



Bongard-OpenWorld:

Few-Shot Reasoning for Free-form Visual Concepts in the Real World

Rujie Wu ^{*1}, Xiaojian Ma ^{*2}, Zhenliang Zhang ², Wei Wang ², Qing Li ², Song-Chun Zhu ^{2,3,4}, Yizhou Wang ^{1,4}

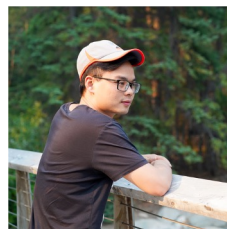
¹ School of Computer Science, Peking University

² National Key Laboratory of General Artificial Intelligence, BIGAI

³ School of Intelligence Science and Technology, Peking University

⁴ Institute for Artificial Intelligence, Peking University

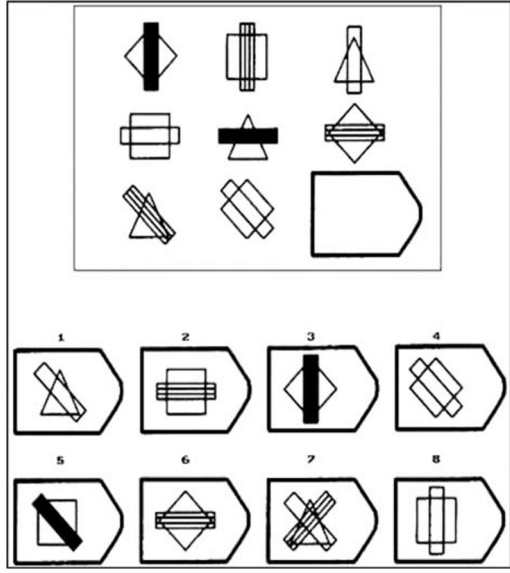
* Equal contribution



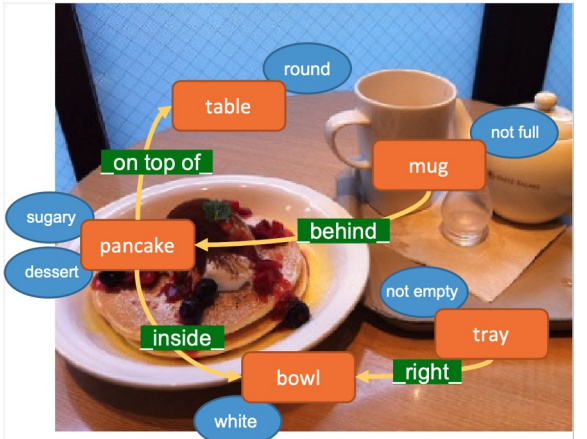
Few-shot learning and visual reasoning: niche of visual intelligence



Few-shot learning in visual recognition



Visual IQ test or abstract visual reasoning



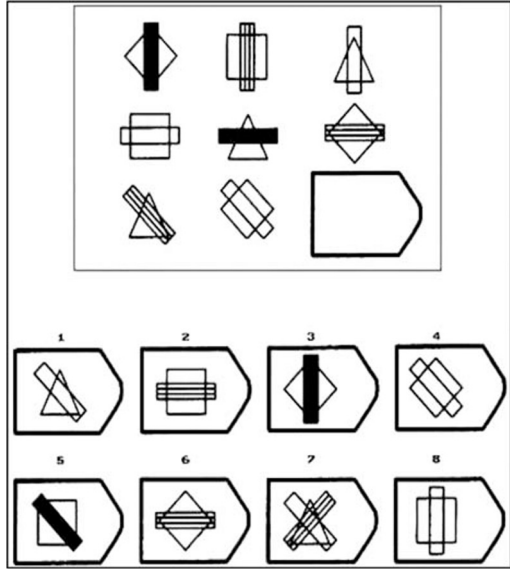
Q: What is inside the thing that to the right of the tray?

Relational reasoning, ex. VQA

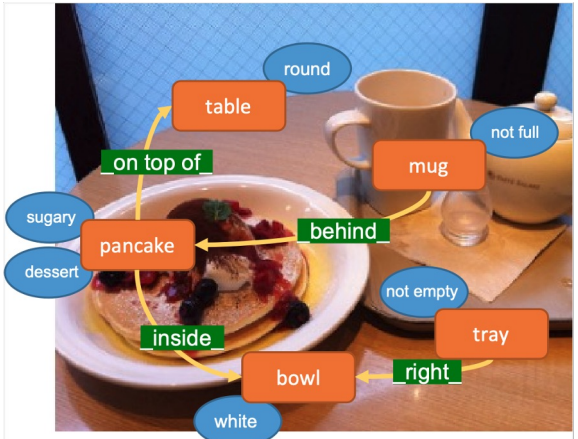
Few-shot learning and visual reasoning: niche of visual intelligence



few-shot induction identifying the *visual concept* (category) with a handful of examples;
meta learning generalizing this to novel concepts.



few-shot induction inferring the hidden *patterns* from very few demonstrations;
relational learning the *patterns* sometime control multiple images rather than just one.



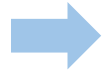
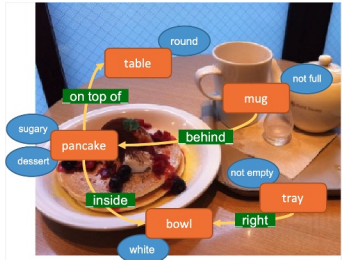
Q: What is inside the thing that to the right of the tray?

relational learning need to answer questions about the relationships between entities (objects);
meta learning generalization to novel relationships.

