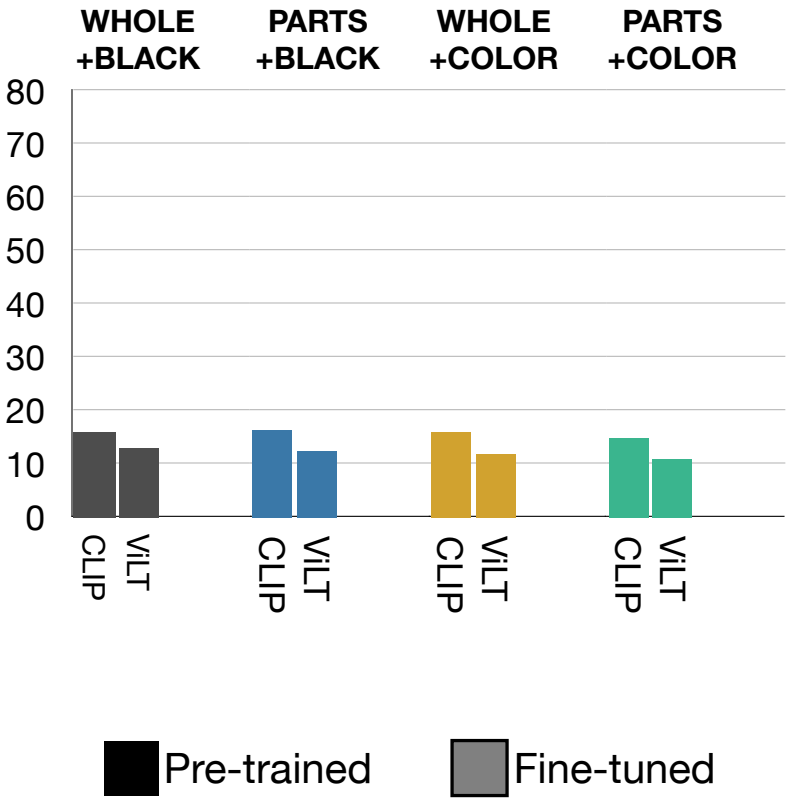
Experimental setups:

1 WHOLE + BLACK	2 PARTS + BLACK	3 WHOLE + COLOR	**4 PARTS + COLOR**

Evaluating Vision-Language Models

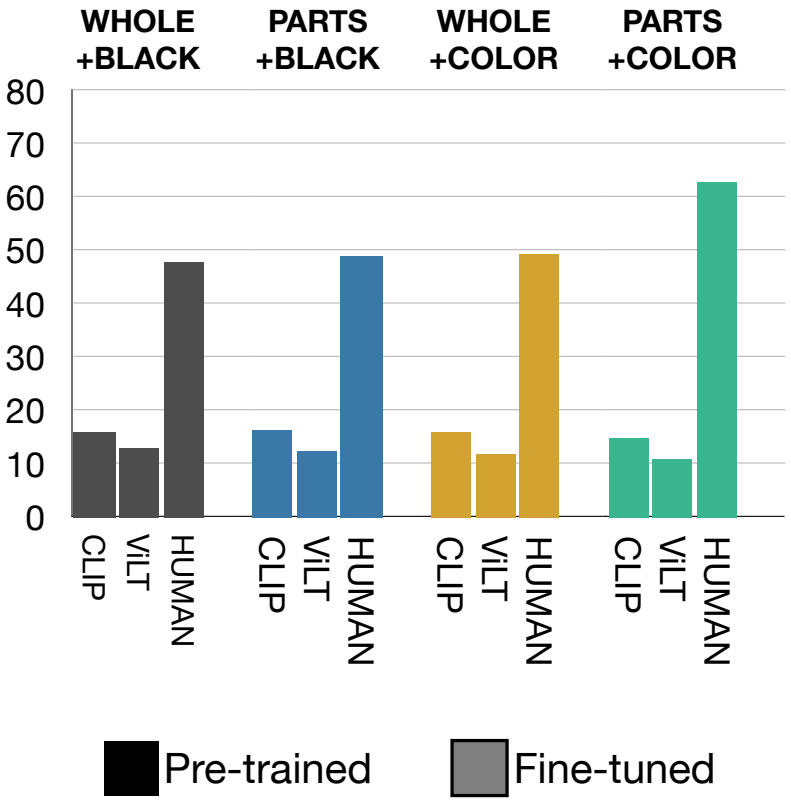
- CLIP (Radford et al., 2021): **separate** encoding of image and text
- ViLT (Kim et al., 2021): **joint** encoding of image and text
- Zero-shot and fine-tuned evaluation using reference games
- 10 tangrams per game

