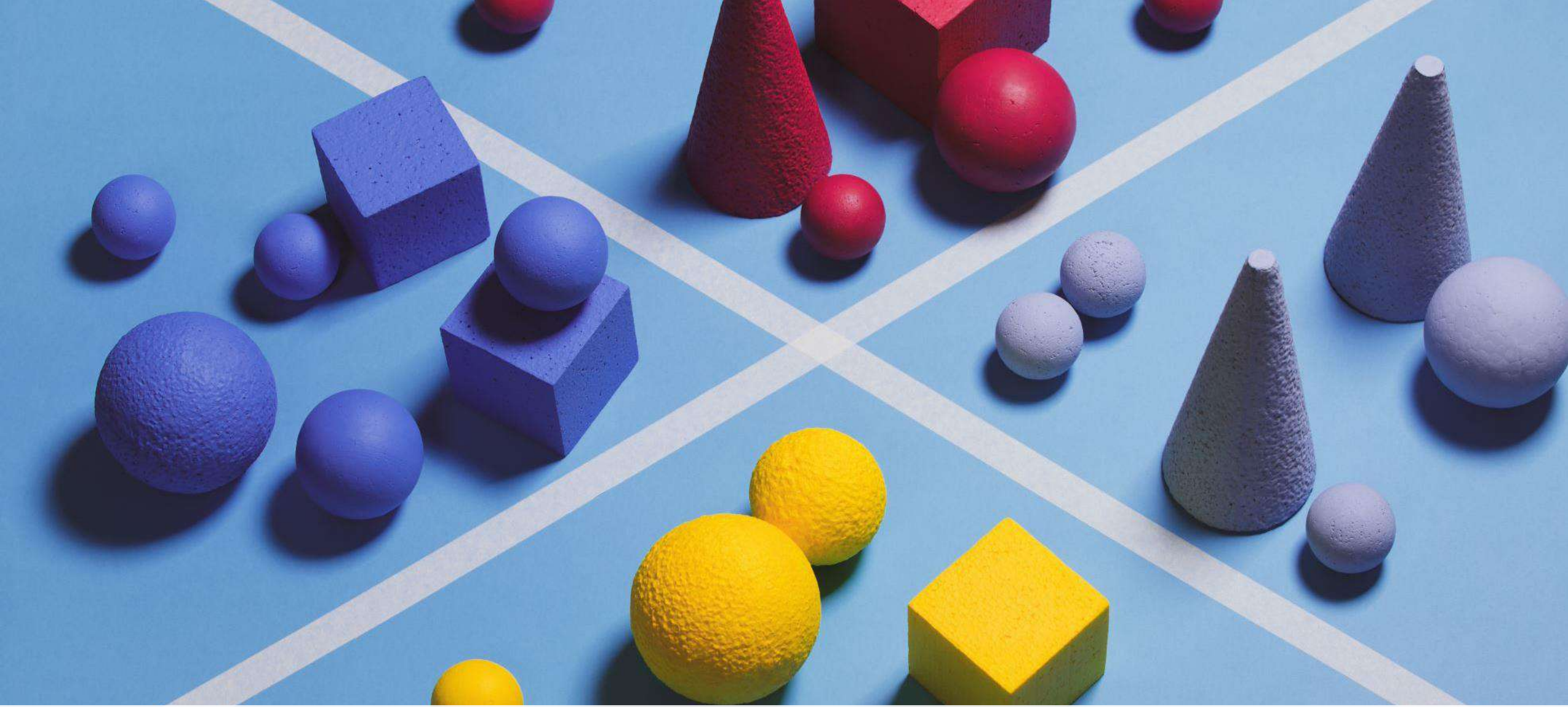
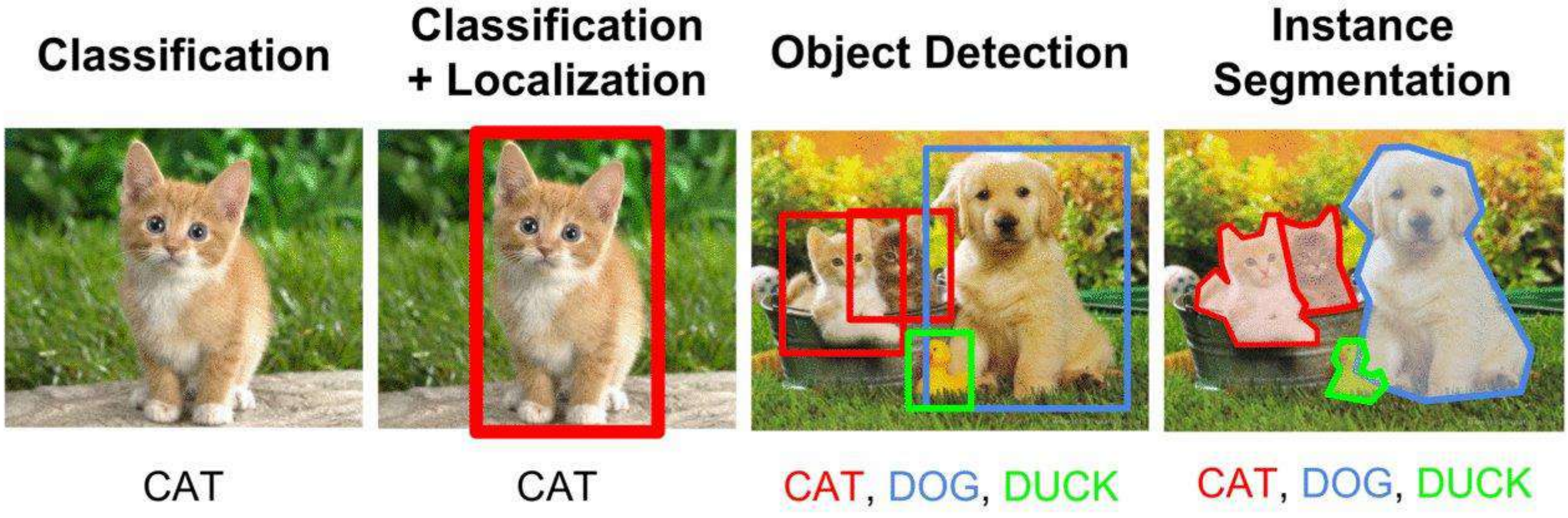