Reconciling few-shot learning and visual reasoning



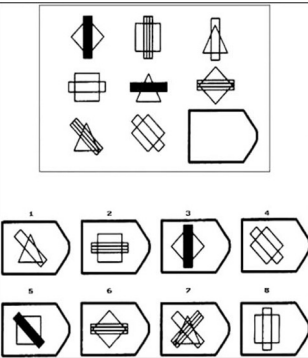
Σ



Few-shot induction
Inducing arbitrary concepts, just from very few examples.

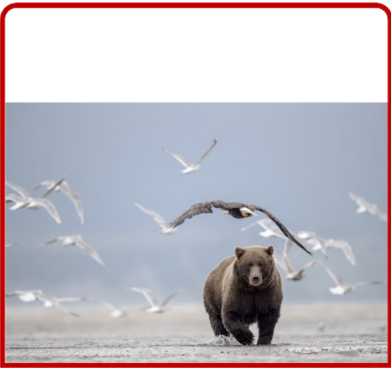
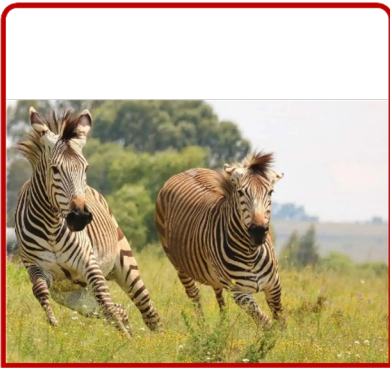
Relational learning
Learning to comprehend and extend relationships among entities.

Meta learning
Generalizing the learned “learning algorithm” to novel scenarios.



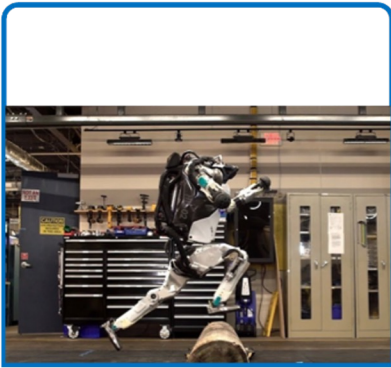
Introducing Bongard-OpenWorld

positive set \mathcal{P}



...

negative set \mathcal{N}



...

query image I_q



- 1 I_q belongs to \mathcal{P} or \mathcal{N} ?
- (optional) What is the concept exclusively depicted by \mathcal{P} ?

Introducing Bongard-OpenWorld

positive set \mathcal{P}

Two zebras are running in the grass.



A bear running with many birds flying around it.




A dog is running on the beach.



negative set \mathcal{N}

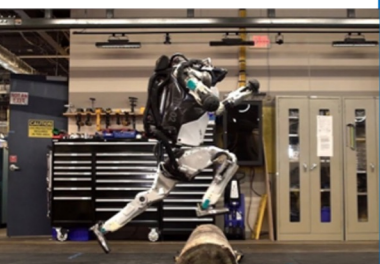
A herd of zebras are drinking by the river.



A bear is catching fish in the shallow river.



A humanoid robot is running.



query image I_q

A herd of wildebeest are running in the grassland.



- 1 I_q belongs to \mathcal{P} or \mathcal{N} ?
- (optional) What is the concept exclusively depicted by \mathcal{P} ?

Ground truth concept c (as sentence):
animals are running

Reconciling few-shot learning and visual reasoning

Few-shot induction

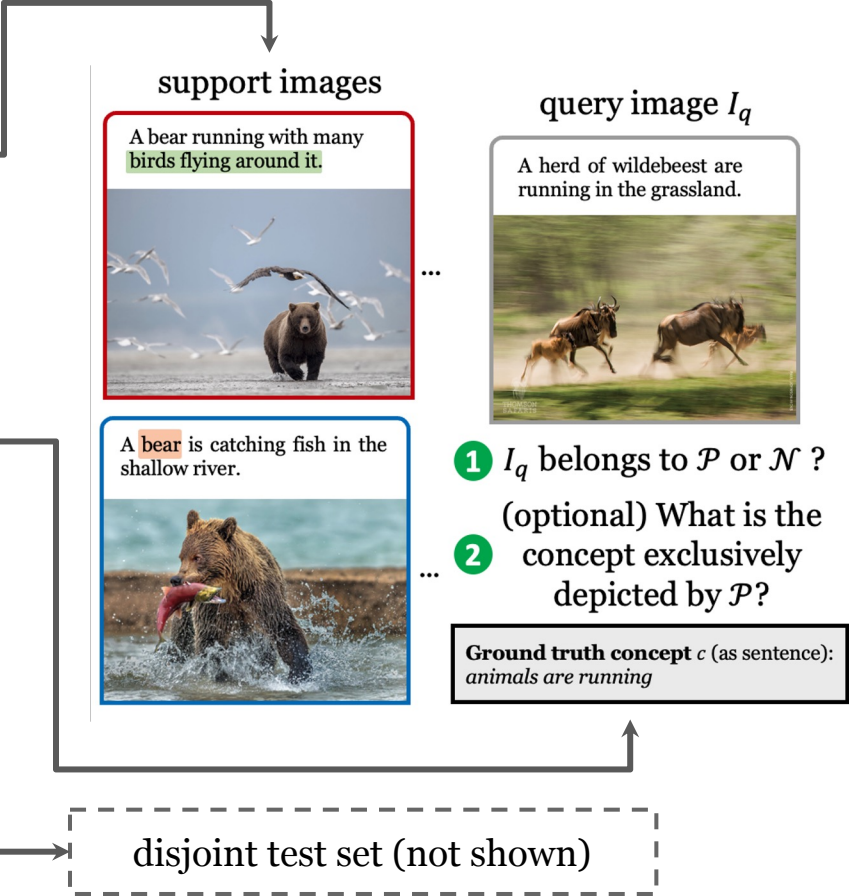
Inducing arbitrary concepts, but just from very few examples.

