Bongard-OpenWorld:

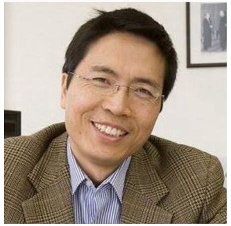
Few-Shot Reasoning for Free-form Visual Concepts in the Real World

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

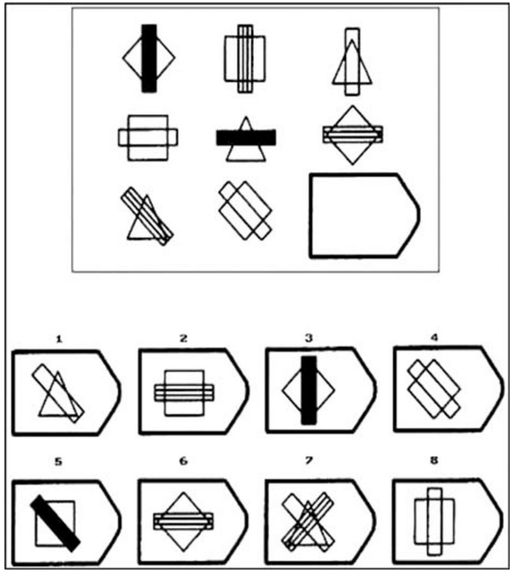




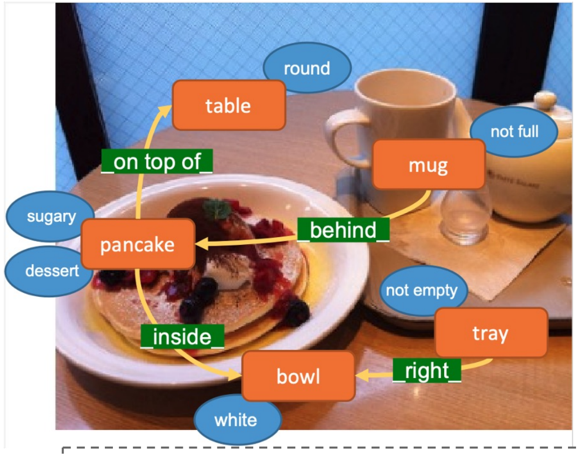




Few-shot learning and visual reasoning: niche of visual intelligence

Few-shot learning in visual recognition



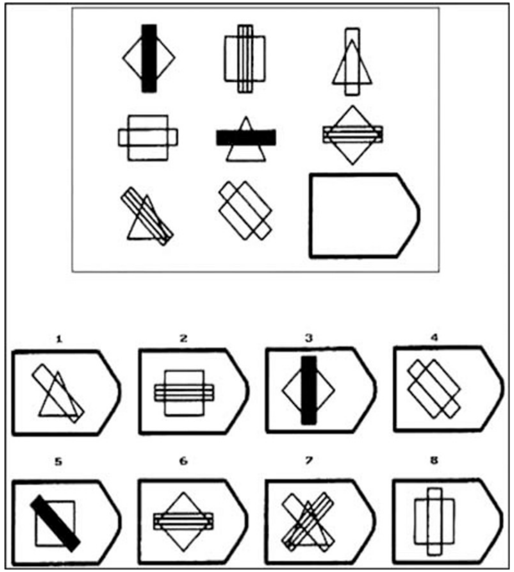
Visual IQ test or abstract visual reasoning

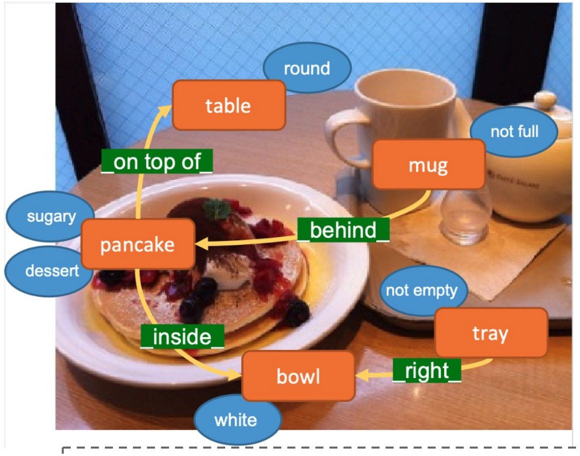


Q: What is inside the thing that to the right of the tray?

Relational reasoning, ex. VQA

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





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

few-shot induction identifying the *visual concept* (category) with a handful of examples;

meta learning generalizing this to novel concepts.

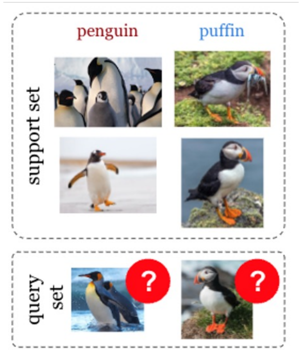
few-shot induction inferring the hidden *patterns* from very few demonstrations;

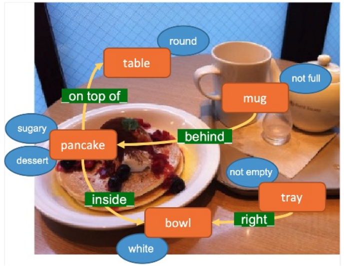
relational learning the *patterns* sometime control multiple images rather than just one.

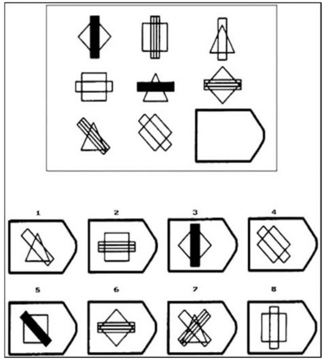
relational learning need to answer questions about the relationships between entities (objects);

meta learning generalization to novel relationships.

Reconciling few-shot learning and visual reasoning







Few-shot induction

Inducing arbitrary concepts, just from very few examples.