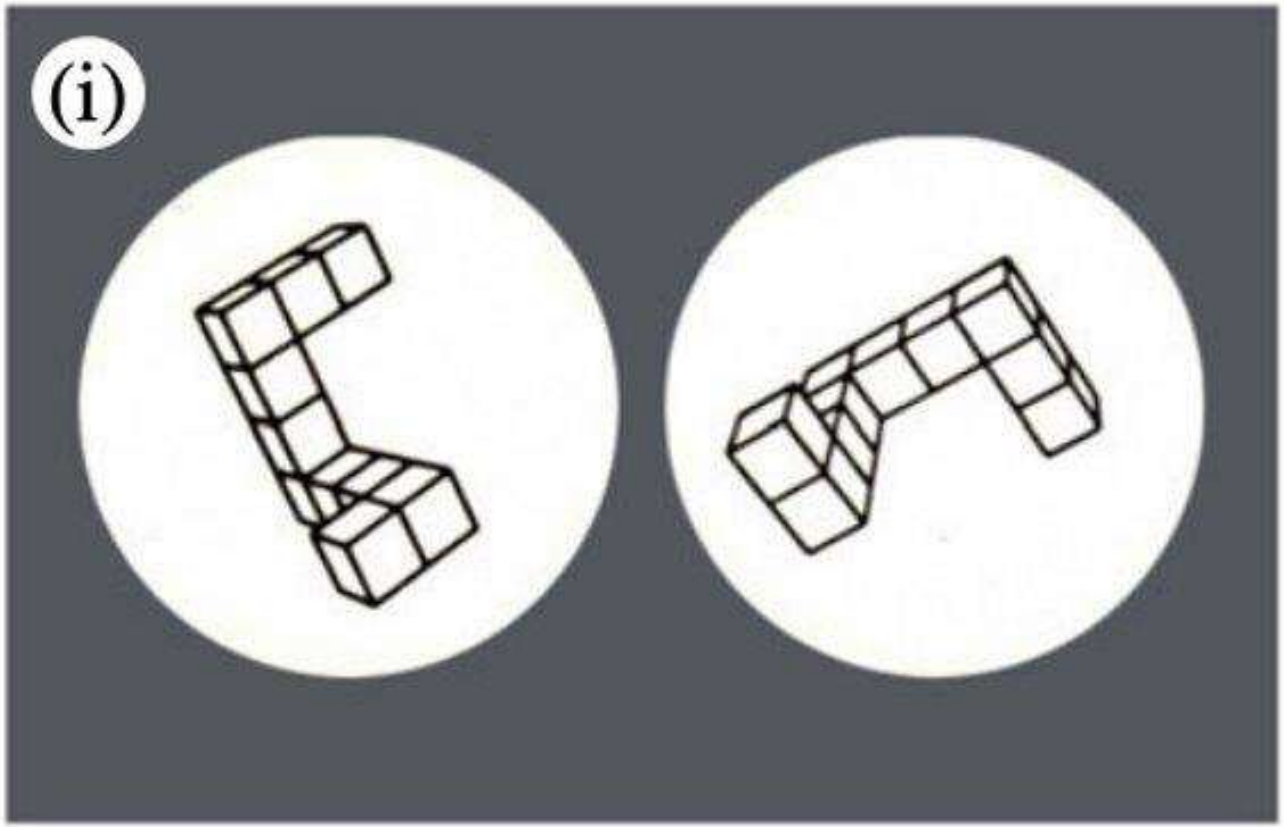
UNDERSTANDING VISUAL REASONING

Mohit Vaishnav

DEEP LEARNING IN COMPUTER VISION



SAME OR DIFFERENT?



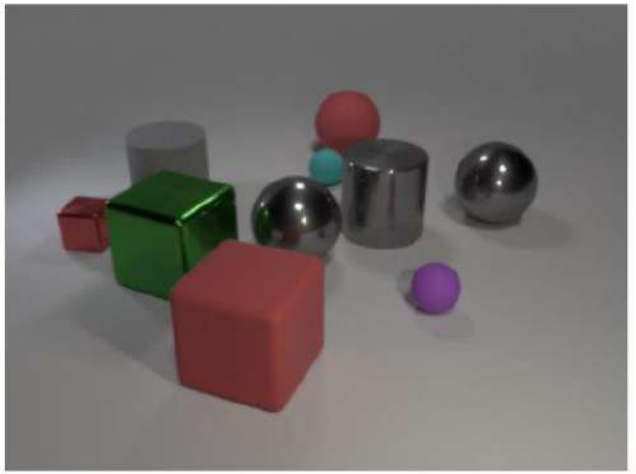
WHAT IS VISUAL REASONING

V+L research is about how to train a smart AI system that can see and talk

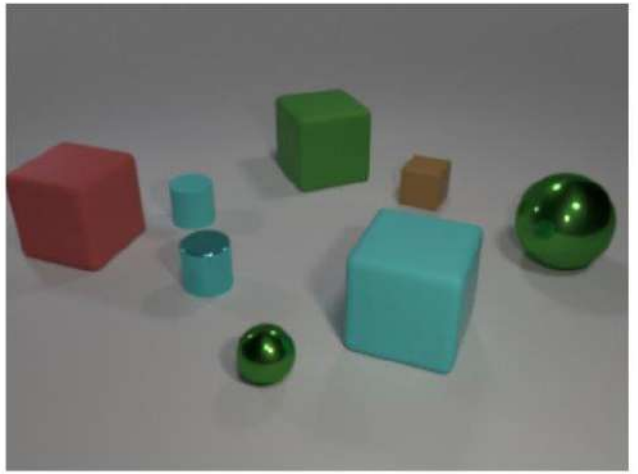
WHY IS VISUAL REASONING CHALLENGING

Compositionality:

Understanding how objects/parts combine to form scenes and meaning



(a) (i) *Remove all gray spheres. How many spheres are there? (3)*, (ii) *Take away 3 cubes. How many objects are there?* (7), (iii) *How many blocks must be removed to get 1 block? (2)*



(b) *Take away all large green metallic spheres. Now remove all cyan objects. How many objects are left? (4)*

WHY IS VISUAL REASONING CHALLENGING

Abstraction:

Reasoning beyond pixel-level details to conceptual understanding.

Example from the V-PROM dataset by Teney et al. (2019). In this example, the images in the Context Panels are related to each other along the rows by their shape. One image is left blank, and the correct image, the heart shape, has to be selected from the Answer Panels. Adapted from the original article.

WHY IS VISUAL REASONING CHALLENGING

Need for World Knowledge /Common Sense:

Images often under specify the scene.

