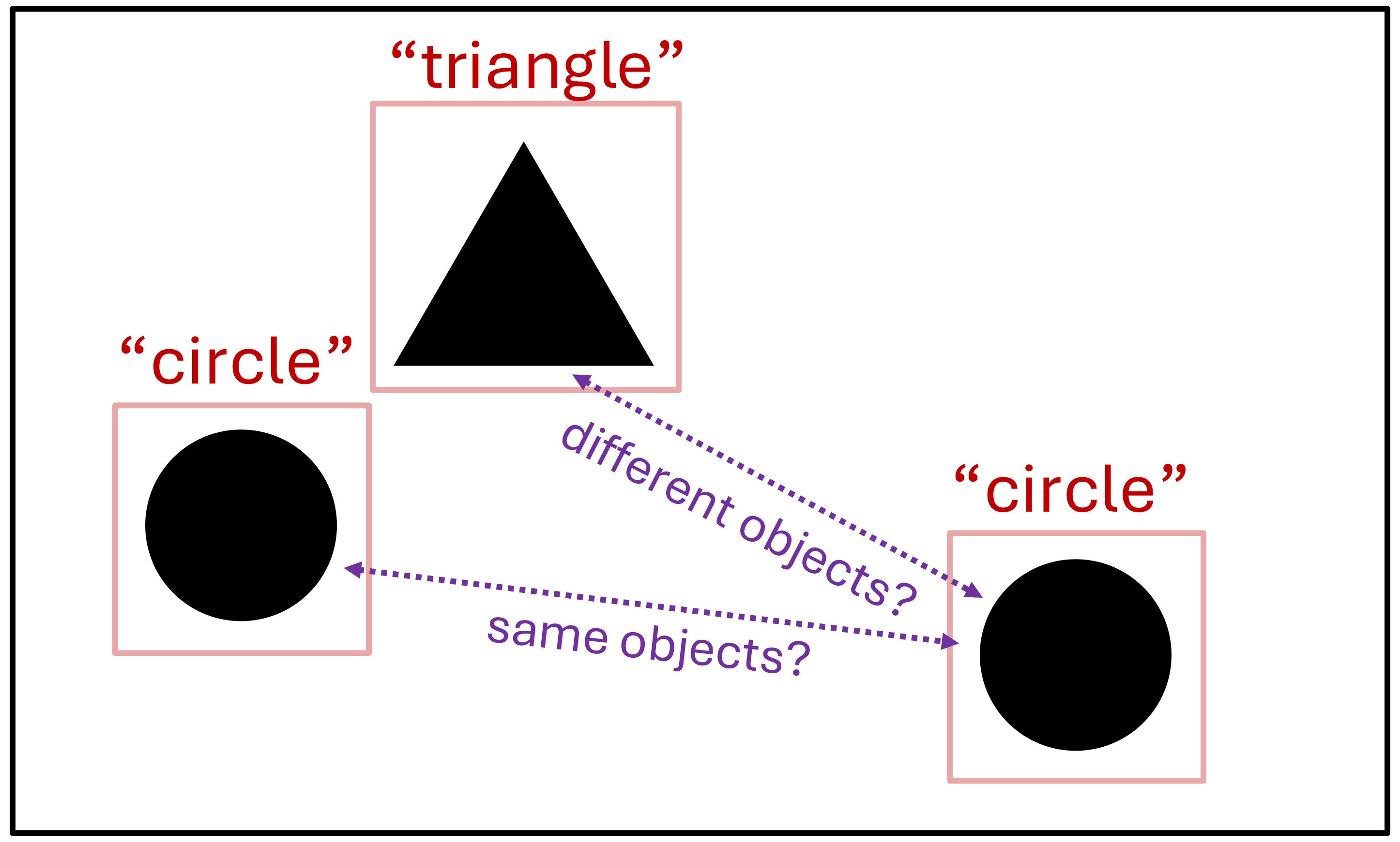
Mind the GAP: Glimpse-based Active Perception improves generalization and sample efficiency of visual reasoning

Oleh Kolner[1,2], Thomas Ortner[1], Stanisław Woźniak[1], Angeliki Pantazi[1]
[1]IBM Research Europe
[2]Institute of Machine Learning and Neural Computation, Graz University of Technology