Results



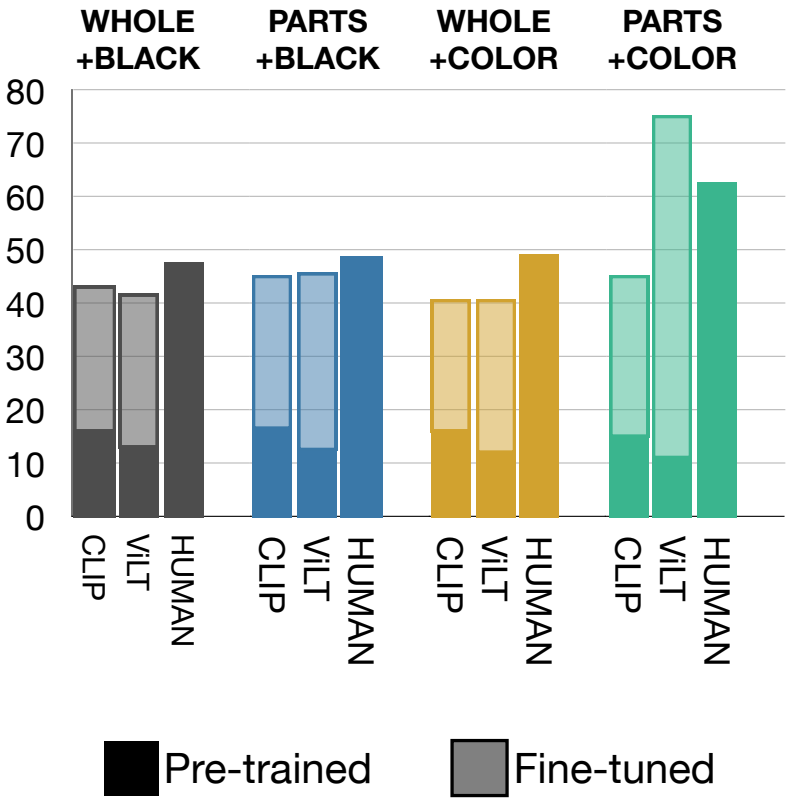
- Pre-trained models show **poor generalization**
- They also show **no use** of part information in every condition