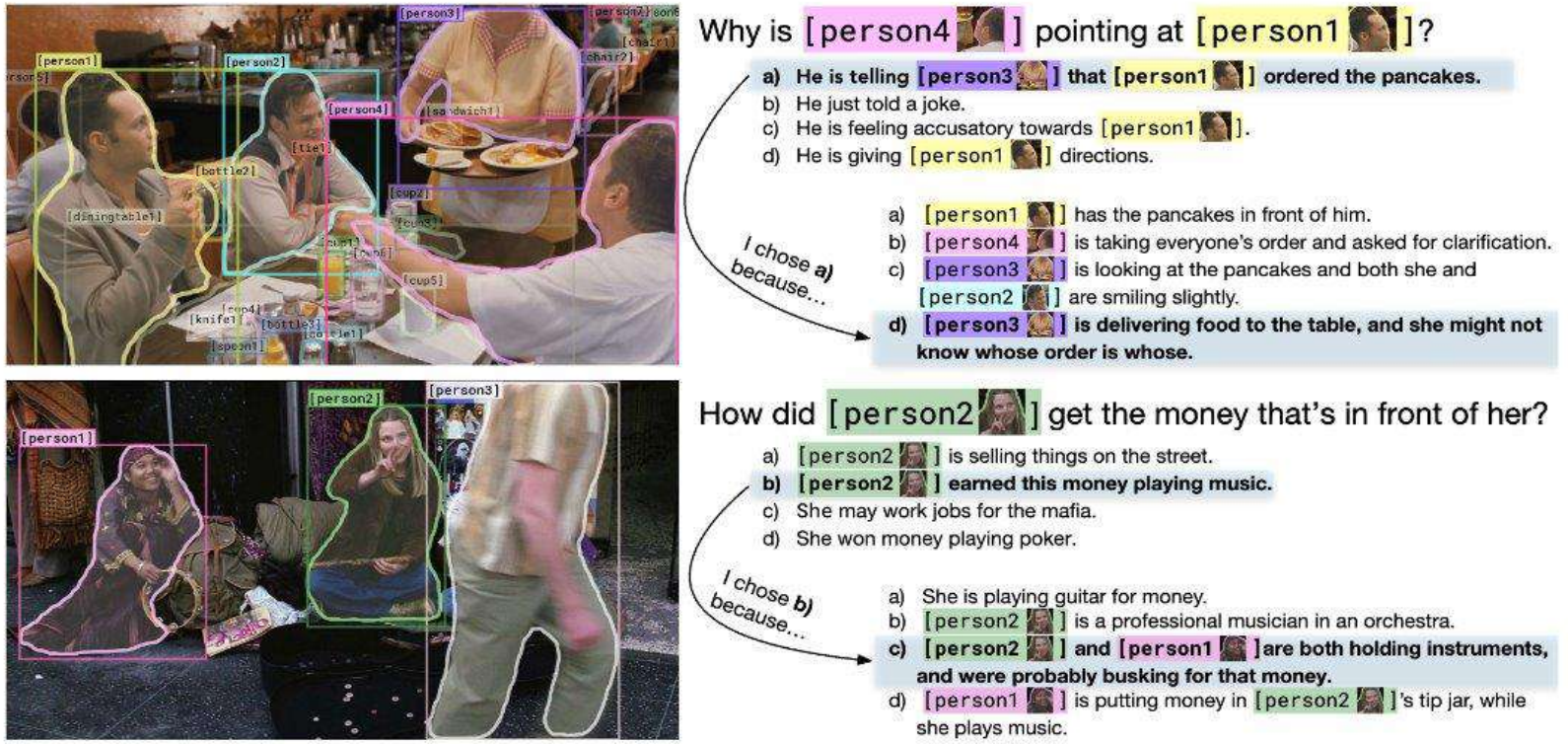