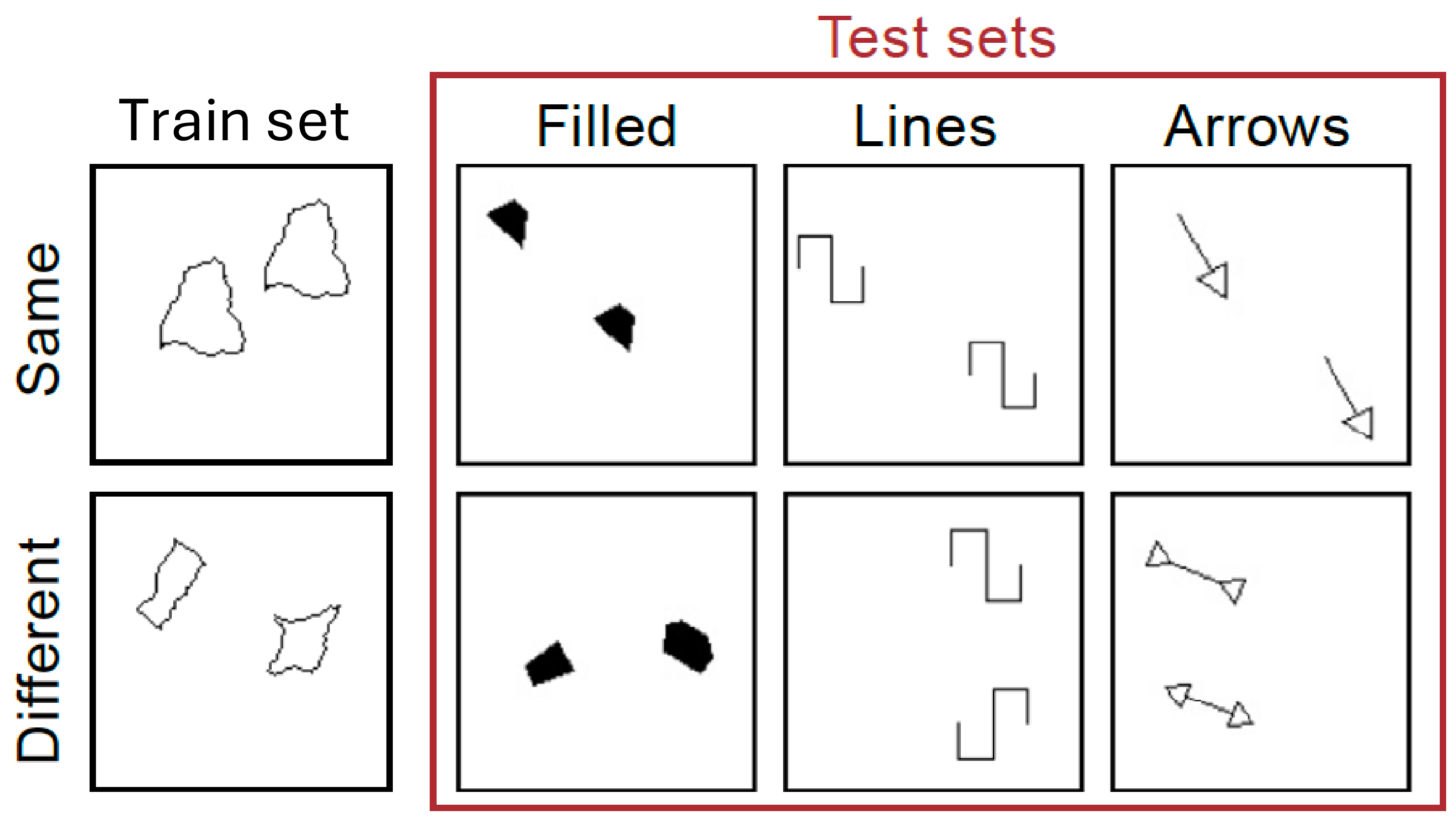
Motivation

Modern vision models can recognize content in an image but cannot reliably reason about relations between content's parts

- Typically, models just overfit to specific visual patterns failing to capture the underlying image structure



Models fail at comparing two out-of-distribution (OOD) objects, i.e. if objects significantly differ from objects from the train set

