Bongard-Tool:
Tool Concept Induction from Few-Shot Visual Exemplars

Team UnCoReable
Guangyuan Jiang* Chuyue Tang* Yuyang Li† Yu Liu†

* Yuanpei Collage, **Peking University**
† Deptartment of Automation, **Tsinghua University**

The water that bears the boat
is the same that swallows it up.

Motivation

What makes a physical object *tool*?

What makes a physical object *tool*?

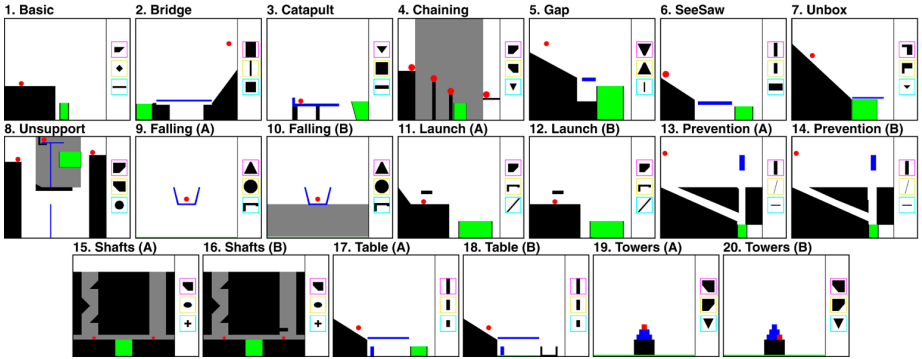
- **Multiple objects** can support the **same** tool use.
- **Compositional** nature of tools makes it hard to **understand** an object's tool-like functionalities **without context**.
- What can a book be?
  - A hammer for smashing
  - A lid covering instant noodles
  - Fuel when making fire
  - A pad for balancing an uneven table.

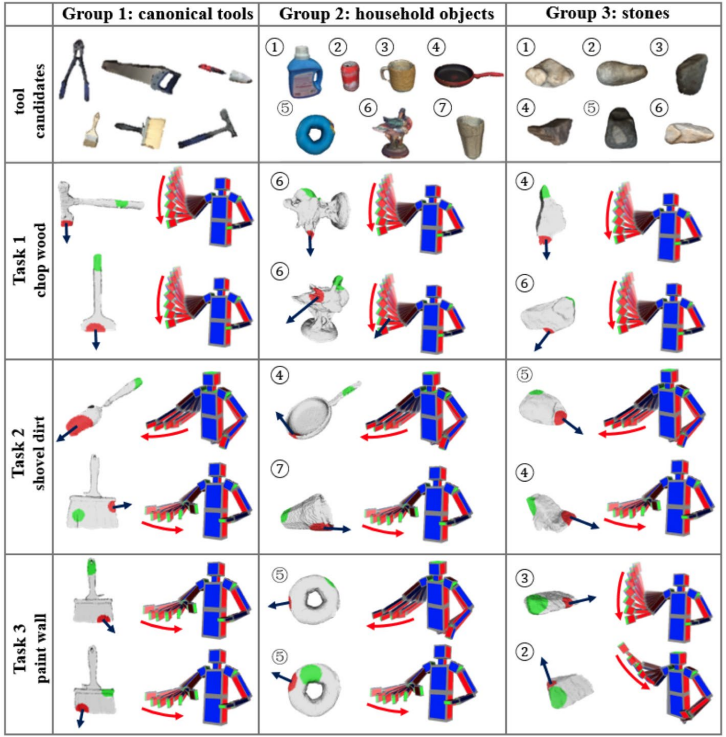
Context of Tools Matter

- A tool is a **composition** of **multiple functions**, and the **context determines** which one comes into action.
- To address this unique property, we introduce the **Bongard-Tool** benchmark for understanding the compositional concepts of tool use.