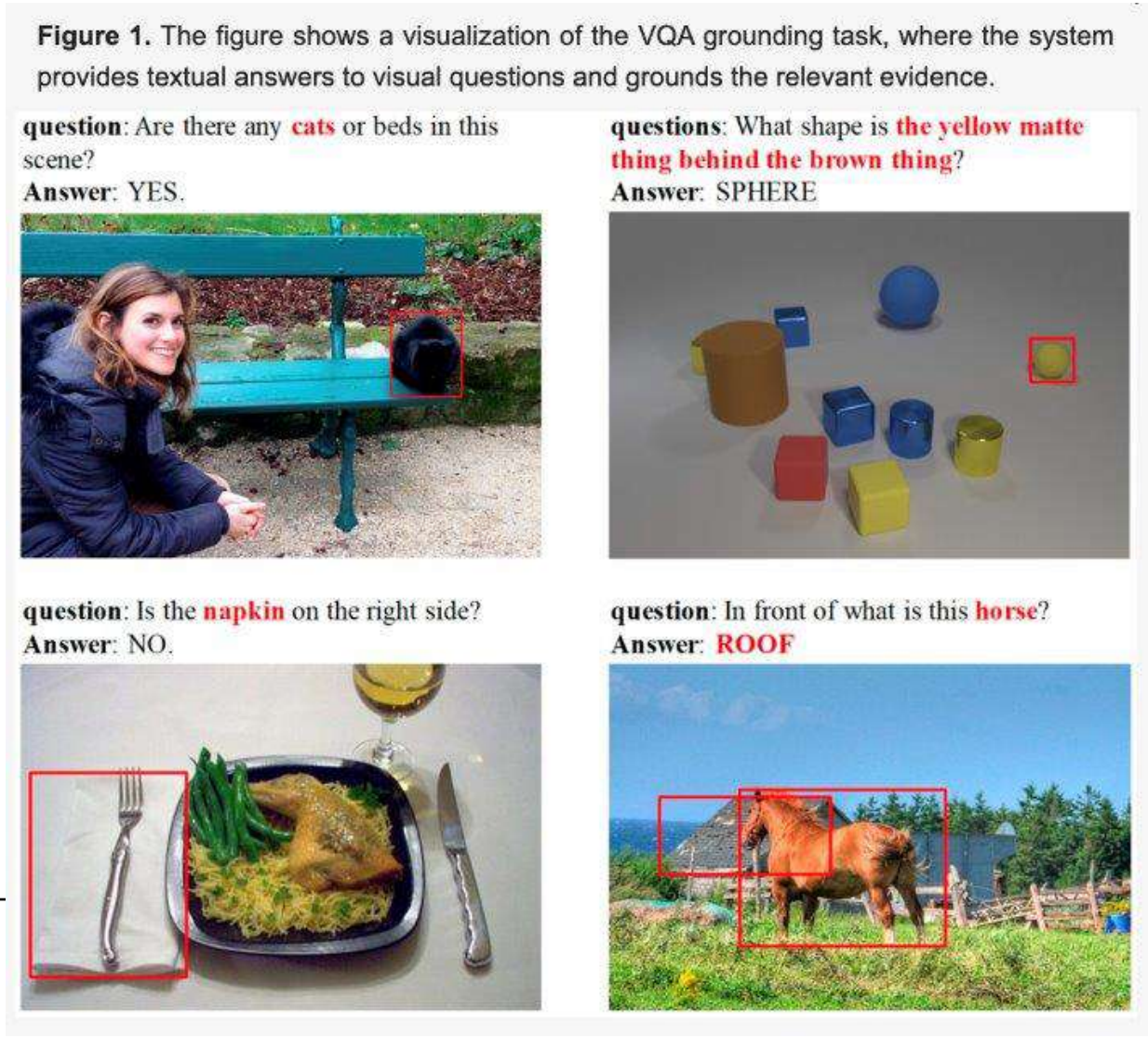