Relational learning

Learning to comprehend and extend relationships among entities.

Meta learning

Generalizing the learned “learning algorithm” to novel scenarios.



Constructing Bongard-OpenWorld



the stars in the night sky
a worker helps to clear the debris
food market showing the traditional food
pop artist performs at the festival in a city



stars night sky
worker help clear debris
food market traditional food
pop artist performs festival city

streamlining text &
grid-sampling + sliding window
(2 / 3 / 4 / 5)

Concept Category	ID	Example
Anything else (non-CS*)	0	Animals are running.
HOI	1	A person playing the guitar.
Taste / Nutrition	2	A plate of high-calorie food.
Color / Material / Shape	3	A wooden floor in the living room.
Functionality / Status / Affordance	4	An animal capable of flying in the tree.

Concept Category	ID	Example
And / Or / Not	5	A man without beard.
Factual Knowledge	6	A building in US capital.
Meta Class	7	Felidae animals.
Relationship	8	A bench near trees.
Unusual observations	9	Refraction of light on a glass cup.

image-text pairs database:
LAION-5B, CC3M...

1

2

3

concept augmentation

Constructing Bongard-OpenWorld

child playing ball room

concepts (as n-tuples)



ChatGPT



User 1) Expand a word tuple into positive sentences by inserting distracting objects, attributes, etc.
User 2) Reduce a word tuple into negative sentences by partially removing a word from it and optionally adding distracting words.

Assistant Here you go!

positives

child playing ball with friends
child playing ball on the beach
child playing ball in the pool
child play ball with a coach
(distractors)

negatives

child playing dolls in the room
child playing video games
child playing with dogs
a dog *playing with ball*
(*partially overlapping*)

Constructing Bongard-OpenWorld

positives

child playing ball with friends
child playing ball on the beach
child playing ball in the pool
child play ball with a coach
(distractors)

negatives

child playing dolls in the room
child playing video games
child playing with dogs
adults playing with ball
a dog playing with ball
(*partially overlapping*)



support images

A bear running with many birds flying around it.

...

query image I_q

A herd of wildebeest are running in the grassland.

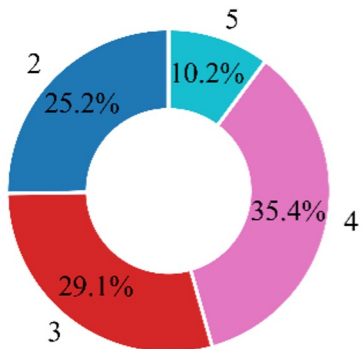
1 I_q belongs to \mathcal{P} or \mathcal{N} ?
(optional) What is the concept exclusively depicted by \mathcal{P} ?

2

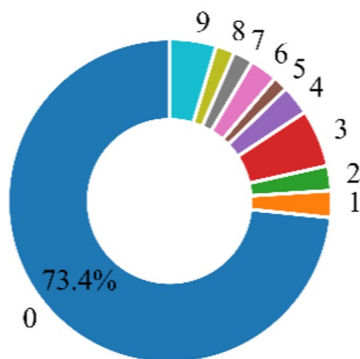
Ground truth concept c (as sentence):
animals are running

Constructing Bongard-OpenWorld

a) Concept Length Distribution



b) Concept Category Distribution



c) Word Distribution



Concept Category	ID	Example	Concept Category	ID	Example
Anything else (non-CS*)	0	Animals are running.	And / Or / Not	5	A man without beard.
HOI	1	A person playing the guitar.	Factual Knowledge	6	A building in US capital.
Taste / Nutrition	2	A plate of high-calorie food.	Meta Class	7	Felidae animals.
Color / Material / Shape	3	A wooden floor in the living room.	Relationship	8	A bench near trees.
Functionality / Status / Affordance	4	An animal capable of flying in the tree.	Unusual observations	9	Refraction of light on a glass cup.

The Bongard Trilogy

A		B	
Test			

positive examples
ride bicycle

negative examples
!ride bicycle

Query images:

Labels: **positive** **negative**

support images

A bear running with many birds flying around it.

query image I_q

A herd of wildebeest are running in the grassland.

- 1 I_q belongs to \mathcal{P} or \mathcal{N} ?
- (optional) What is the concept exclusively depicted by \mathcal{P} ?

Ground truth concept c (as sentence):
animals are running

Bongard-LOGO (2020)
(very close to the original Bongard problem)