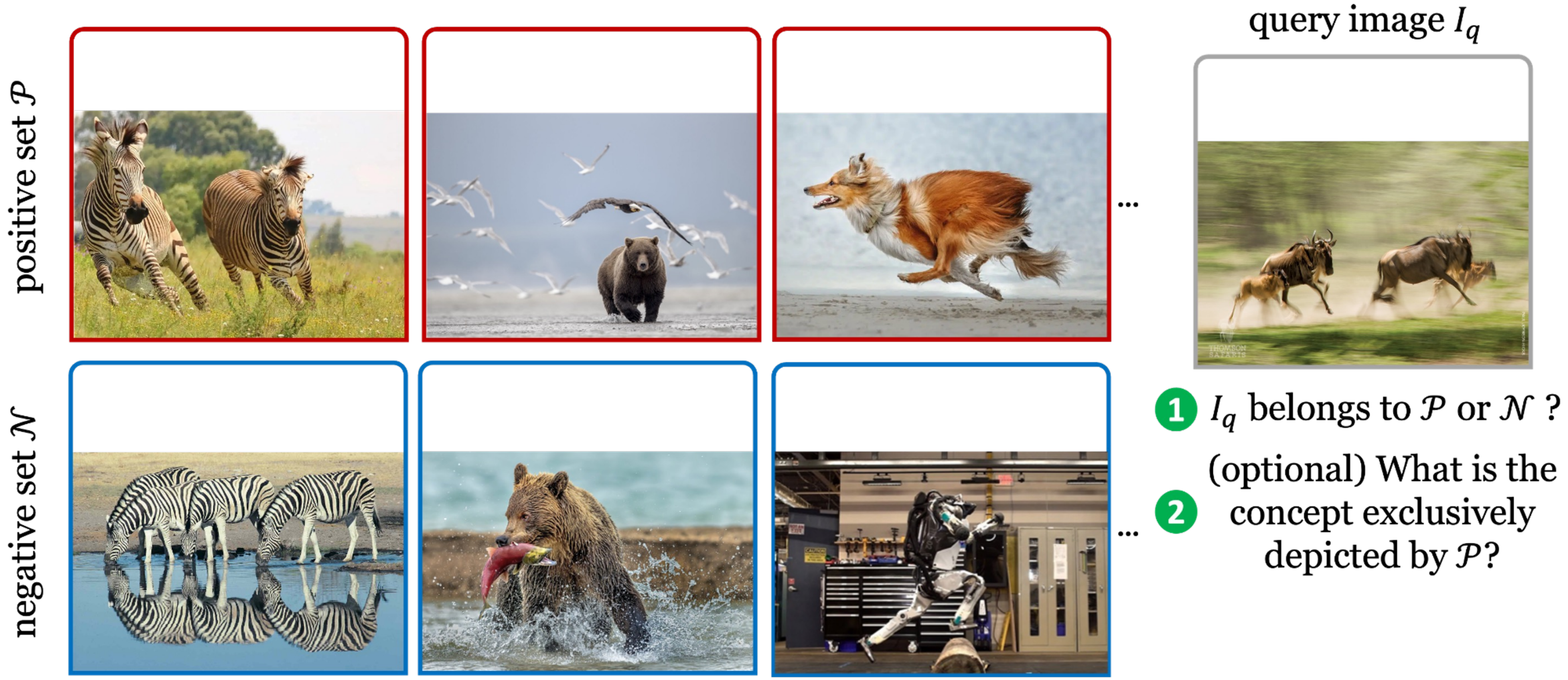
Relational learning

Learning to comprehend and extend relationships among entities.

Meta learning

Generalizing the learned “learning algorithm” to novel scenarios.

Introducing Bongard-OpenWorld



Introducing Bongard-OpenWorld

Reconciling few-shot learning and visual reasoning

Few-shot induction

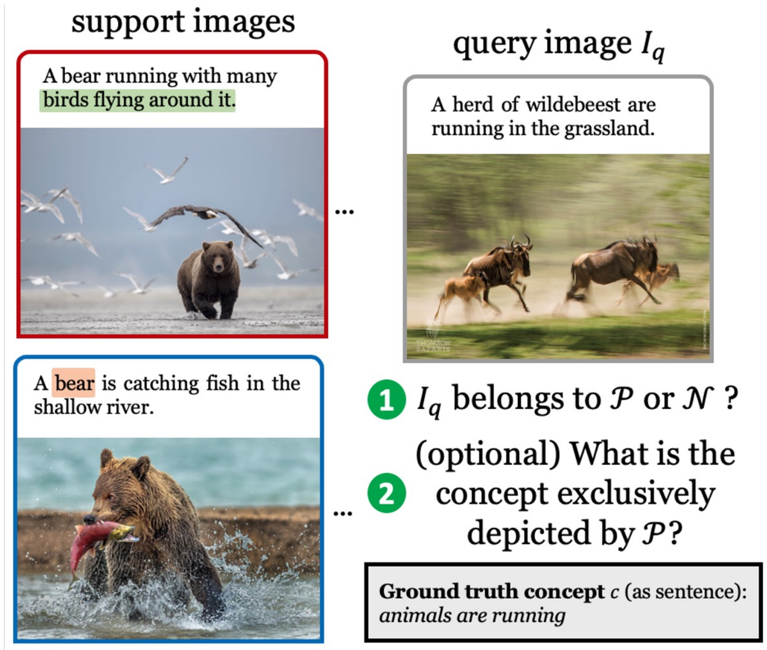
Inducing arbitrary concepts, but just from very few examples.

Relational learning

Learning to comprehend and extend relationships among entities.

Meta learning

Generalizing the learned “learning algorithm” to novel scenarios.



disjoint test set (not shown)

Constructing Bongard-OpenWorld

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

1

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)

2

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.

concept augmentation

3

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

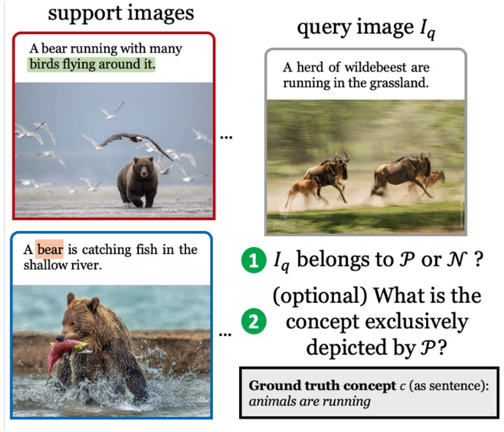
negatives
child playing dolls in the room
child playing video games
child playing with dogs
a dog *playing with ball*
(*partially overlapping*)