Figure 1: VCR: Given an image, a list of regions, and a question, a model must answer the question and provide a *rationale* explaining why its answer is right. Our questions challenge computer vision systems to go beyond recognition-level understanding, towards a higher-order cognitive and commonsense understanding of the world depicted by the image.

WHY IS VISUAL REASONING CHALLENGING

Integration with Language/Other Modalities:

Many tasks require grounding language in vision.

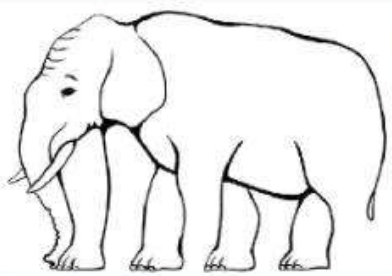


WHY IS VISUAL REASONING CHALLENGING

Ambiguity &

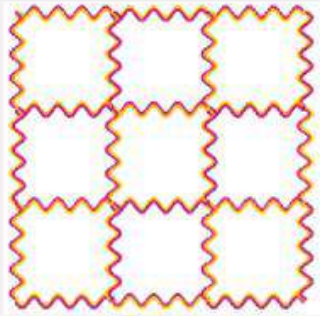
Under specification

in visual data.



Q: What is unusual about this line drawing of an elephant?

A. The elephant has five or six legs
B. The elephant is using its trunk as a fifth leg
C. The elephant is merging with the background in some regions
D. The elephant has six legs while the rest of its body is normal



Q: How many pale yellow regions are in this image?

A. Four, the four corner regions
B. Zero, all the regions are white
C. Five, the center and the four corner regions
D. One, only the center
E. Nine, all the regions are pale yellow

Q: What is unusual about this image?

A. The left cup is a miniature and the right cup is a normal-sized cup
B. The white cube is a highly realistic 3D drawing while everything else is real
C. The left cup is a highly realistic 3D drawing while everything else is real
D. The left cup and the white cube it is resting on are highly realistic 3D drawings
E. The right cup is absurdly large and possibly a highly realistic 3D drawing

CORE TASK IN VISUAL REASONING

- Visual Question Answering (VQA)
- Visual Grounding
- Visual Relationship Detection
- Abstract Visual Reasoning
- Visual Dialog
- Visual Entailment

THE ENIGMA OF VLM VISUAL REASONING

BEYOND "WHAT": UNDERSTANDING *HOW* VLM REASON

VLMs achieve impressive feats (image captioning, VQA, summary etc.).

However, their internal reasoning processes, especially for complex, multi-step visual tasks, remain largely opaque.

Current evaluations often focus on end-task accuracy, offering limited insight into *why* a model succeeds or fails.

This is a fundamental challenge for building more robust, generalizable, and human-like AI.

"How do VLMs bridge the gap between raw pixels and conceptual understanding, particularly for these complex visual reasoning scenarios?"

EXAMPLE BONGARD OW TASK

Negative Set

Positive Set

Query Image

EXAMPLE WINOGROUND TASK

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

MOTIVATION: WHY A NEW EVALUATION APPROACH?

Need for Deeper Diagnostics: Current benchmarks often don't isolate specific failure points (e.g., is it poor perception or flawed reasoning?).

Inspiration from Human Cognition: Humans employ diverse strategies (holistic, deductive, analytical). Can we use these as lenses to evaluate AI?

Goal: To develop a framework that:

- Dissects the perception-reasoning interface.
- Allows systematic probing of different processing stages.
- Provides insights beyond simple accuracy scores.
- Can handle multi-image reasoning evaluation even for single-image architectures.

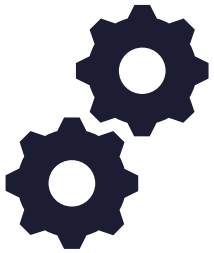
OUR CONTRIBUTION: A COGNITIVELY-INSPIRED EVALUATION FRAMEWORK

- We propose a structured **evaluation framework** grounded in cognitive science.
- This framework employs **three distinct evaluation paradigms**, each mirroring established human problem-solving strategies.
- These paradigms systematically vary the cognitive demands placed on the VLM.

THE THREE COGNITIVE PARADIGMS (OVERVIEW)

Direct Visual Rule Learning

Deductive Rule Learning



Componential Analysis

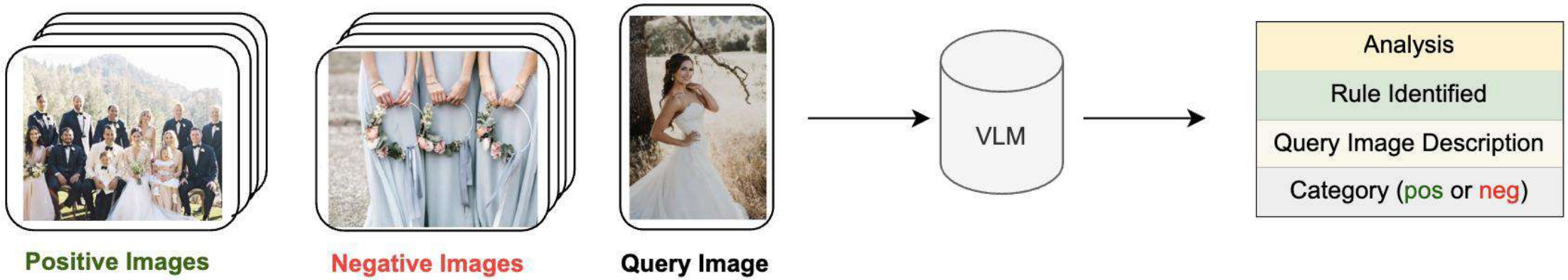
DIRECT VISUAL RULE LEARNING (DVRL)

Mechanism: Holistic, concurrent processing of all images (positive, negative, query).