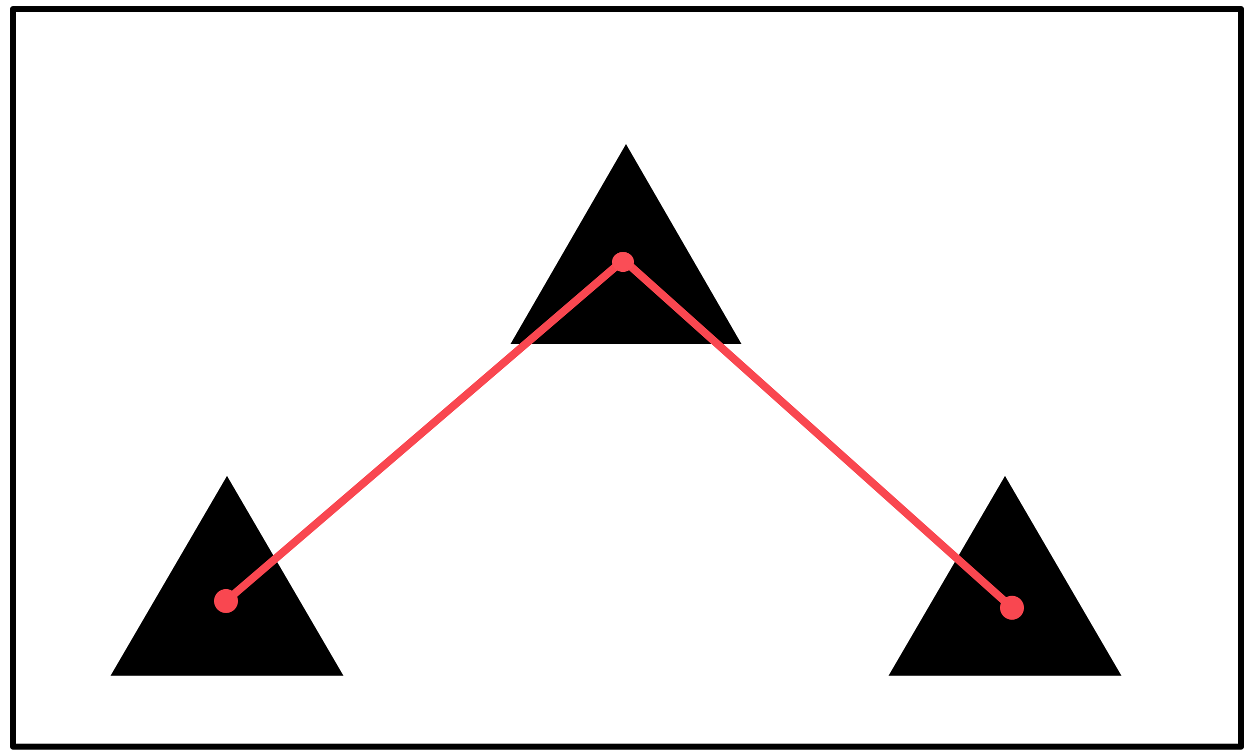


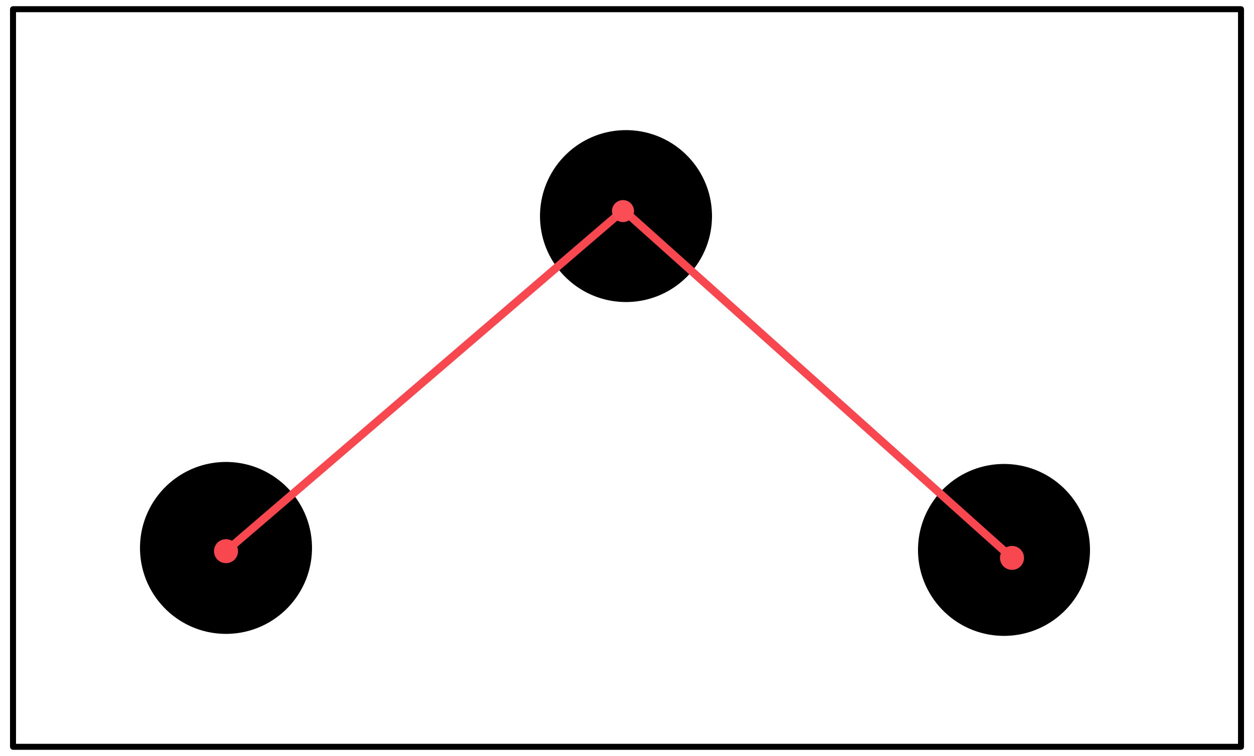
Inspiration from human brain

Typical AI systems process images all at once

In contrast, humans process images sequentially by moving their eyes to their important parts