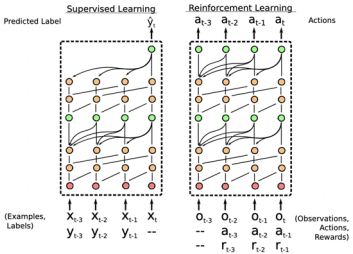
Related Work

Tool Understanding

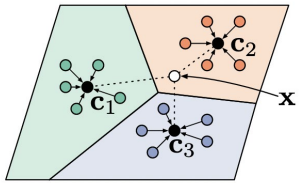




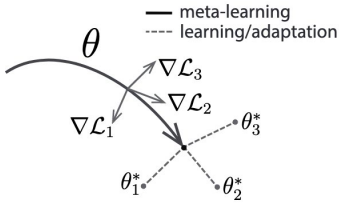
Few-Shot Learning Methods



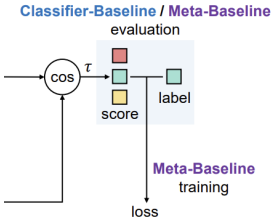
SNAIL (Mishra, 2017)



ProtoNet (Snell, 2017)

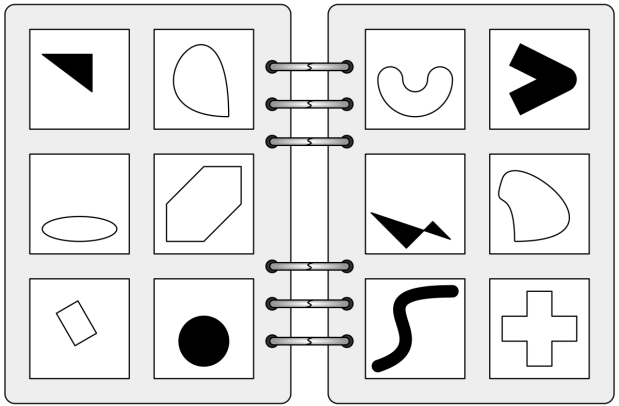


MAML (Finn, 2017)



Meta-Baseline (Chen, 2020)

The Bongard Problem (Bongard, 1970)

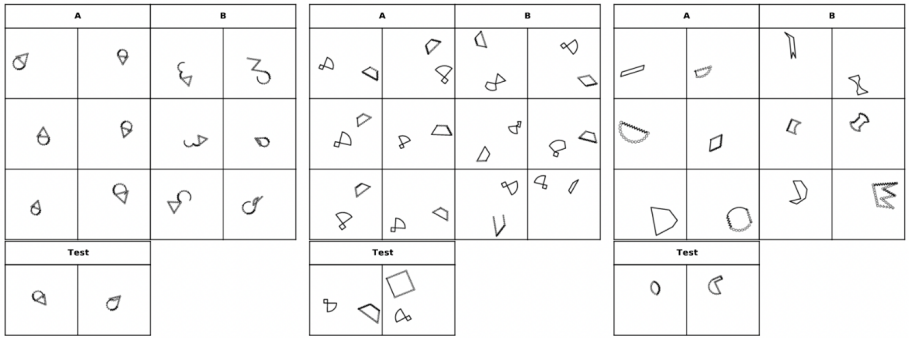


Share some *property*
E.g. shape, color, algorithmic rules, ...

No such *property*

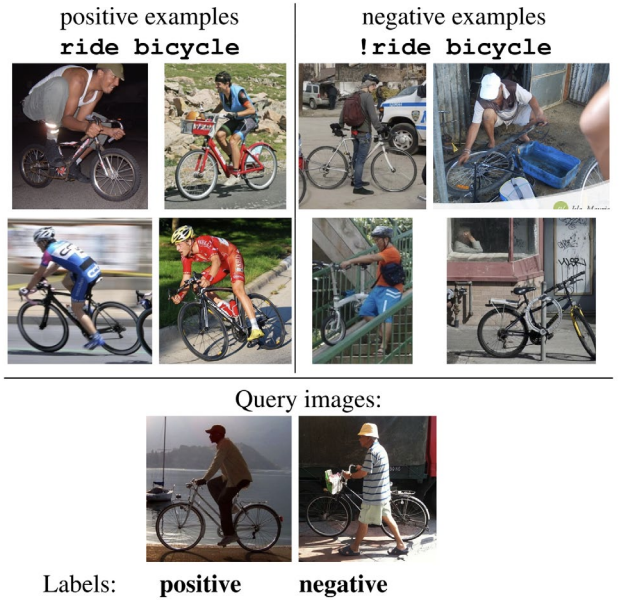
Having this *property* ?