Results



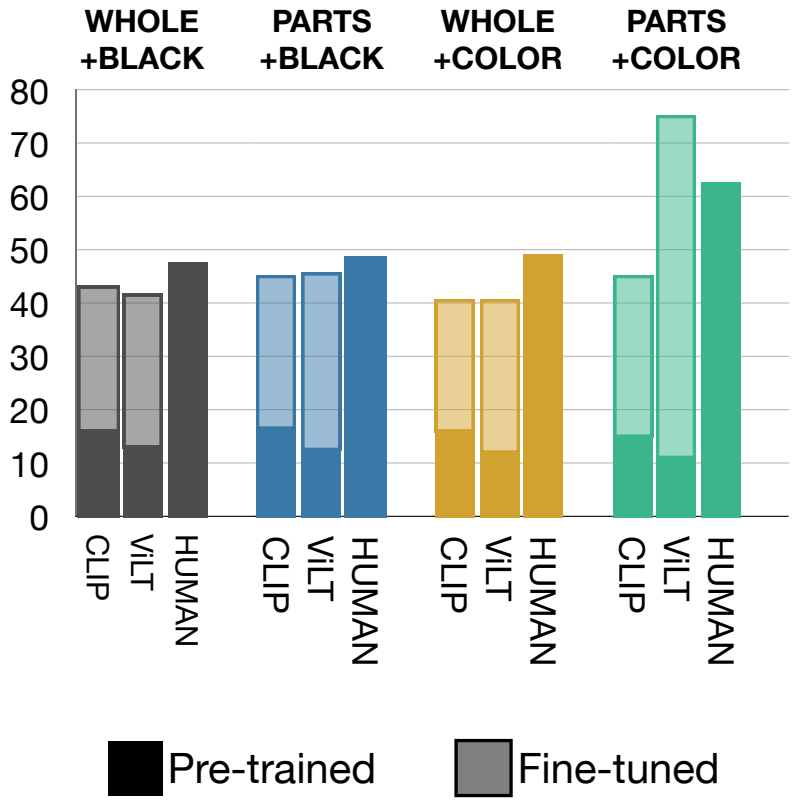
- Pre-trained models show **poor generalization**
- They also show **no use** of part information in every condition

Results



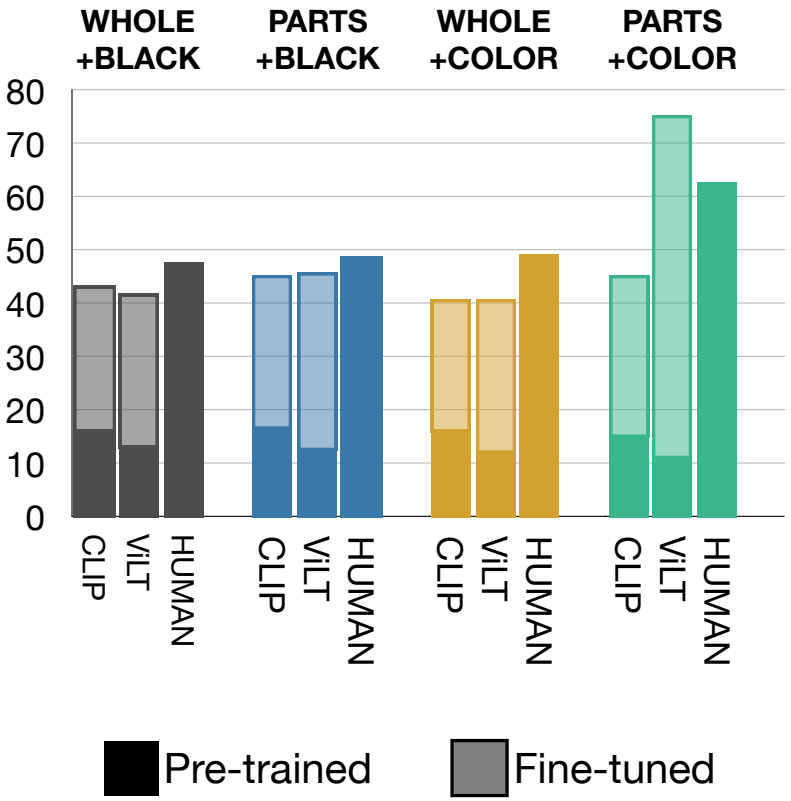
- Pre-trained models show **poor generalization**
- They also show **no use** of part information in every condition
- **Fine-tuning** dramatically increases model performance
- Fine-tuned models **benefit** from **part information**, especially ViLT