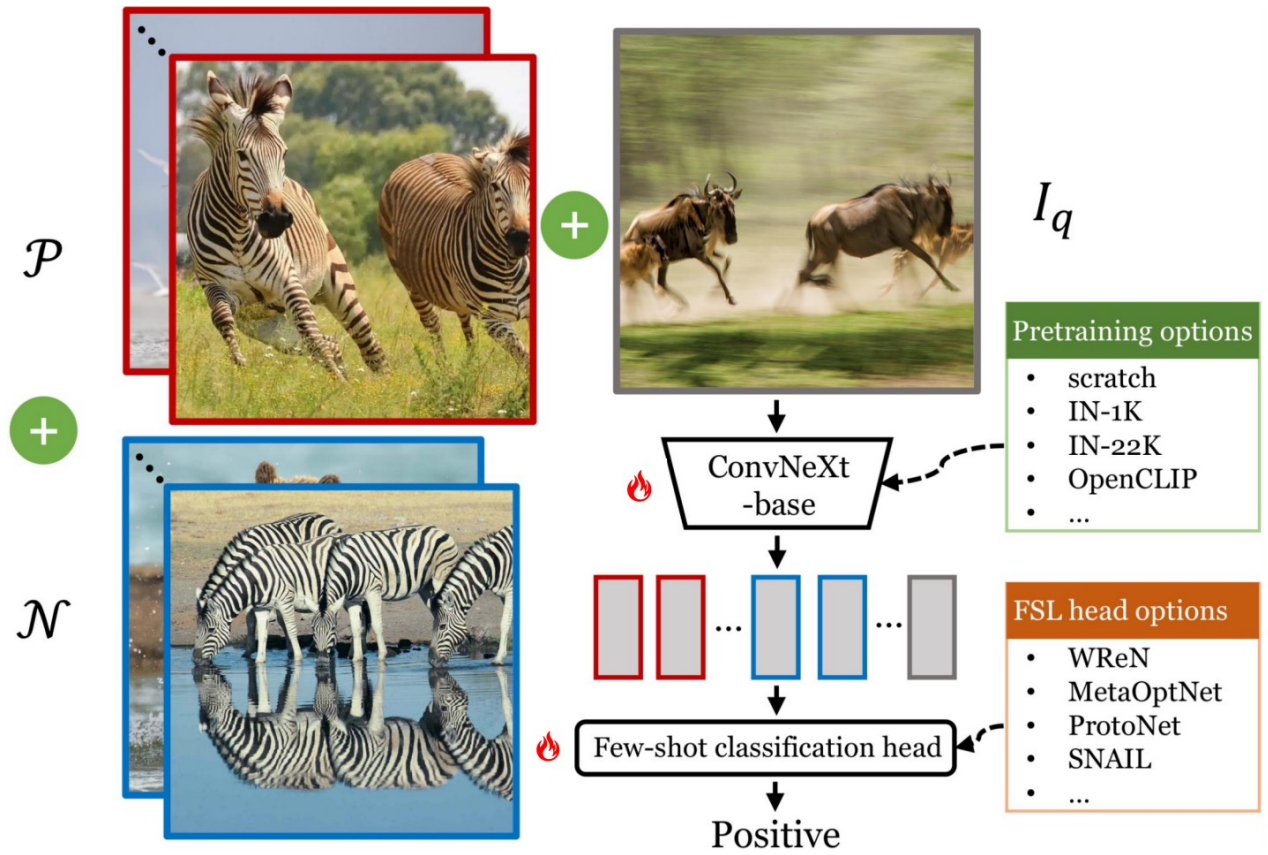


Bongard-HOI (2021)
+ real-world images
+ hard negatives
+ generalization tests



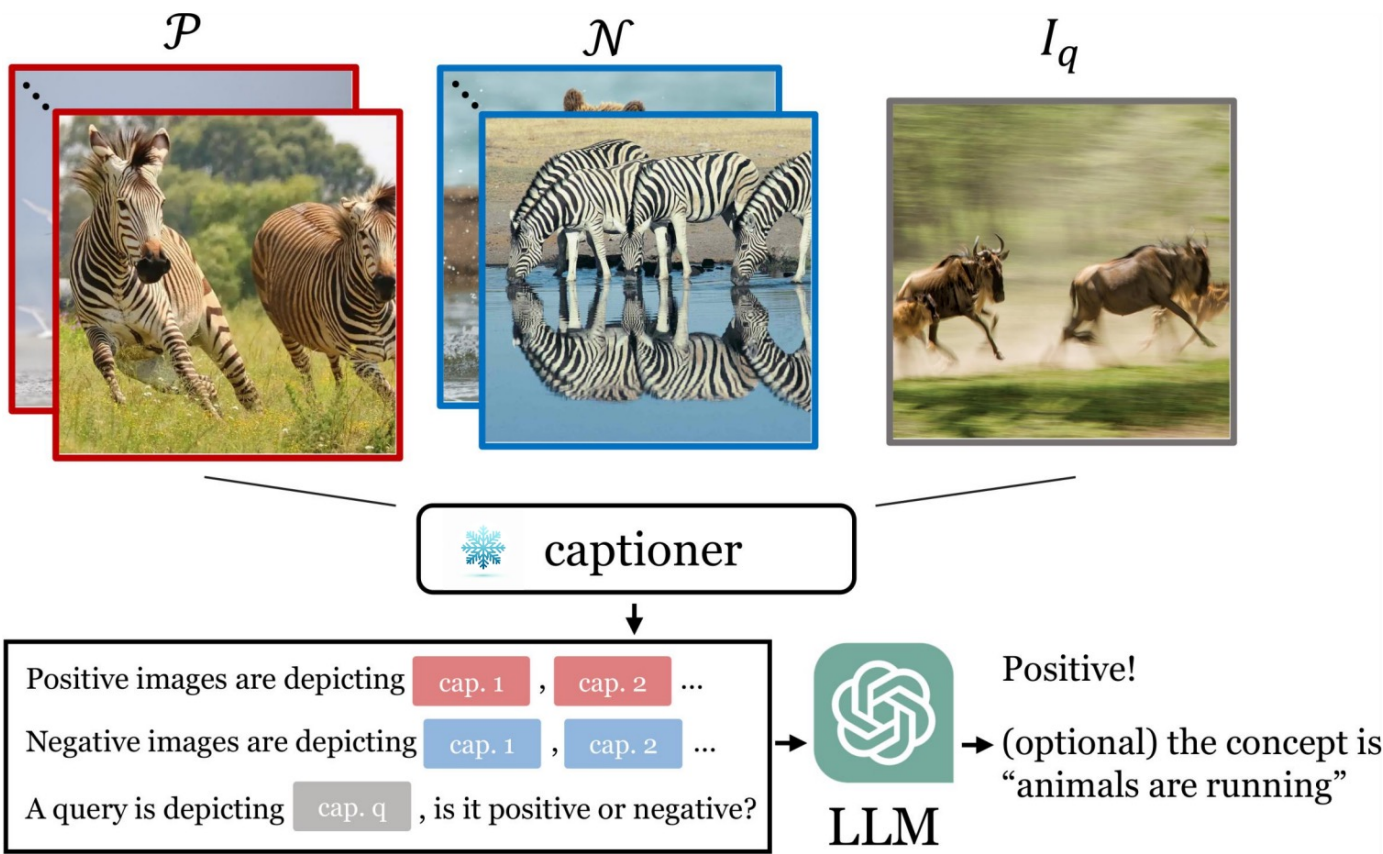
Bongard-OpenWorld (2023)
+ open vocabulary
+ free-form concepts/relationships
+ explicit concept induction