Constructing Bongard-OpenWorld

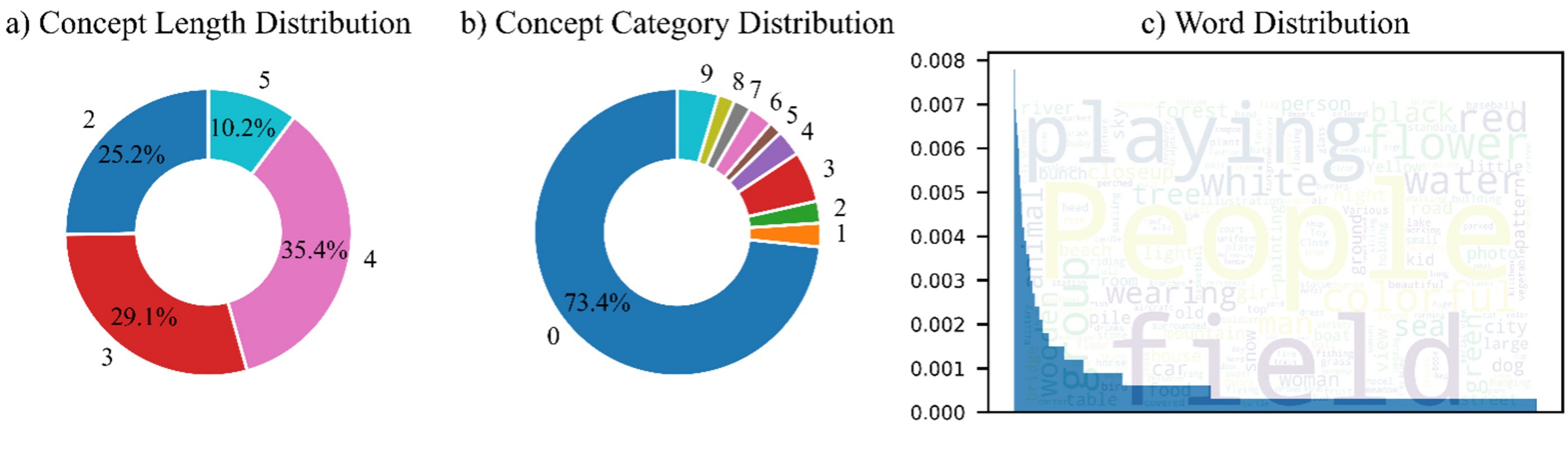
positives
child playing ball with friends
child playing ball on the beach
child playing ball in the pool
child play ball with a coach
(distractors)

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*)

Google Images
gettyimages
flickr

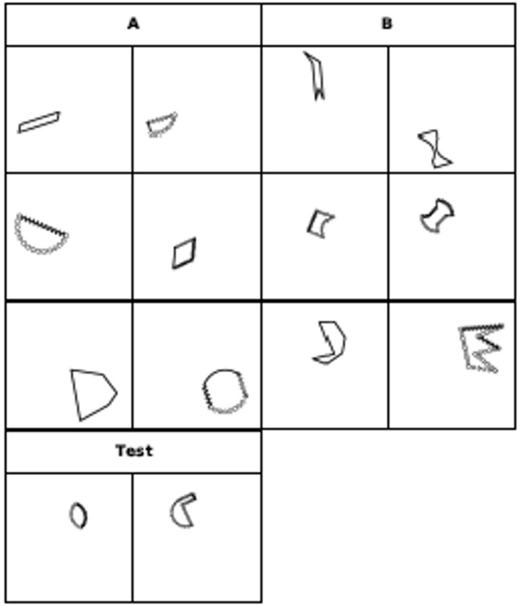


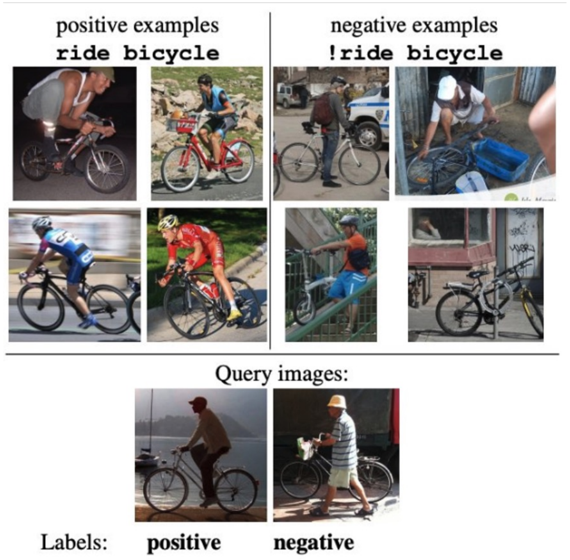
Constructing Bongard-OpenWorld

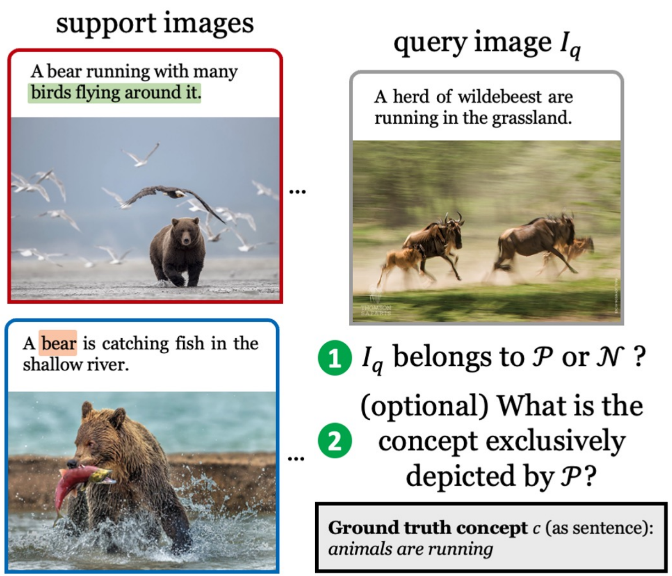


Concept Category	ID	Example	Concept Category	ID	Example
Anything else (non-CS*)	0	Animals are running.	And / Or / Not	5	A man without beard.
HOI	1	A person playing the guitar.	Factual Knowledge	6	A building in US capital.
Taste / Nutrition	2	A plate of high-calorie food.	Meta Class	7	Felidae animals.
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