Information about the eye movements helps to understand image structure beyond the visual content





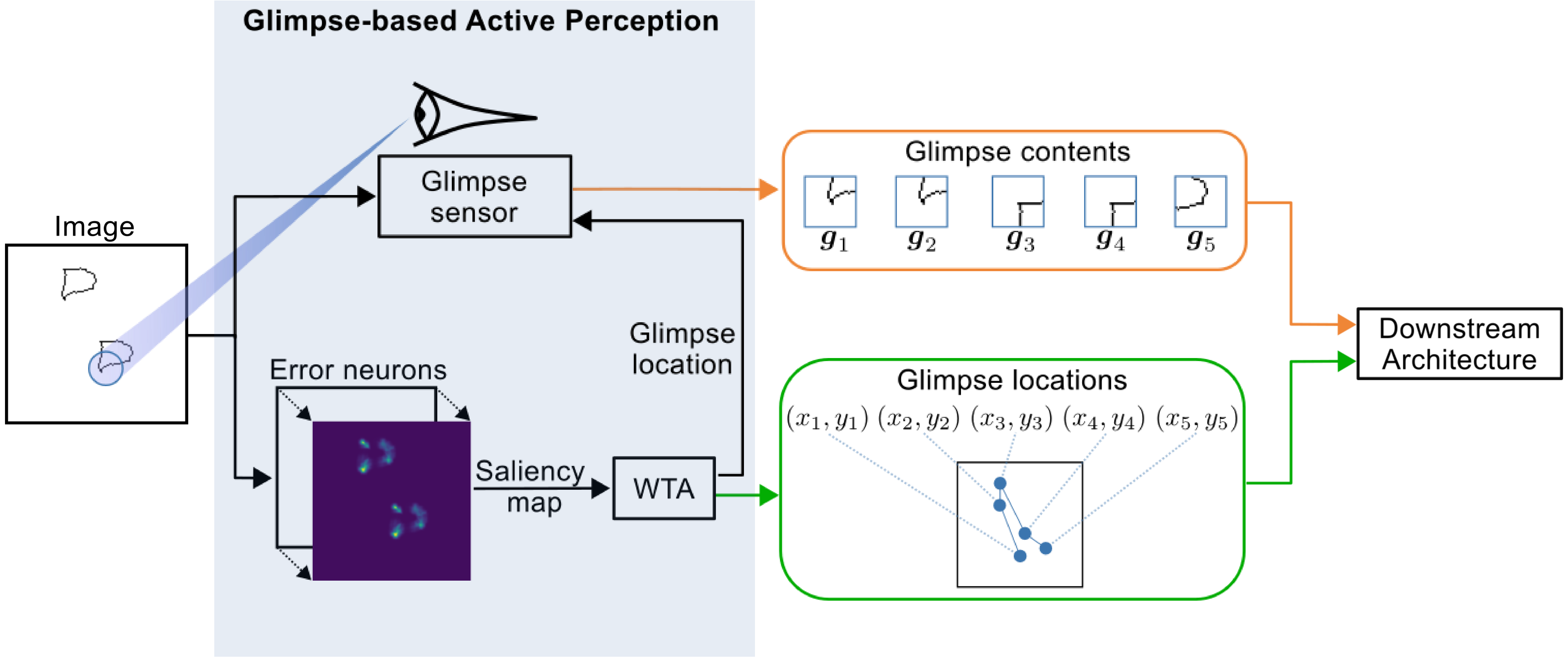
Approach

Error neurons extract a saliency map that highlights most important image parts whose locations (=glimpse locations) are obtained in a winner-takes-all (WTA) manner