Bongard-LOGO (Nie *et al.*, 2020)



(a) free-form shape problem (b) basic shape problem (c) abstract shape problem

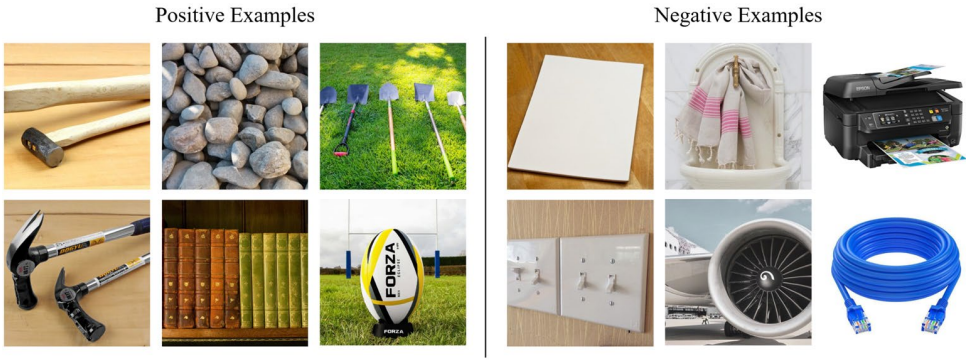
Bongard-HOI (Jiang *et al.*, 2022)



The Bongard-Tool Benchmark

Benchmark Formulation

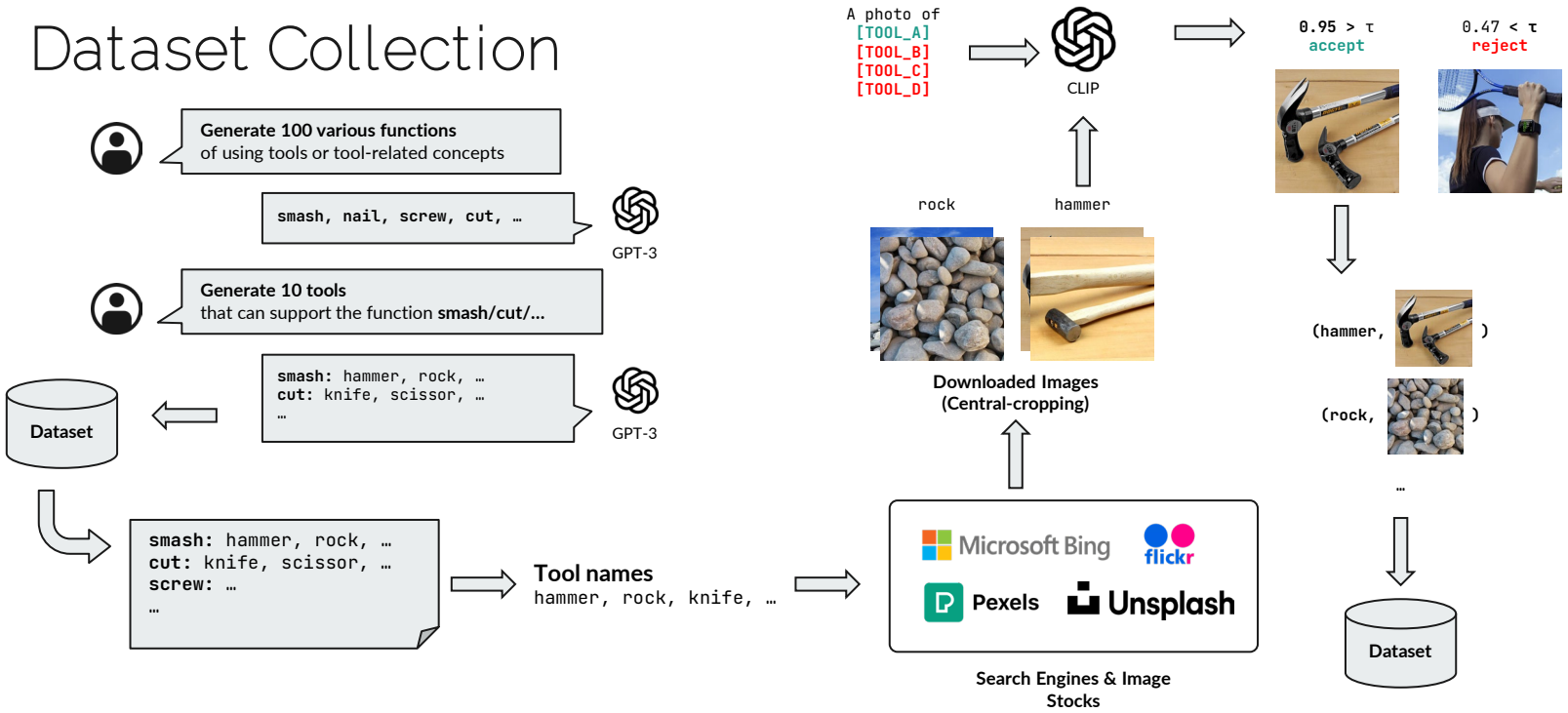
M=6 Tools
that **can** support the same function