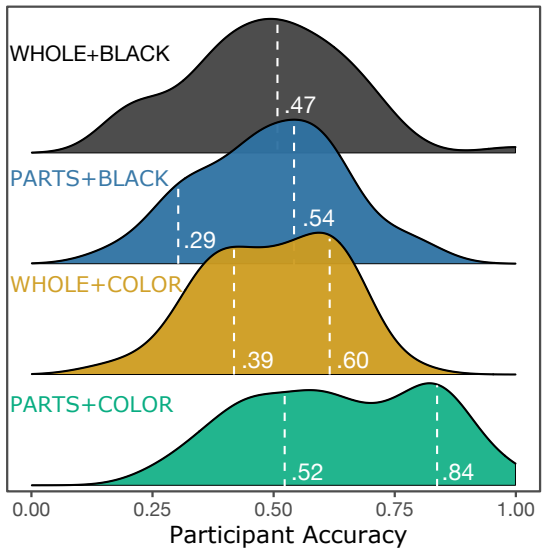
Results



- Pre-trained models show **poor generalization**
- They also show **no use** of part information in every condition
- **Fine-tuning** dramatically increases model performance
- Fine-tuned models **benefit** from **part information**, especially ViLT
- Fine-tuned models approach or even **surpass** human performance

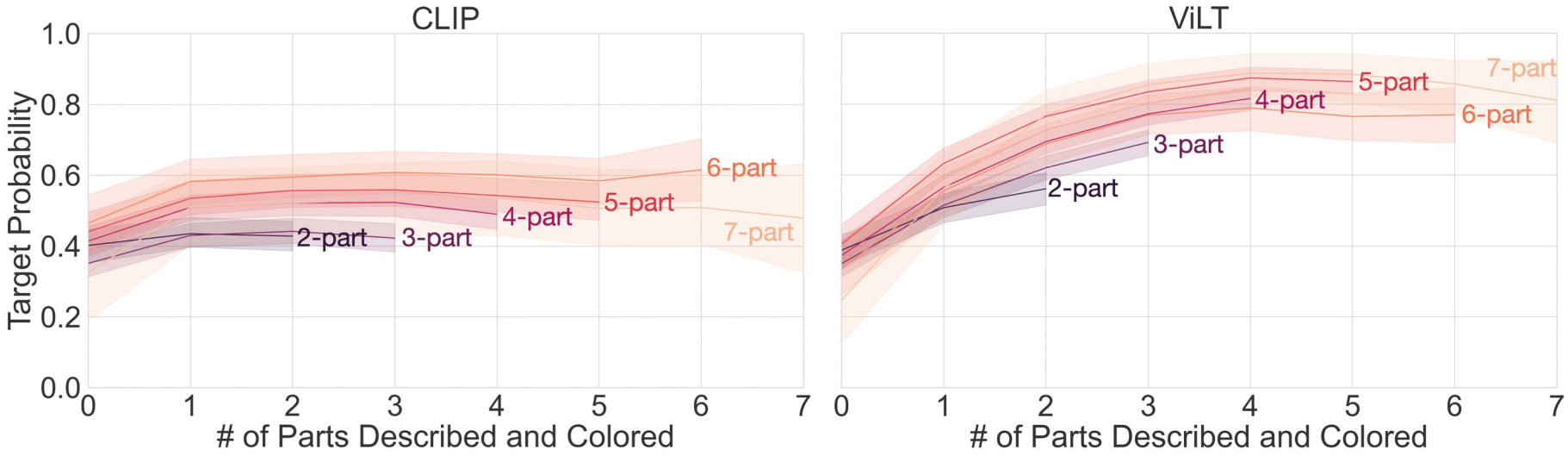
Human Performance





- **High-performing sub-population** still **outperforms** ViLT in every condition
- Low-performing sub-population may have not made full use of part correspondence information

How does adding part information help?



Part information is beneficial, but with a **diminishing return** as more part information is added

Thank you!

- KiloGram: a larger and richer tangrams resource
- Pre-trained models fail to generalize via abstraction
- Reasoning about parts improves both human and model performance

lil.nlp.cornell.edu/kilogram