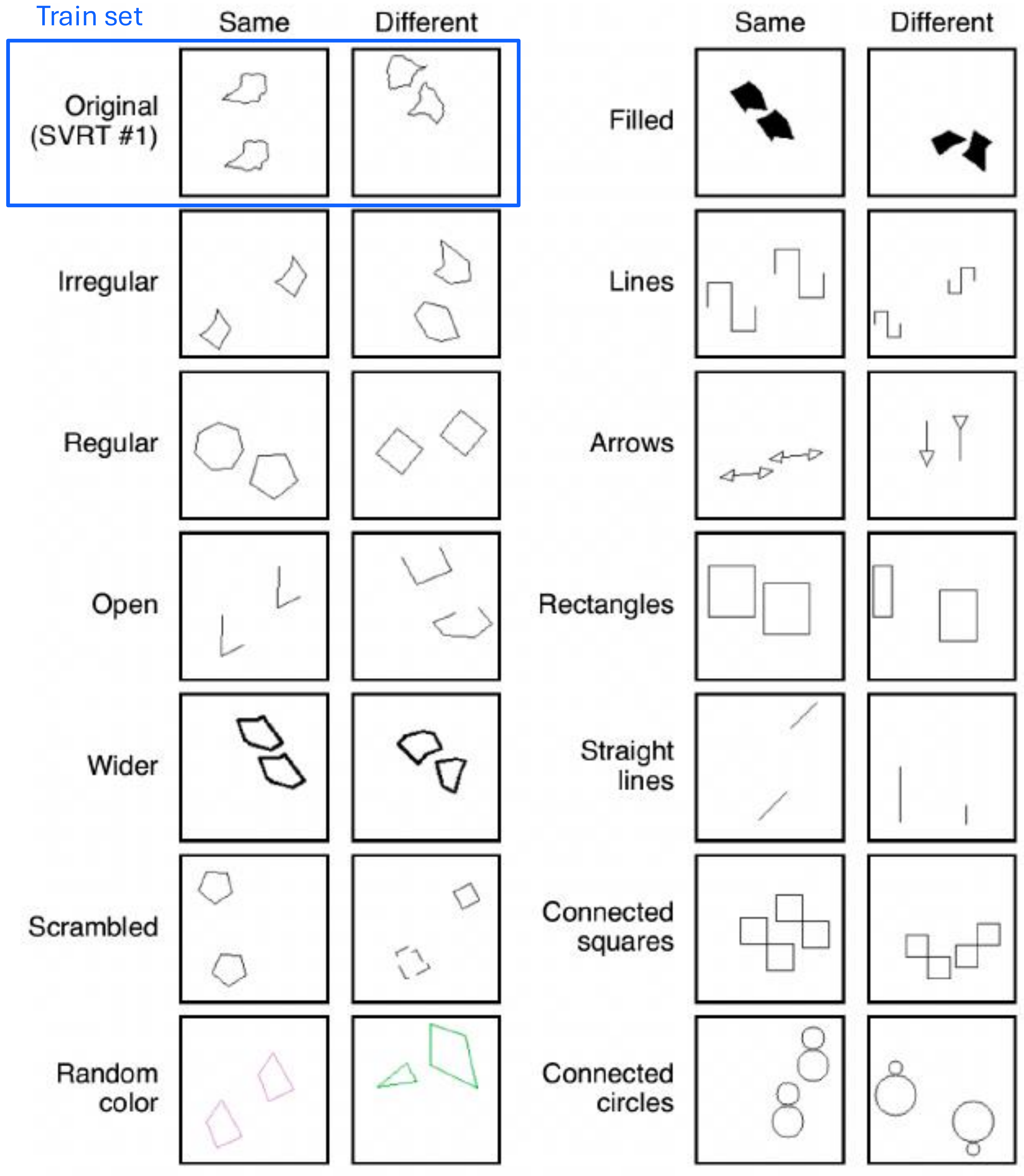
Glimpse sensor extracts the local visual content (=glimpse content) at the glimpse locations

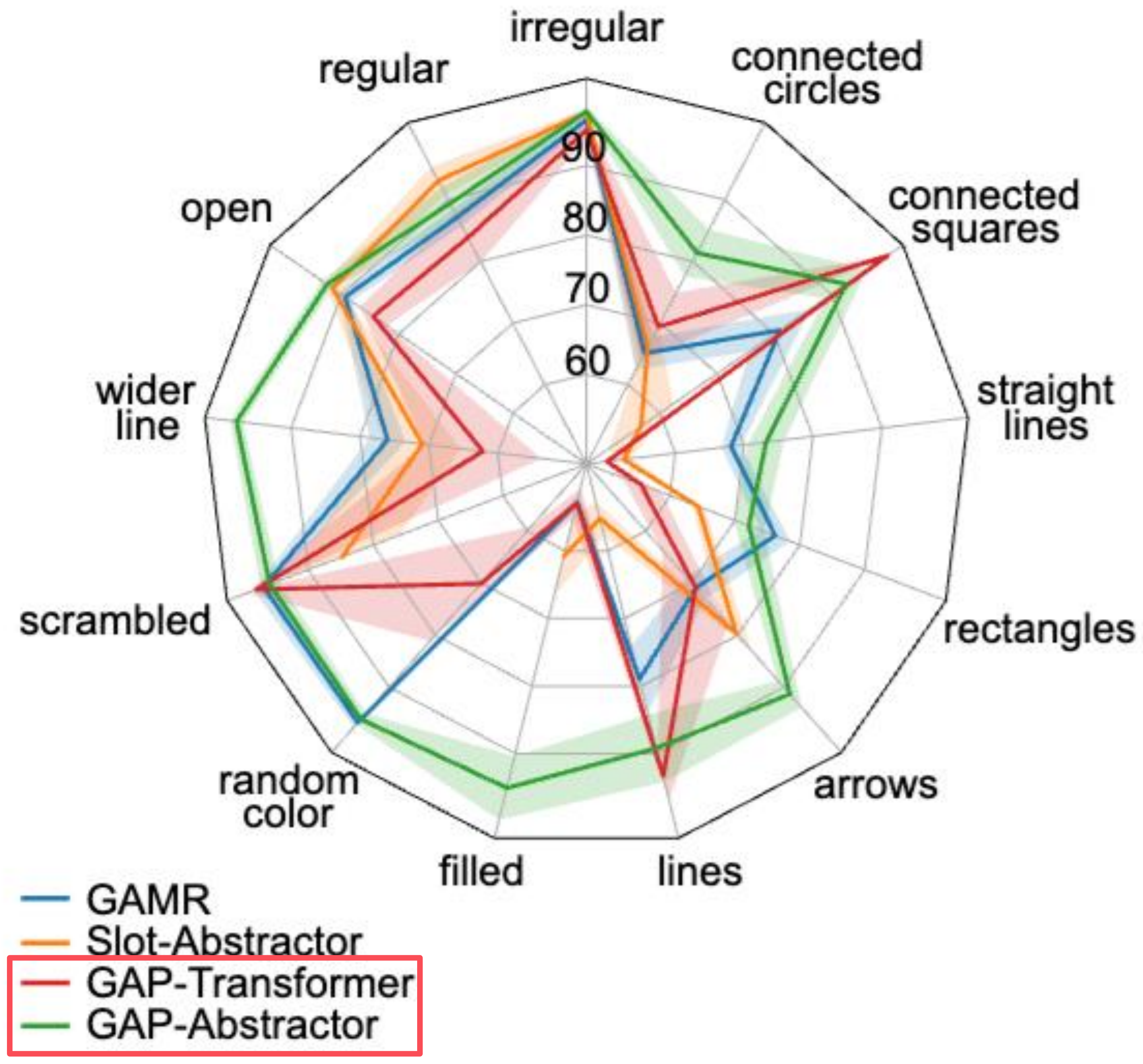
Glimpse-based Active Perception (GAP) produces two sequences
- Glimpse contents
- Glimpse locations