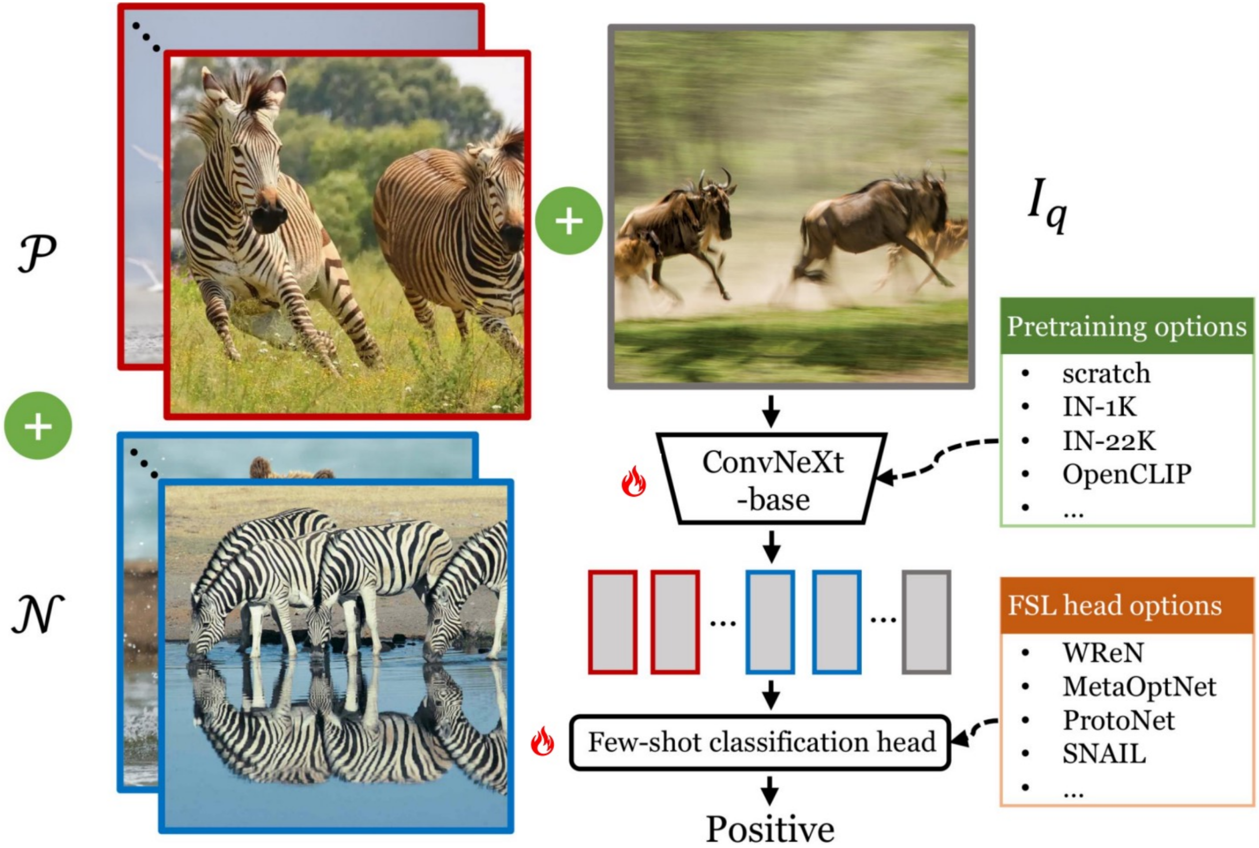


Bongard-LOGO (2020)
(very close to the original Bongard problem)

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

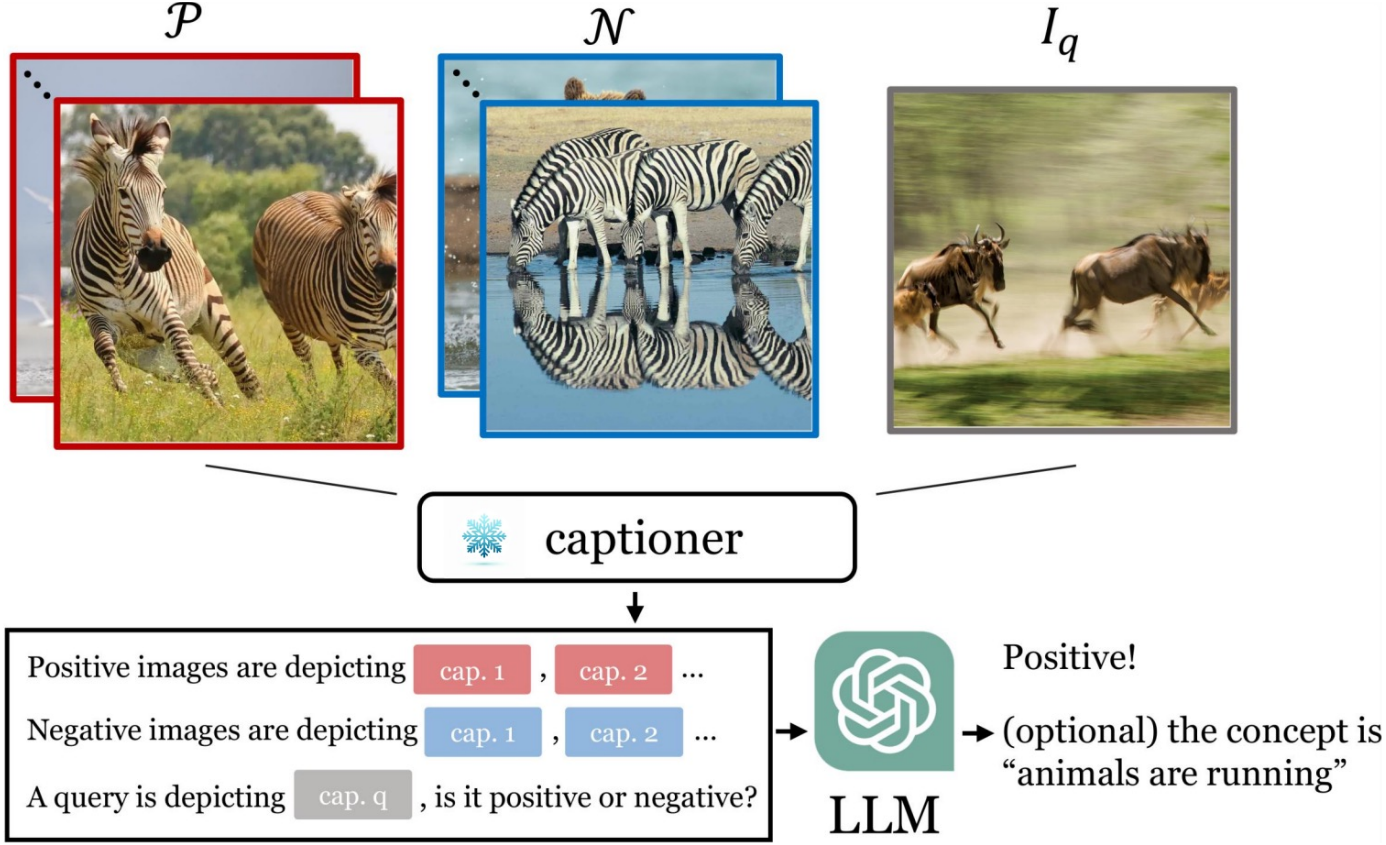
Bongard-OpenWorld (2023)
\+ open vocabulary
\+ free-form concepts/relationships
\+ explicit concept induction