Models for Bongard-OpenWorld?



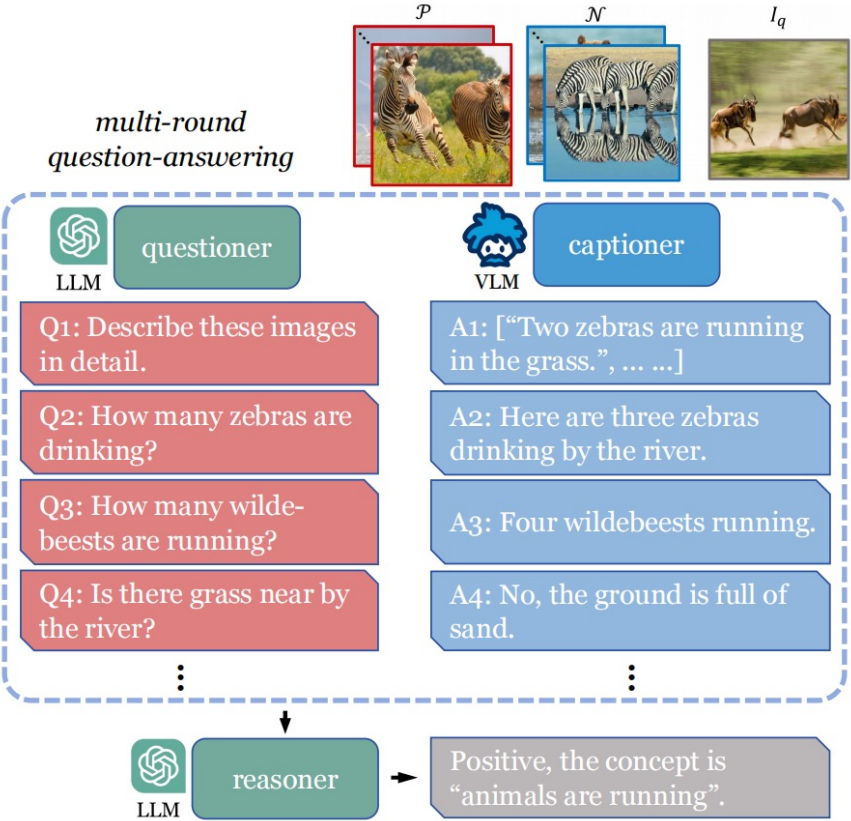
(a) Few-shot learning for Bongard-OpenWorld

Models for Bongard-OpenWorld?



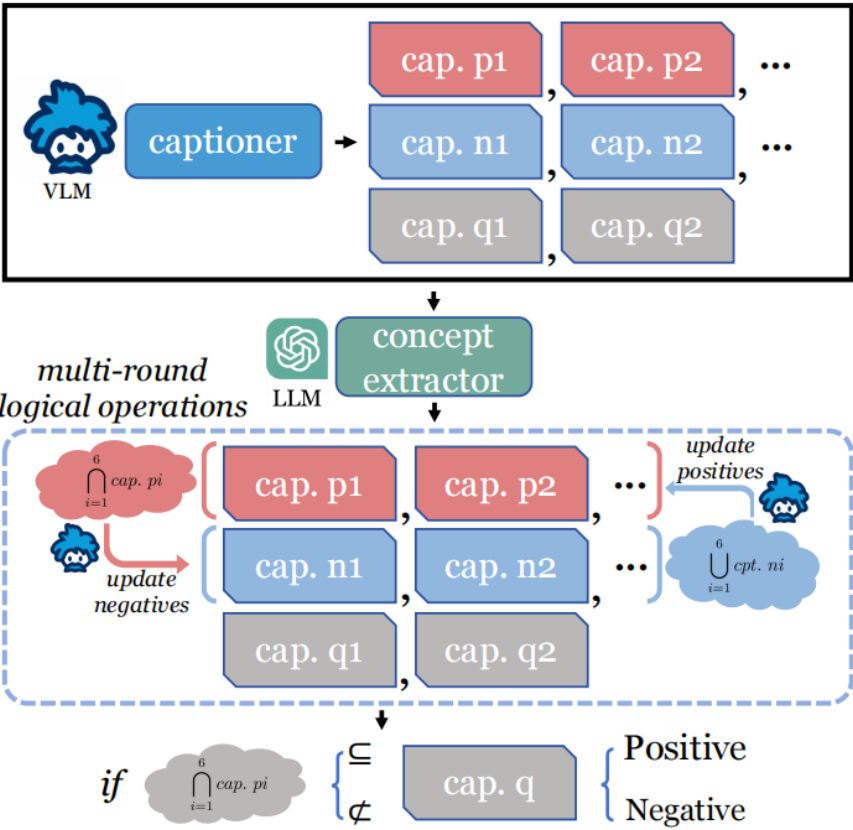
(b) VLM+LLM (single-round) for Bongard-OpenWorld

Models for Bongard-OpenWorld?