The downstream architecture makes the final decision



Result 1: OOD generalization



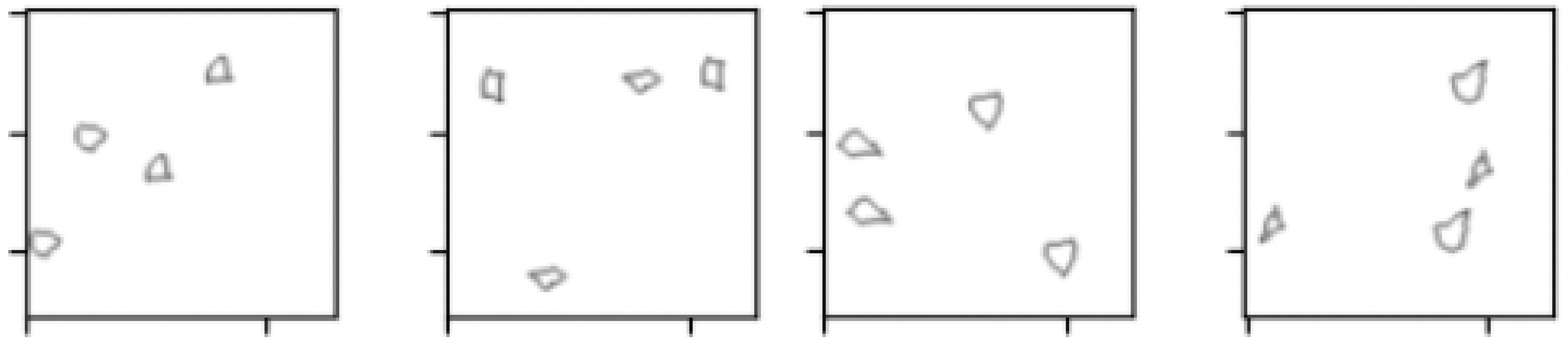


Result 2: sample-efficiency

23 different visual reasoning tasks are used

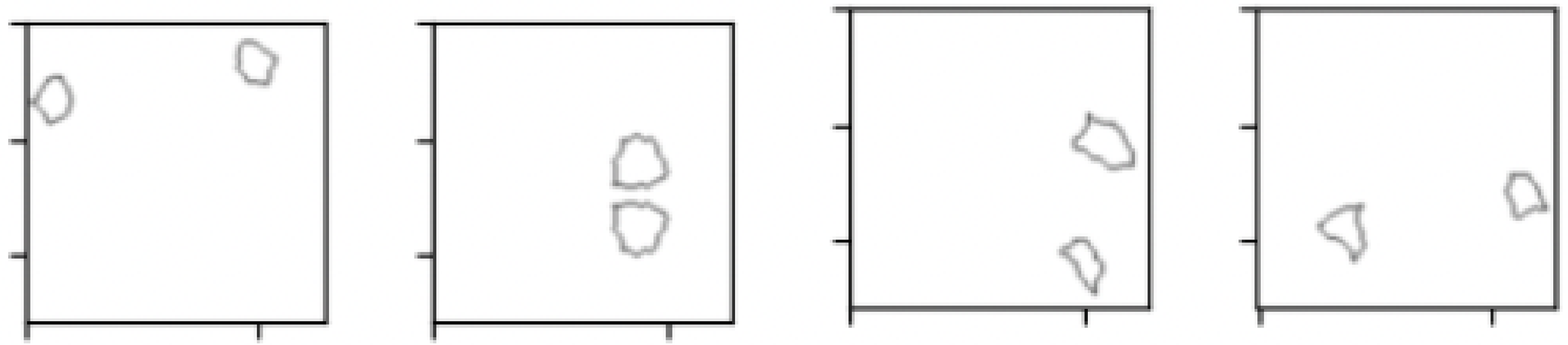
Models are trained with datasets of different sizes