Models for Bongard-OpenWorld?



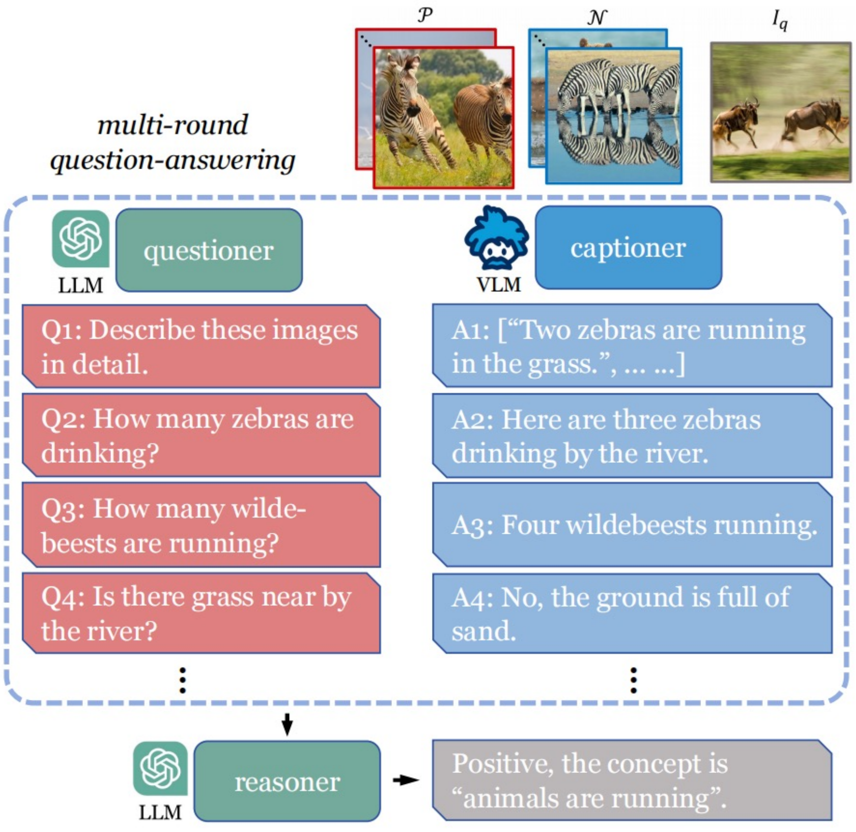
(a) Few-shot learning for Bongard-OpenWorld

Models for Bongard-OpenWorld?



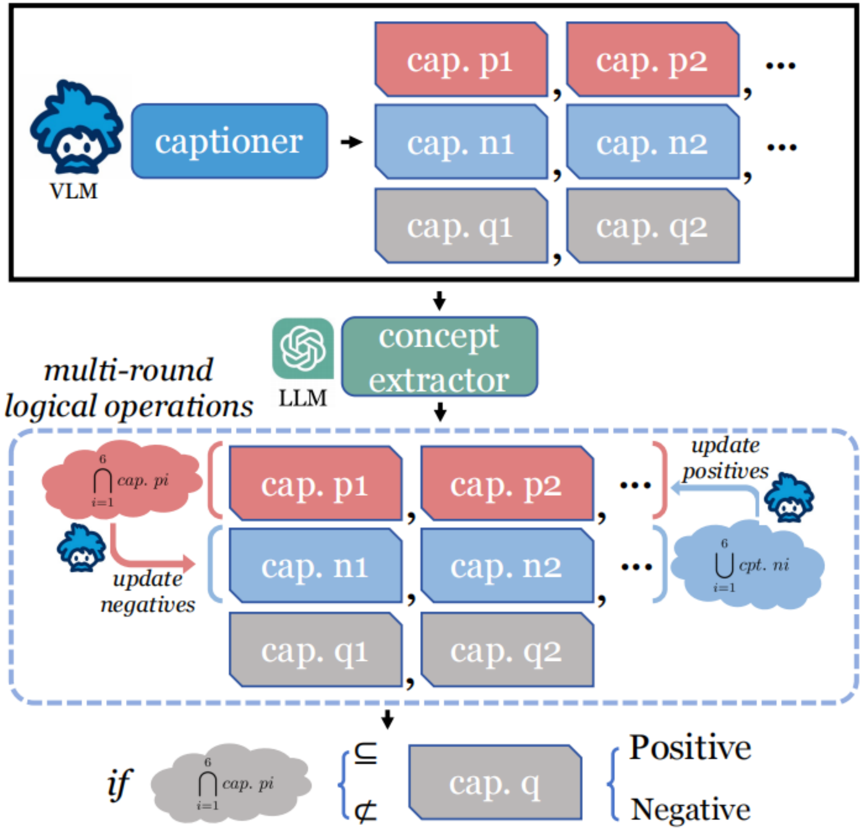
(b) VLM+LLM (single-round) for Bongard-OpenWorld

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: short concept	long concept	CS* concept	non-CS* concept	avg.
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	✗	56.9	54.9	51.8	57.6	56.0
	IN-22K	✗	62.4	51.6	54.5	58.6	57.5
	OpenCLIP	✗	61.9	53.8	59.1	57.9	58.3
	OpenCLIP	✓	59.2	57.7	51.8	61.0	58.5
SNAIL [25]	scratch	✗	52.8	46.2	50.9	49.3	49.8
	IN-1K	✗	61.5	54.9	48.2	62.4	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

Challenge of free-form visual concepts. longer, knowledge-extensive concepts generally impose greater challenge to the learners.