(c) VLM+LLM (multi-round) for Bongard-OpenWorld

Models for Bongard-OpenWorld?



(d) Neuro-symbolic approach for Bongard-OpenWorld

What can we learn from the results?

method	image representation	aux. task?	splits				avg.
			short concept	long concept	CS* concept	non-CS* concept	
MetaOptNet [17]	scratch	×	52.3	51.6	54.5	51.0	52.0
	IN-1K	×	60.6	47.3	54.5	54.5	54.5
	IN-22K	×	61.5	51.5	53.6	57.9	56.8
	OpenCLIP	×	63.3	51.6	50.9	60.7	58.0
	OpenCLIP	✓	62.8	51.1	51.8	59.7	57.5
ProtoNet [31]	Challenge of free-form visual concepts. longer, knowledge-extensive	×	57.8	50.5	48.2	56.9	54.5
		×	56.9	54.9	51.8	57.6	56.0
		×	62.4	51.6	54.5	58.6	57.5
		×	61.9	53.8	59.1	57.9	58.3
		✓	59.2	57.7	51.8	61.0	58.5
SNAIL [25]	concepts generally impose greater challenge to the learners.	×	52.8	46.2	50.9	49.3	49.8
		×	61.5	54.9	48.2	62.4	58.5
		×	62.8	57.7	54.5	62.8	60.5
		×	64.2	57.7	57.3	62.8	61.3
		✓	66.1	61.5	63.6	64.1	64.0
ChatGPT [27]	URL	N/A	54.6	52.7	60.0	51.4	53.8
	captions	N/A	60.6	56.6	55.5	60.0	58.8
GPT-4 [4]	captions	N/A	64.5	58.0	57.3	63.2	61.6

What can we learn from the results?

method	image representation	aux. task?	splits				avg.
			short concept	long concept	CS* concept	non-CS* concept	
MetaOptNet [17]	scratch	✗	52.3	51.6	54.5	51.0	52.0
	IN-1K	✗	60.6	47.3	54.5	54.5	54.5
	IN-22K	✗	61.5	51.5	53.6	57.9	56.8
	OpenCLIP	✗	63.3	51.6	50.9	60.7	58.0
	OpenCLIP	✓	62.8	51.1	51.8	59.7	57.5
ProtoNet [31]	scratch	✗	57.8	50.5	48.2	56.9	54.5
	IN-1K	✗	60.6	47.3	51.8	57.6	56.0
	IN-22K	✗	61.5	51.5	53.6	57.9	57.5
	OpenCLIP	✗	63.3	51.6	50.9	60.7	58.3
	OpenCLIP	✓	62.8	51.1	51.8	61.0	58.5
SNAIL [25]	scratch	✗	57.8	50.5	48.2	49.3	49.8
	IN-1K	✗	60.6	47.3	51.8	62.4	58.5
	IN-22K	✗	61.5	51.5	53.6	62.8	60.5
	OpenCLIP	✗	64.2	57.7	57.3	62.8	61.3
	OpenCLIP	✓	66.1	61.5	63.6	64.1	64.0
ChatGPT [27]	URL	N/A	54.6	52.7	60.0	51.4	53.8
	captions	N/A	60.6	56.6	55.5	60.0	58.8
GPT-4 [4]	captions	N/A	64.5	58.0	57.3	63.2	61.6

Open vocabulary representations. from IN-1k to OpenCLIP, more open vocabulary image representations help deliver better results.

What can we learn from the results?

method	image representation	aux. task?	splits				avg.
			short concept	long concept	CS* concept	non-CS* concept	
MetaOptNet [17]	scratch	✗	52.3	51.6	54.5	51.0	52.0
	IN-1K	✗	60.6	47.3	54.5	54.5	54.5
	IN-22K	✗	61.5	51.5	53.6	57.9	56.8
	OpenCLIP	✗	63.3	51.6	50.9	60.7	58.0
	OpenCLIP	✓	62.8	51.1	51.8	59.7	57.5
ProtoNet [31]	scratch	✗	The role of captioning.			56.9	54.5
	IN-1K	✗	1) captioning as an auxiliary task			58.6	56.0
	IN-22K	✗	generally help boost the			59.1	57.5
	OpenCLIP	✗	performances;			51.8	58.3
	OpenCLIP	✓	2) even with off-the-shelf BLIP-2			61.0	58.5
SNAIL [25]	scratch	✗	captions, LLM-based method still			62.4	49.8
	IN-1K	✗	cannot achieve significantly			62.8	58.5
	IN-22K	✗	better results over counterparts.			62.8	60.5
	OpenCLIP	✗	66.1	61.5	63.6	64.1	61.3
	OpenCLIP	✓	66.1	61.5	63.6	64.1	64.0
ChatGPT [27]	URL	N/A	54.6	52.7	60.0	51.4	53.8
	captions	N/A	60.6	56.6	55.5	60.0	58.8
GPT-4 [4]	captions	N/A	64.5	58.0	57.3	63.2	61.6

What can we learn from the results?