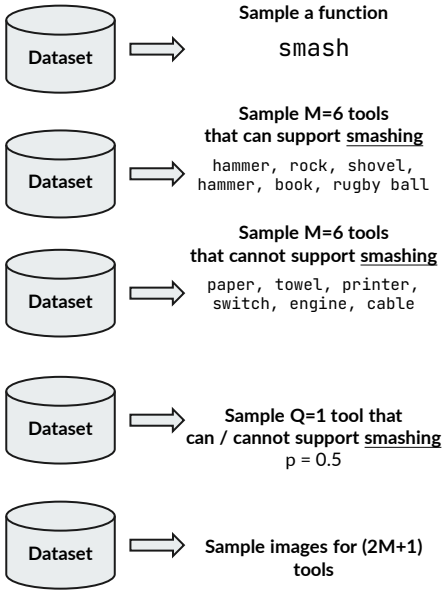
M=6 Tools
that **cannot** support the function

1 Query Image
to be predicted

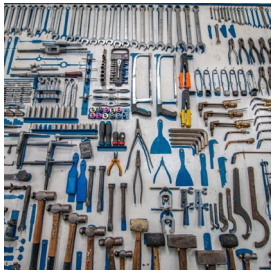
Dataset Collection



Bongard Instance Generation



Text-Image Mismatch in Dataset Construction



Tool Cluster (Many-to-One)

Hammer? Piler? Saw? ...

Text Polysemy (One-to-Many)

Rock? Johnson the Rock? Rock'n'Roll? ...

Grounding Disagreement

Camera? Camera App? Camera Logo? ...

Rare / Abstract Tools

Brick Hammer? PDF File? ...

?

Baseline Experiments
with Existing Methods

Normal Split v.s. Conceptual Generalization Split

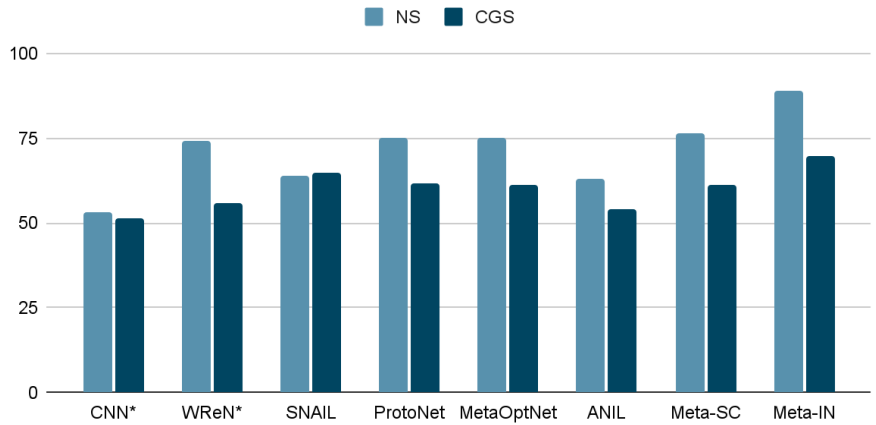
Normal Split:

- sample Bongard-Tool instances from the whole dataset
- randomly select: 80% for training and 20% for testing
- **models have seen all concepts (functions) in training.**

Conceptual Generalization Split:

- randomly select: 80% of the functions for training 20% for testing
- sample Bongard-Tool instances from training concepts (functions) and testing concepts respectively.
- **models have never seen the testing concepts**

Test Accuracy / %



* Non-Episodic Methods

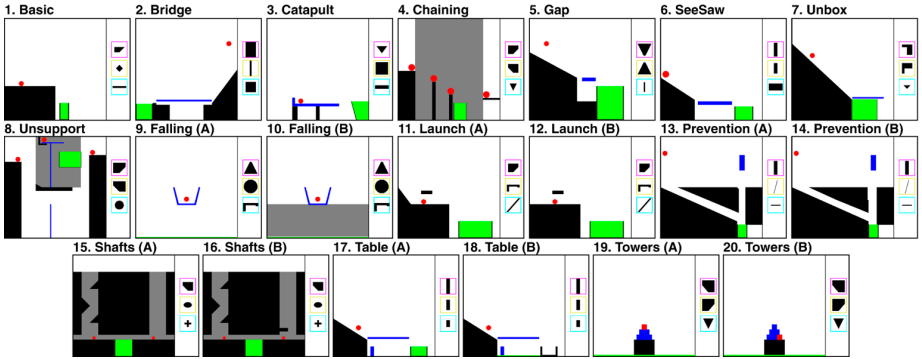
Split	Non-Episodic		Meta-Learning Methods					
	CNN	WReN	SNAIL	ProtoNet	MetaOptNet	ANIL	Meta-SC	Meta-IN
NS	52.97	74.10	63.91	**75.18**	75.13	63.03	76.54	**88.82**
CGS	51.35	55.90	**64.42**	61.83	61.02	54.29	61.37	**69.78**

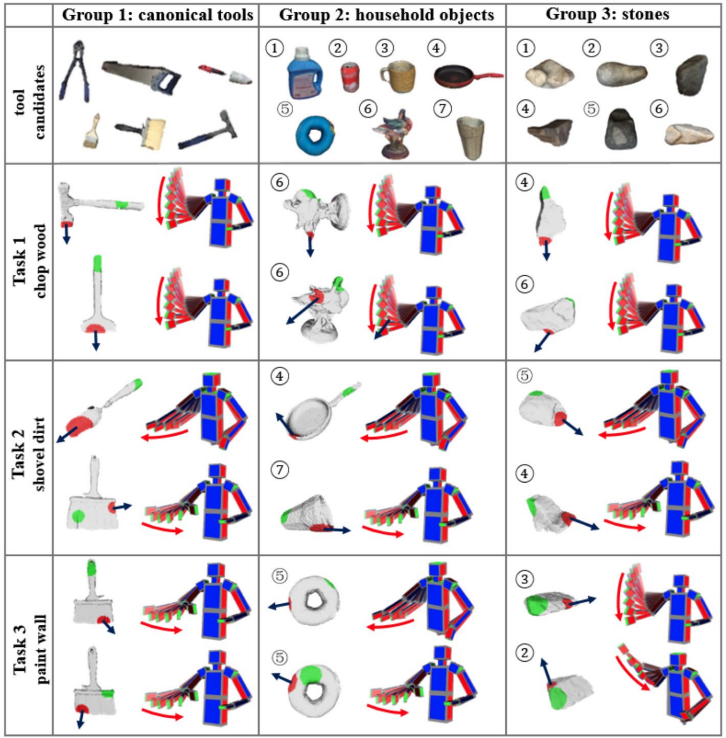
Table 1: **Test accuracy on baseline methods.** We report the test accuracy of baseline methods on our benchmark. Bold fonts indicate the best result in the group.