What can we learn from the results?

method	image representation	aux. task?	splits: short concept	long concept	CS* concept	non-CS* concept	avg.
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	✗	[illegible]	[illegible]	51.8	57.6	56.0
	IN-22K	✗	62.4	51.6	54.5	58.6	57.5
	OpenCLIP	✗	61.9	53.8	59.1	57.9	58.3
	OpenCLIP	✓	59.2	57.7	51.8	61.0	58.5
SNAIL [25]	scratch	✗	52.8	46.2	50.9	49.3	49.8
	IN-1K	✗	61.5	54.9	48.2	62.4	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

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: short concept	splits: long concept	splits: CS* concept	splits: non-CS* concept	avg.
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	✗	[illegible]	[illegible]	[illegible]	56.9	54.5
	IN-1K	✗	[illegible]	[illegible]	[illegible]	[illegible]	56.0
	IN-22K	✗	62.4	51.6	54.5	58.6	57.5
	OpenCLIP	✗	61.9	53.8	59.1	57.9	58.3
	OpenCLIP	✓	59.2	57.7	51.8	61.0	58.5
SNAIL [25]	scratch	✗	52.8	46.2	50.9	49.3	49.8
	IN-1K	✗	61.5	54.9	48.2	62.4	58.5
	IN-22K	✗	62.8	57.7	54.5	62.8	60.5
	OpenCLIP	✗	64.2	57.7	57.5	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

The role of captioning.
1) captioning as an auxiliary task generally help boost the performances;
2) even with off-the-shelf BLIP-2 captions, LLM-based method still cannot achieve significantly better results over counterparts.

What can we learn from the results?

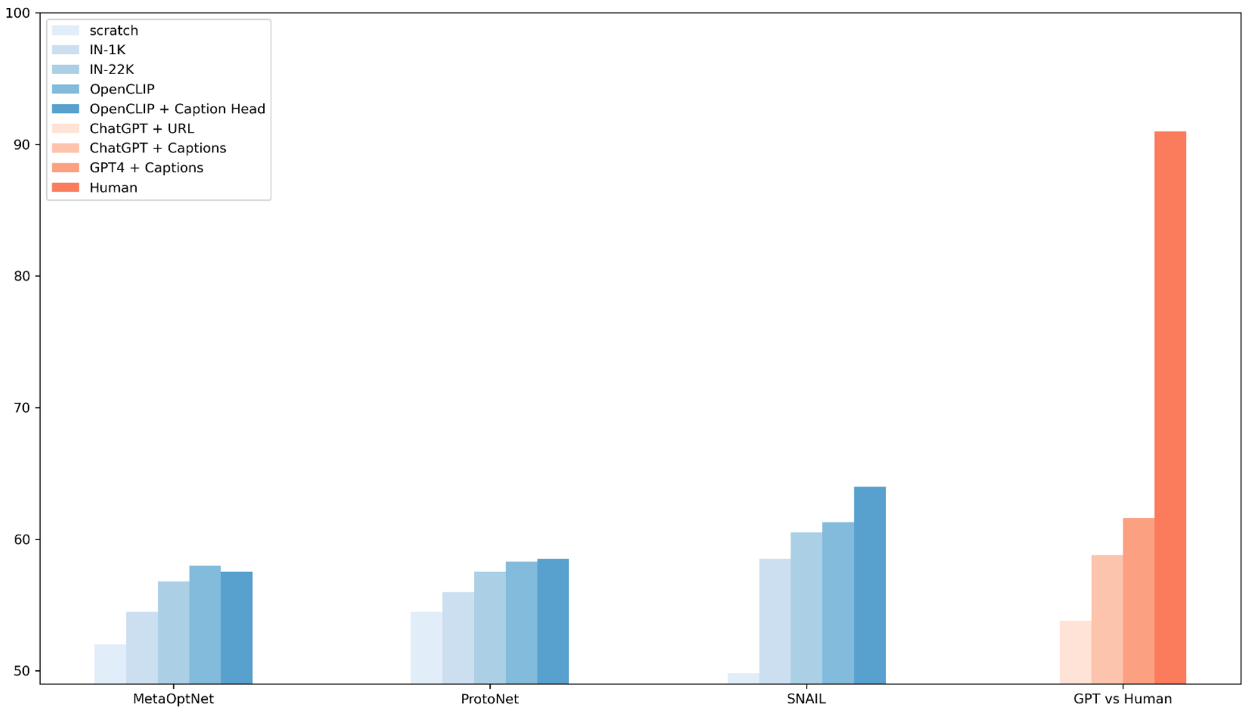
method	image representation	aux. task?	splits: short concept	splits: long concept	splits: CS* concept	splits: non-CS* concept	avg.
Meta-Baseline [6]	scratch	✗	56.9	50.0	58.2	52.1	53.8
	IN-1K	✗	58.3	47.3	50.9	54.1	53.3
	IN-22K	✗	59.6	52.7	55.5	56.9	56.5
	OpenCLIP	✗	57.8	52.5	53.6	55.9	55.3
	OpenCLIP	✓	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
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	60.7	58.0
	OpenCLIP	[illegible]	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	[illegible]	56.9	54.9	51.8	57.6	56.0
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	57.5
	OpenCLIP	✗	61.9	53.8	59.1	57.9	58.3
	OpenCLIP	[illegible]	59.2	57.7	51.8	61.0	58.5
SNAIL [25]	scratch	[illegible]	[illegible]	[illegible]	50.9	49.3	49.8
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	62.4	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);