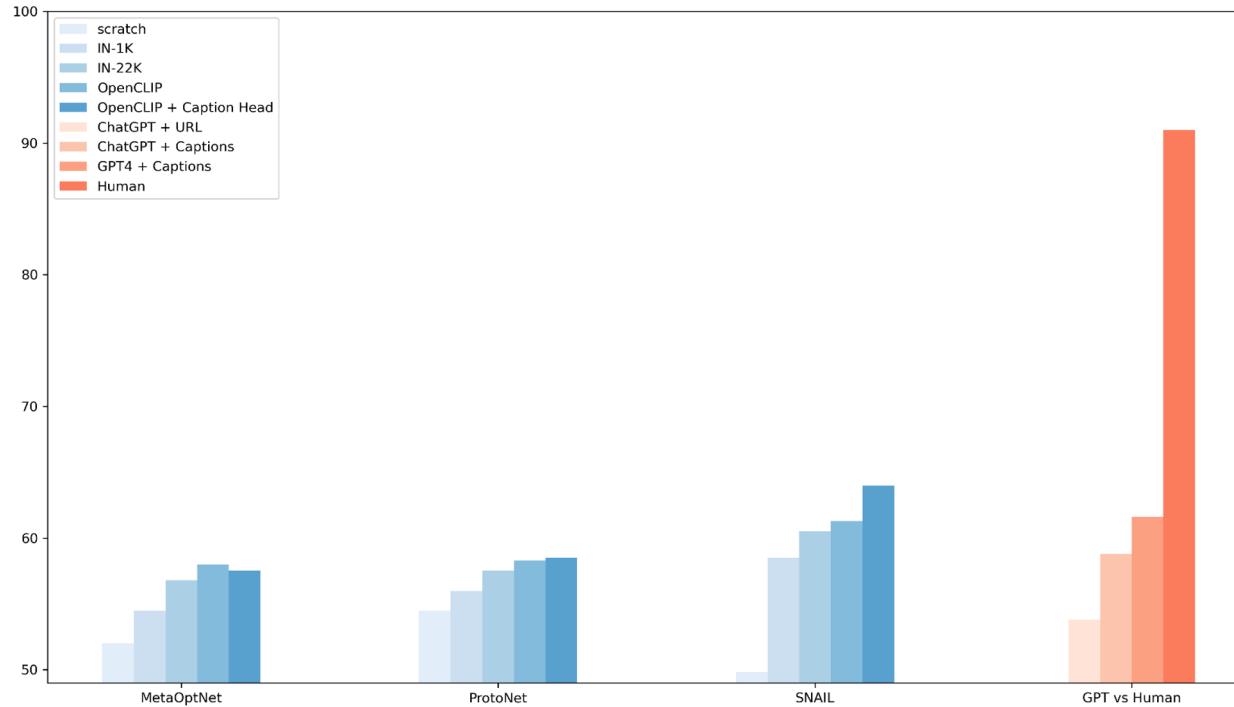
method	image representation	aux. task?	splits				avg.
			short concept	long concept	CS* concept	non-CS* concept	
Meta-Baseline [6]	scratch	✗	56.9	50.0	58.2	52.1	53.8
	IN-1K	✗	58.3	47.3	50.9	54.1	53.3
	IN-22K	✗	59.6	52.7	55.5	56.9	56.5
	OpenCLIP	✗	57.8	52.5	53.6	55.9	55.3
	OpenCLIP	✓	58.3	57.1	64.5	55.2	57.8
MetaOptNet [17]	scratch	✗	52.3	51.6	54.5	51.0	52.0
	IN-1K	✗	60.6	47.3	54.5	54.5	54.5
	IN-22K	✗	61.5	51.5	53.6	57.9	56.8
	OpenCLIP	✗	59.2	51.9	51.8	60.7	58.0
	OpenCLIP	✓	59.2	51.9	59.7	59.7	57.5
ProtoNet [31]	scratch	✗	51.9	53.8	59.1	57.9	54.5
	IN-1K	✗	51.9	53.8	59.1	57.9	56.0
	IN-22K	✗	59.2	57.7	51.8	61.0	57.5
	OpenCLIP	✗	59.2	57.7	51.8	61.0	58.3
	OpenCLIP	✓	59.2	57.7	51.8	61.0	58.5
SNAIL [25]	scratch	✗	50.9	49.3	50.9	49.3	49.8
	IN-1K	✗	62.8	57.7	54.5	62.8	58.5
	IN-22K	✗	62.8	57.7	54.5	62.8	60.5
	OpenCLIP	✗	64.2	57.7	57.3	62.8	61.3
	OpenCLIP	✓	66.1	61.5	63.6	64.1	64.0
ChatGPT [27]	URL	N/A	54.6	52.7	60.0	51.4	53.8
	captions	N/A	60.6	56.6	55.5	60.0	58.8
GPT-4 [4]	captions	N/A	64.5	58.0	57.3	63.2	61.6
Human	N/A	N/A	91.7	90.1	89.1	91.7	91.0

Machine vs. Machine vs Human.