Cognitive Link: Gist perception, similarity-based judgment.

Constraint: Requires VLM to support multi-image input.

DIRECT VISUAL RULE LEARNING (DVRL)



Mirrors a human's initial intuitive attempt to grasp the core difference between the two sets

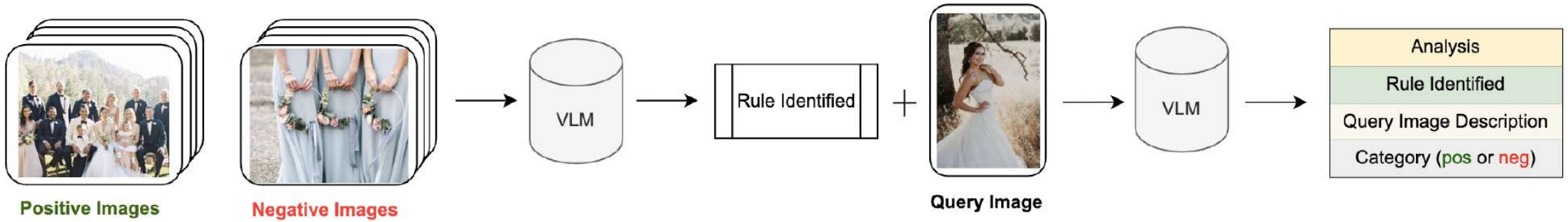
Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

DEDUCTIVE RULE LEARNING (DRL)

Mechanism: Two-stage: (1) Extract rule from examples, (2) Apply rule to query.

Cognitive Link: Explicit, deliberative deduction.

DEDUCTIVE RULE LEARNING (DRL)



Akin to how humans might generate and test hypotheses.

1. Deriving a set of potential rule
2. Generated rules are tested against the image sets

Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

COMPONENTIAL ANALYSIS (CA)

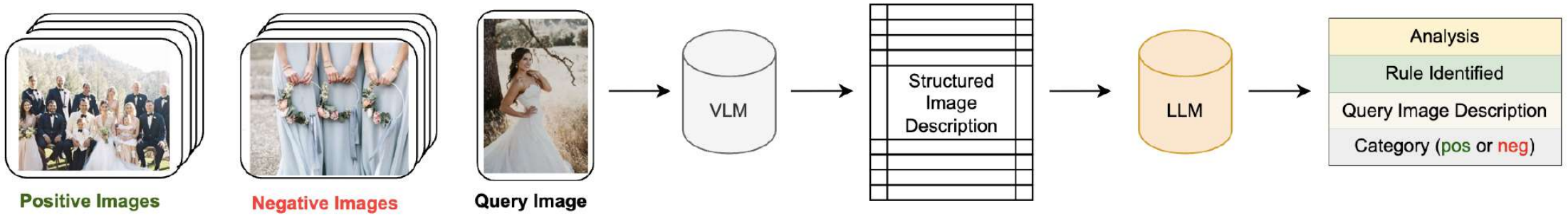
Mechanism: Two-stage:

(1) Generate detailed, **task-agnostic** textual descriptions for each image *individually*.

(2) Reason *solely* over these textual descriptions.

Cognitive Link: Analytical decomposition, reasoning over symbolic representations.

COMPONENTIAL ANALYSIS (CA)



Two-step process - more manageable parts

1. Extract key visual components or features
2. Identify a commonality facilitate understanding and reasoning.

Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

CA: DECOUPLING PERCEPTION AND REASONING

1. **Evaluates Single-Image VLMs on Multi-Image Tasks:** By textualizing images individually, CA bypasses architectural limits for tasks like BPs.
2. **Isolates Reasoning from Perception:**
   - **Stage 1 (Description Generation):** Probes perception.
   - **Stage 2 (Text-Based Reasoning):** Probes reasoning over given representations.

Allows using external, high-fidelity descriptions (e.g., from GPT 4o) to assess "pure" reasoning ability of other models.

CA: DECOUPLING PERCEPTION AND REASONING

3. **Enables LLM Evaluation on Visual Tasks:** Text-only LLMs can perform Stage 2 reasoning.