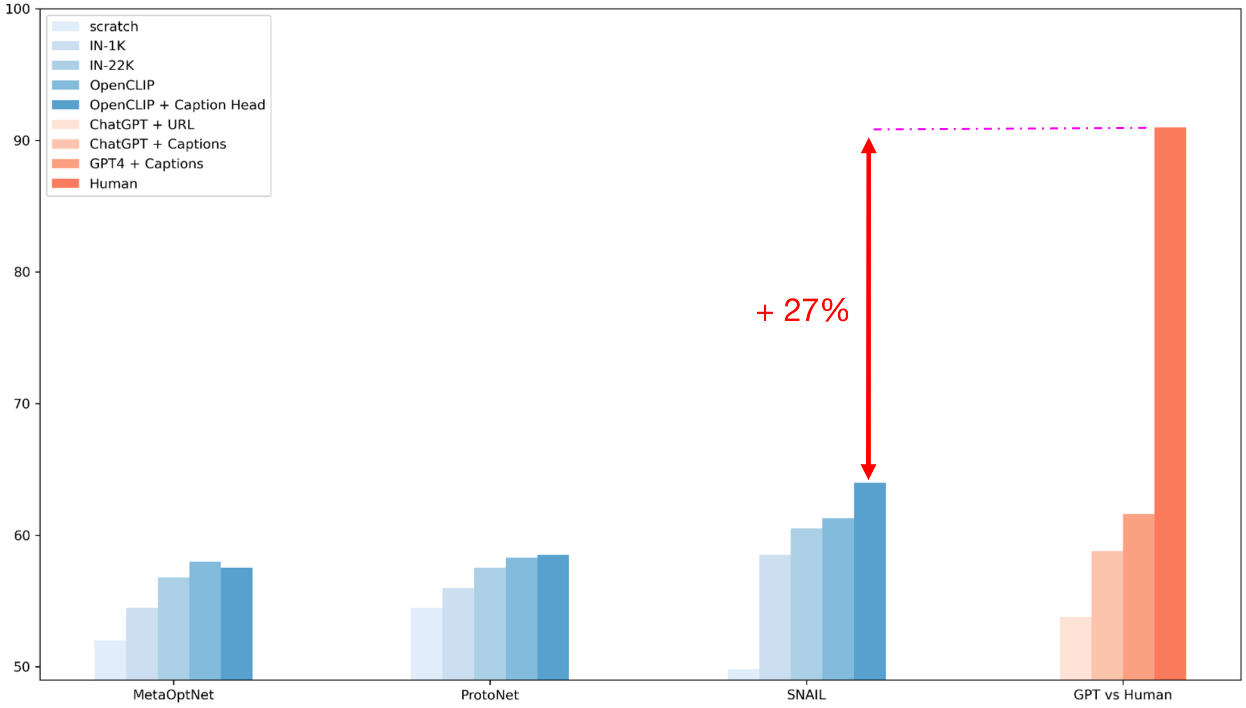
2) The gap to amateur human participants is still quite significant.

What can we learn from the results?

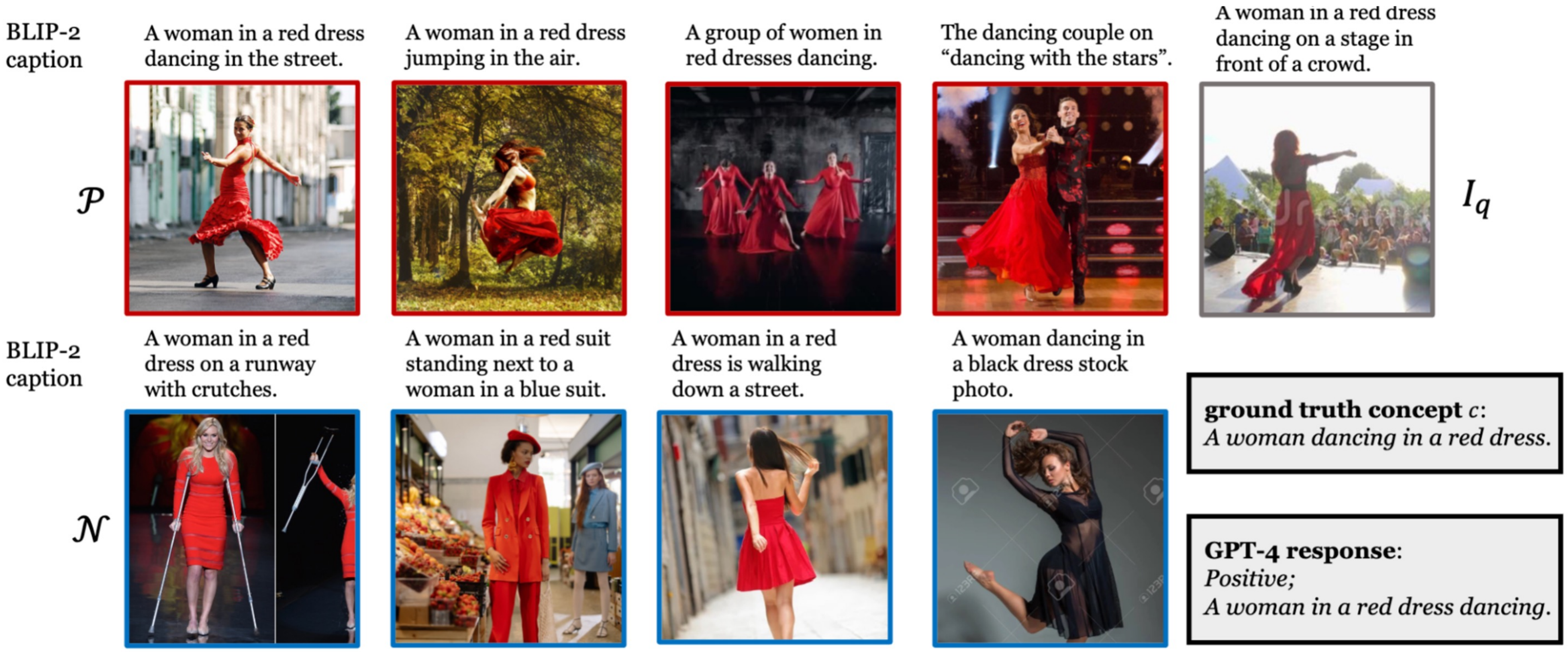


What can we learn from the results?