6

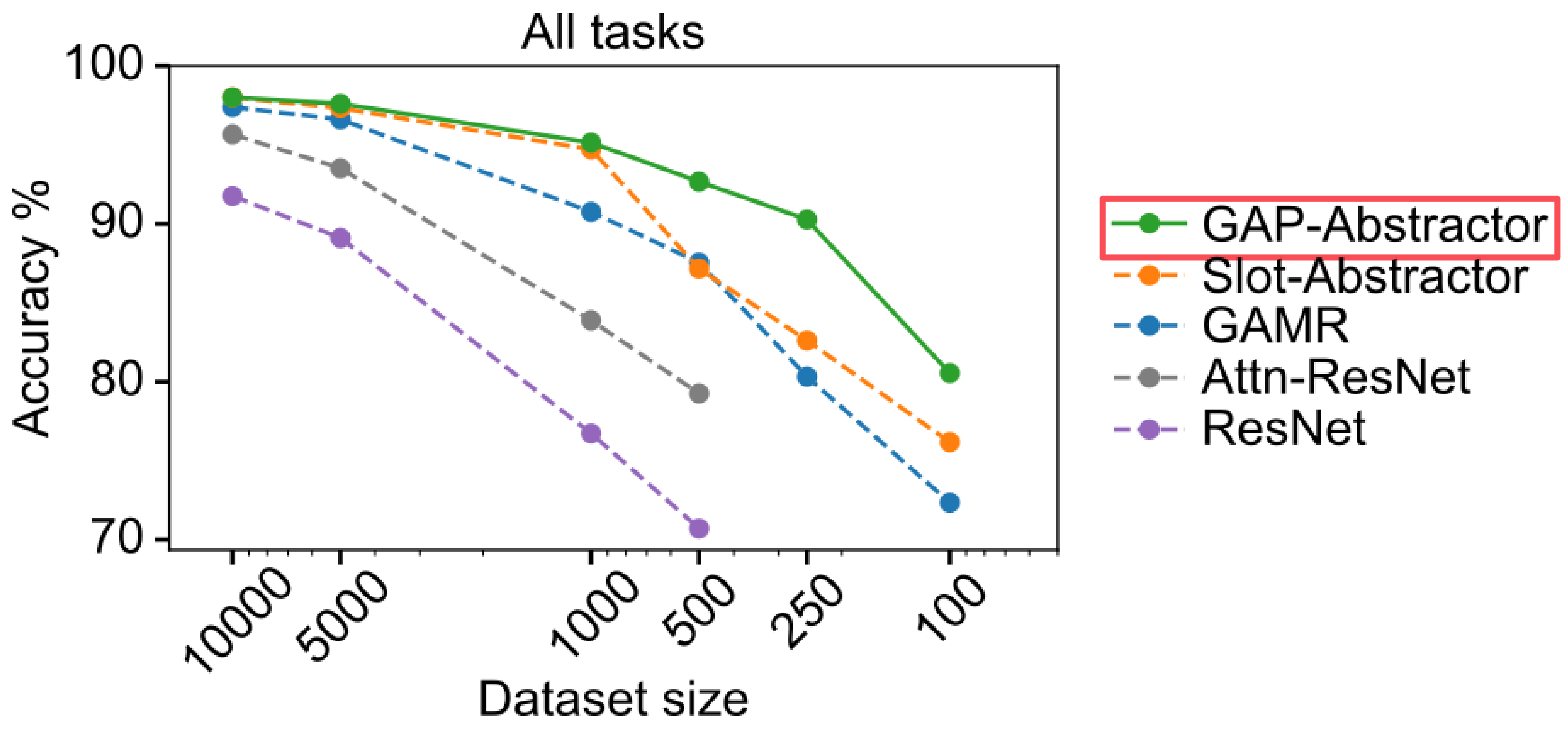


Each image contains two pairs of identical shapes, as with category 1 in problem #5. In category 1, the distance between the two identical shapes is the same for both pairs.

20



Each image contains two shapes. In category 1 one shape can be obtained from the other by reflection around the perpendicular bisector of the line joining their centers.