1) Memory-based learner (SNAIL) wins as it might requires less data (we do not offer a huge amount of training episodes as counterparts);

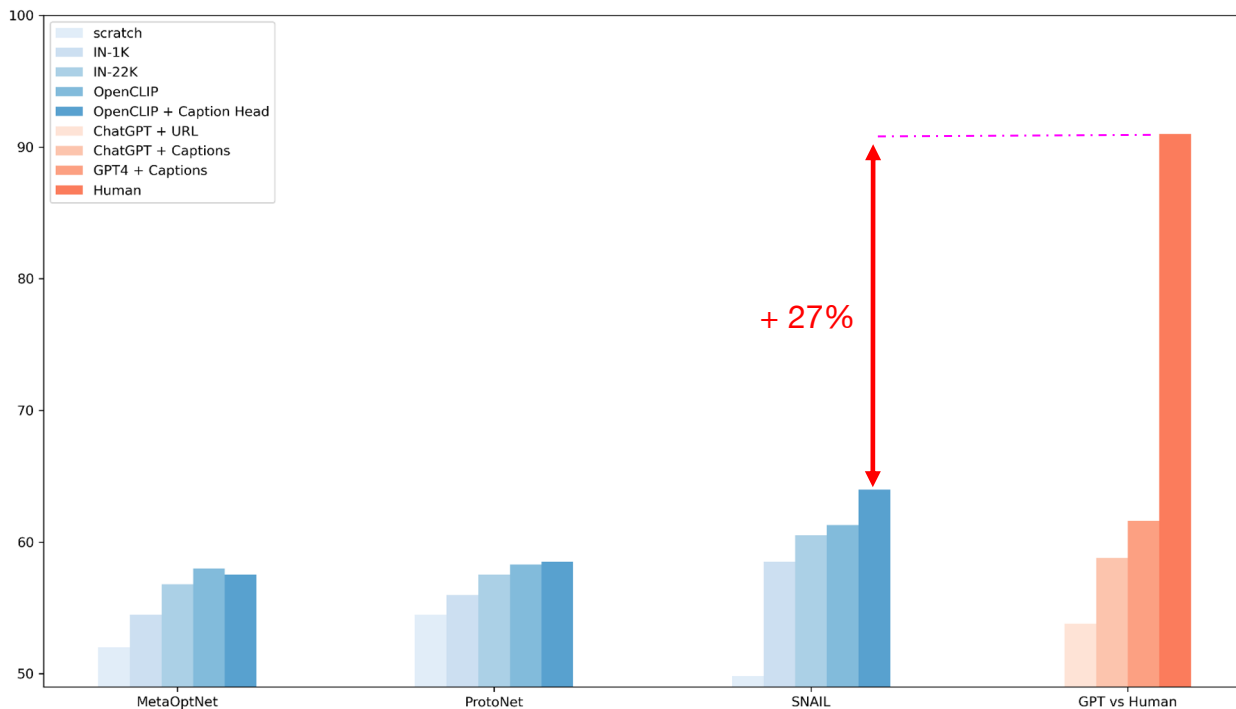
2) The gap to amateur human participants is still quite significant.

What can we learn from the results?



What can we learn from the results?

significant performance gap



Qualitative results

BLIP-2 caption

A woman in a red dress dancing in the street.

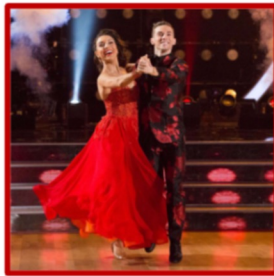
A woman in a red dress jumping in the air.

A group of women in red dresses dancing.

The dancing couple on "dancing with the stars".

A woman in a red dress dancing on a stage in front of a crowd.

\mathcal{P}



I_q

BLIP-2 caption

A woman in a red dress on a runway with crutches.

A woman in a red suit standing next to a woman in a blue suit.

A woman in a red dress is walking down a street.

A woman dancing in a black dress stock photo.

ground truth concept c :
A woman dancing in a red dress.









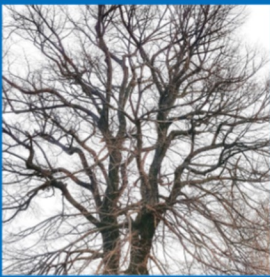
\mathcal{N}



GPT-4 response:
Positive;
A woman in a red dress dancing.

(a) GPT-4 correctly produces both binary prediction and induced visual concept.

Qualitative results

BLIP-2 caption	A snow covered tree in the middle of a forest.	A tree branch covered in snow.	A tree covered in snow against a blue sky.	A tree branch covered in ice and snow.	A red bird sitting on top of a tree branch.
\mathcal{P}					
BLIP-2 caption	The sun is shining through the leaves of a tree.	The top of a tree branch.	A cherry blossom tree in the water.	A bench sitting next to a tree in a park.	I_q
\mathcal{N}					<p>ground truth concept c: <i>Tree branches covered by snow.</i></p>
					<p>GPT-4 response: <i>Negative; A tree or branch covered in snow.</i></p>

(b) BLIP-2 only covers unhelpful content of I_q , GPT-4 makes correct concept induction but fails on binary prediction.

Qualitative results

BLIP-2 caption

A bridge over a body of water at night.

A view of the brooklyn bridge over water at night.

A bridge over a river with fireworks in the night sky.

A bridge over the river at night with a street light on it.

A bench on a bridge at night.

\mathcal{P}



I_q

BLIP-2 caption

A woman standing on a bridge over water.

People walking across a bridge at night.

A view of the golden gate bridge from the water.

A black and white photo of a bridge over a river.

\mathcal{N}



ground truth concept c :
A bridge over the water at night.

GPT-4 response:
Positive;
A bridge at night.

(c) GPT-4 fails on both concept induction and binary prediction due to hard negatives.

Takeaway

We present Bongard-OpenWorld, a benchmark that reconciles few-shot learning and visual reasoning using free-form visual concepts and real world images.

We carefully construct Bongard-OpenWorld problems with interesting visual concepts crawled from the web & augmented by human writers.

Bongard-OpenWorld imposes great challenge to canonical few-shot learners and LLM-based zero-shot learners.

Code & Data: [Bongard-OpenWorld](#)