significant performance gap

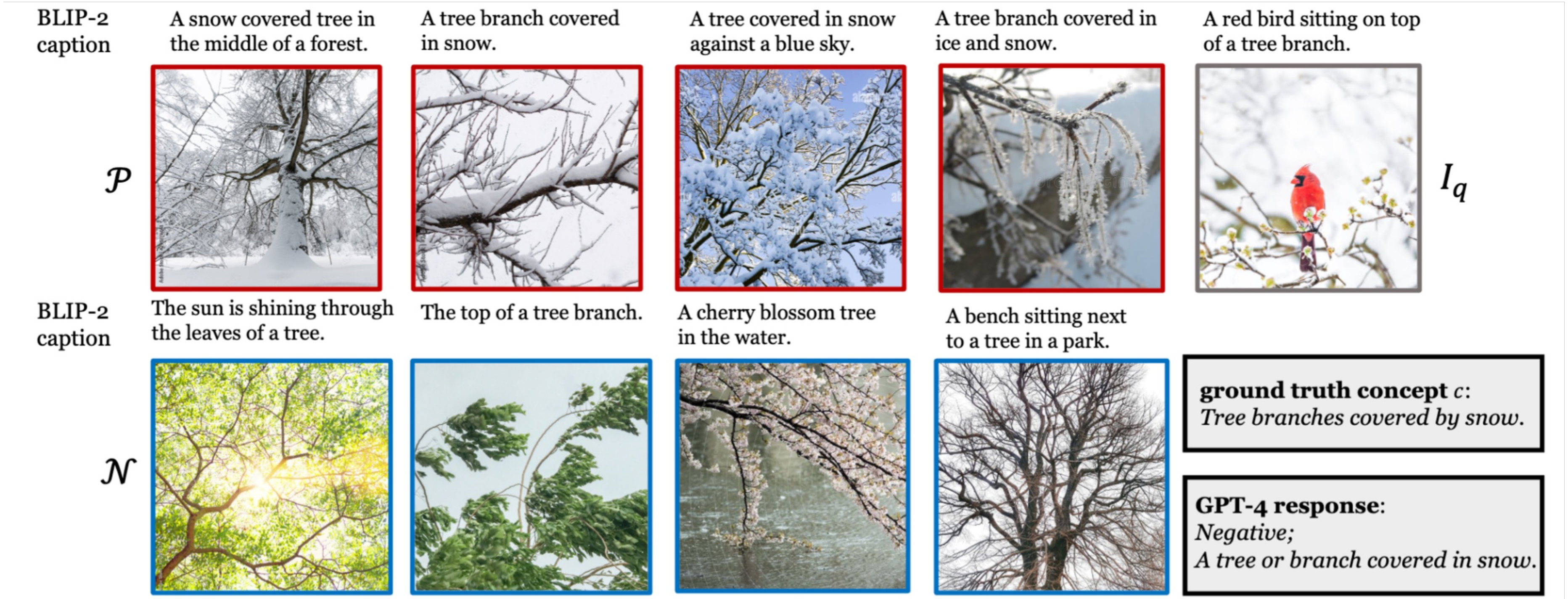


Qualitative results



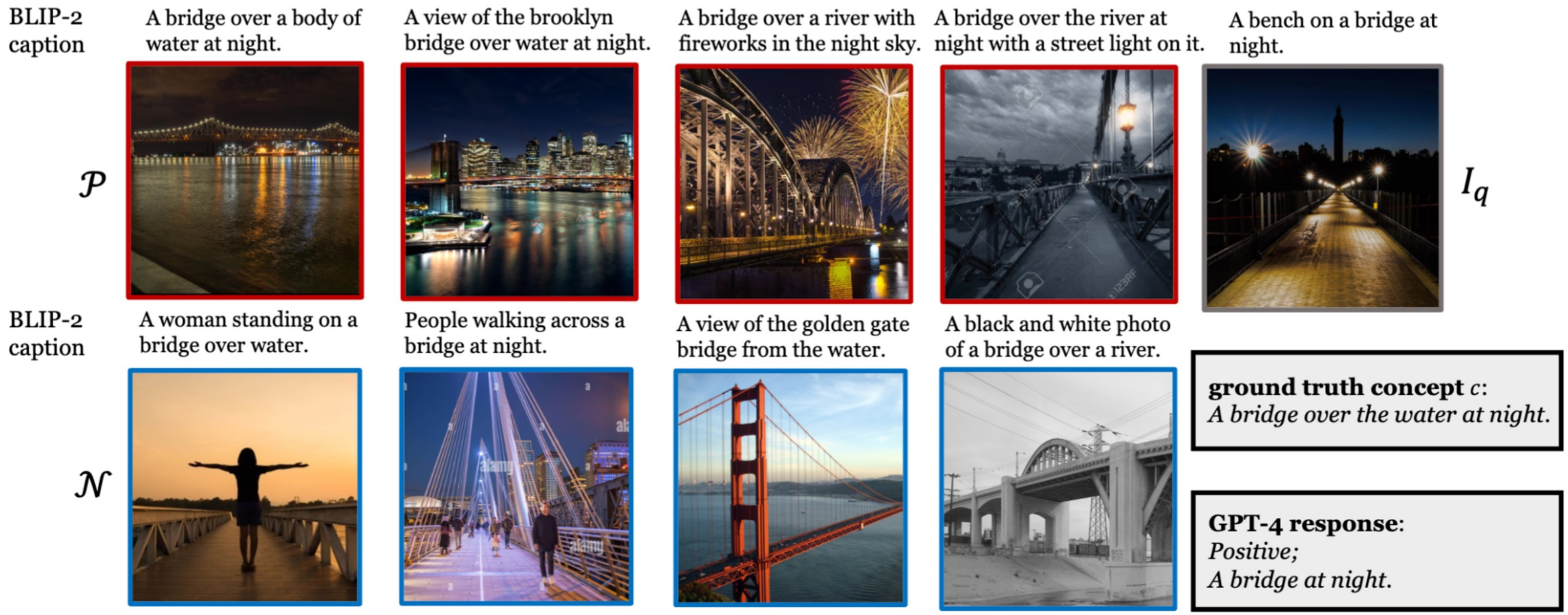
(a) GPT-4 correctly produces both binary prediction and induced visual concept.

Qualitative results



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

Qualitative results



(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