Result 3: scaling to more complex images

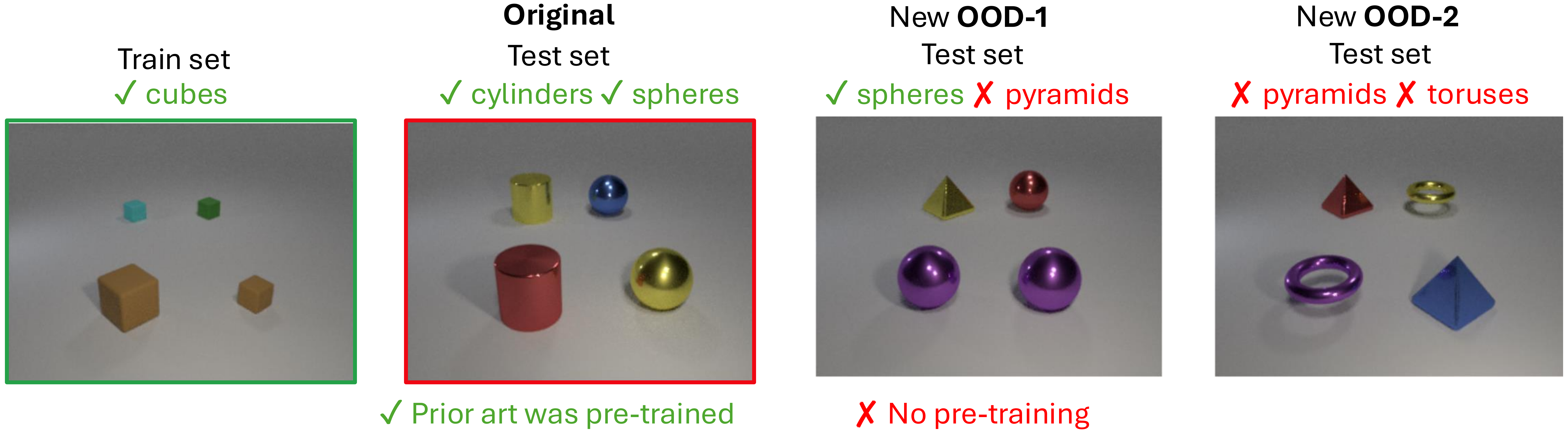
Superior OOD performance on more complex tasks

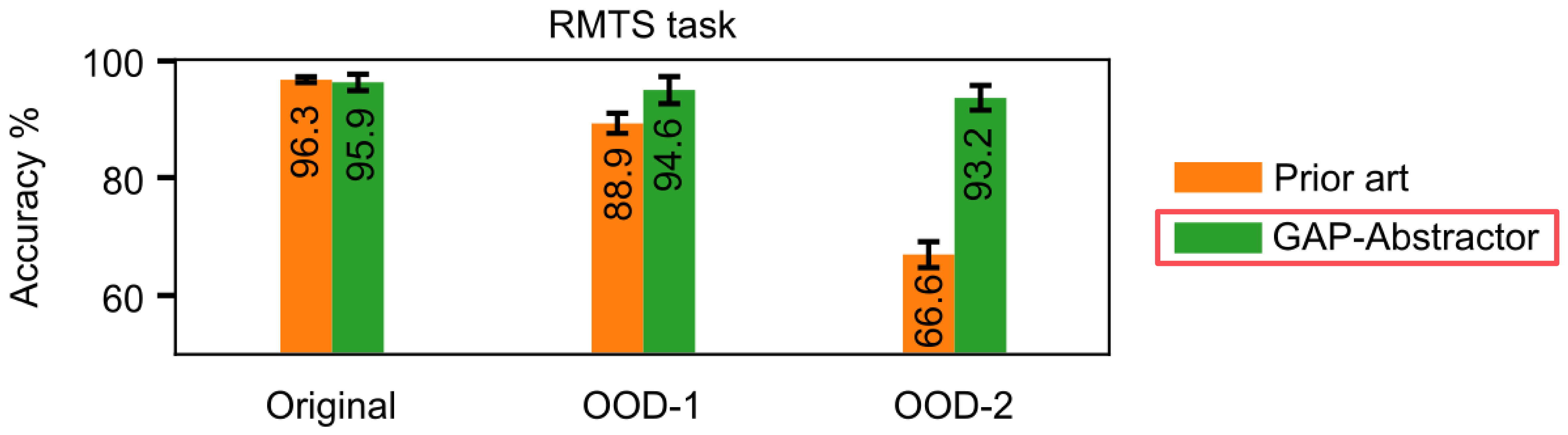
- more abstract relations
- more realistic images

Best performing prior art uses a pre-trained component to pre-segment objects

We define additional datasets, OOD-1 and OOD-2, to test OOD generalization

Our GAP approach achieves superior OOD performance without any pre-training





Thank you for watching!

IBM **Research** Europe