- Meta-learning methods generally outperform non-episodic methods
- These models may find some shortcuts to our Bongard-Tool tasks.
- Visual representations are not enough.

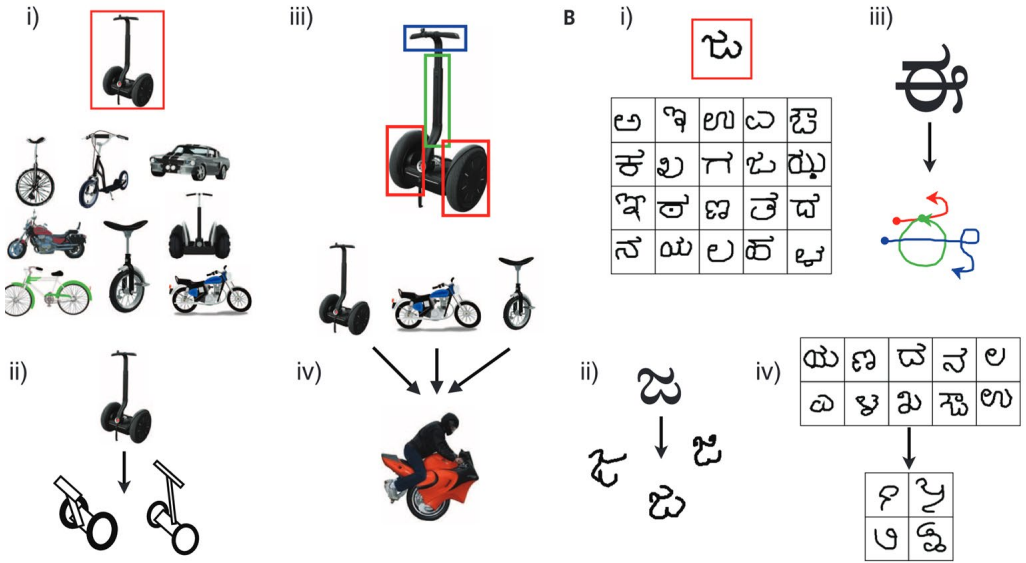
Discussion

Tool Understanding





Concept Induction



Few-Shot Visual Classification v.s. Few-Shot Reasoning

- Requires almost no high-level reasoning ability
- Object-level knowledge
- A test-bed for visual perception ability

- Requires high-level **reasoning** ability
- **Object & function**-level knowledge
- A test-bed for visual reasoning ability

Conclusion

Problem Formulation: What is Bongard-Tool?

- We take the **form** of **Bongard**, but base on the **affordance** and **functionalities** of **tools**.
- Our benchmark requires more knowledge than the information provided by the images.
- For explainable human-level **induction**, many aspects should be taken into consideration, including **priors** of **physical properties of tools**, **experiences** of using tools, etc.
- Existing methods show **acceptable performance**, but it is more likely that they find **shortcuts** between tool functionalities and their image patterns.

Contributions

- We formulate the context-dependent tool understanding as a few-shot concept induction problem, which can represent a broad range of compositional functionalities of tools.
- We propose the Bongard-Tool benchmark for addressing the context-dependent tool understanding in visual reasoning. Extensive experiments on recent few-shot and meta-learning methods show the hardship of understanding compositional tool concepts from pure visual perception.
- We demonstrate the effectiveness of fast constructing large-scale datasets by utilizing large language models for knowledge building, web crawling, and vision-language models for content retrieval and filtering.

Bongard-Tool:
Tool Concept Induction from Few-Shot Visual Exemplars

Team UnCoReable
Guangyuan Jiang* Chuyue Tang* Yuyang Li† Yu Liu†

* Yuanpei Collage, **Peking University**
† Deptartment of Automation, **Tsinghua University**