4. **Task-Agnostic Descriptions:** Stage 1 aims for comprehensive descriptions, *not* descriptions tailored to the specific question/caption (crucial difference from some CoT methods).

DATASETS

Bongard Problems (BPs)

Classic test of abstract rule discovery, multi-image comparison.

1. **Bongard-OpenWorld (BP-OW)**: Natural, real-world images.
   - 500-sample balanced subset
2. **Bongard-HOI (BP-HOI):** Human-Object Interaction rules.
   - 400 samples (100 from each of 4 standard splits, balanced)

Winoground:

Fine-grained visio-linguistic compositional reasoning.

MODELS

VLMs

GPT 4o, Gemini 2.0 flash exp, Pixtral-12B, Llama-Vision (11B, 90B), LLaVA (7B, 13B, 34B), LLaVA-Llama3-8B.

LLMs (for CA Stage 2/Ablations):

Llama3-70B, Qwen2.5 (14B, 32B), Phi-4-14B, etc.

EVALUATION PROTOCOL

- Few-shot prompting (T=0 for deterministic output).
- No model fine-tuning.
- **Metrics**:
  - Accuracy (BPs),
  - Text/Image/Group Scores (Winoground).

PARADIGM TREND

DVRL < DRL < CA

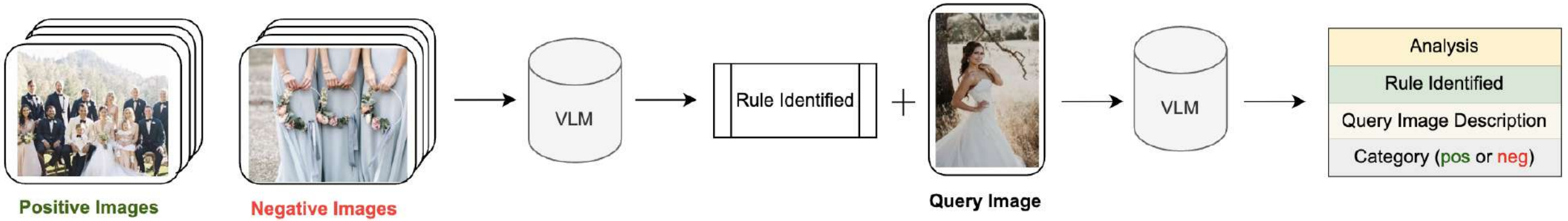
Suggests benefit of structured reasoning and effectiveness of text-based reasoning for this task.

UNCOVERING THE BOTTLENECK

The Critical Role of Perception

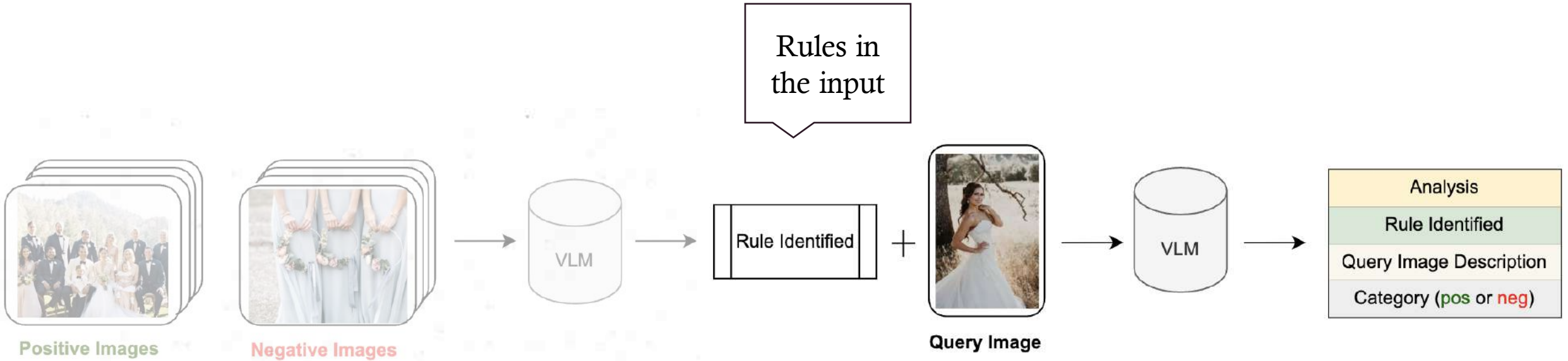
1. RULE BASED EVALUATION

Deductive Rule learning



Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

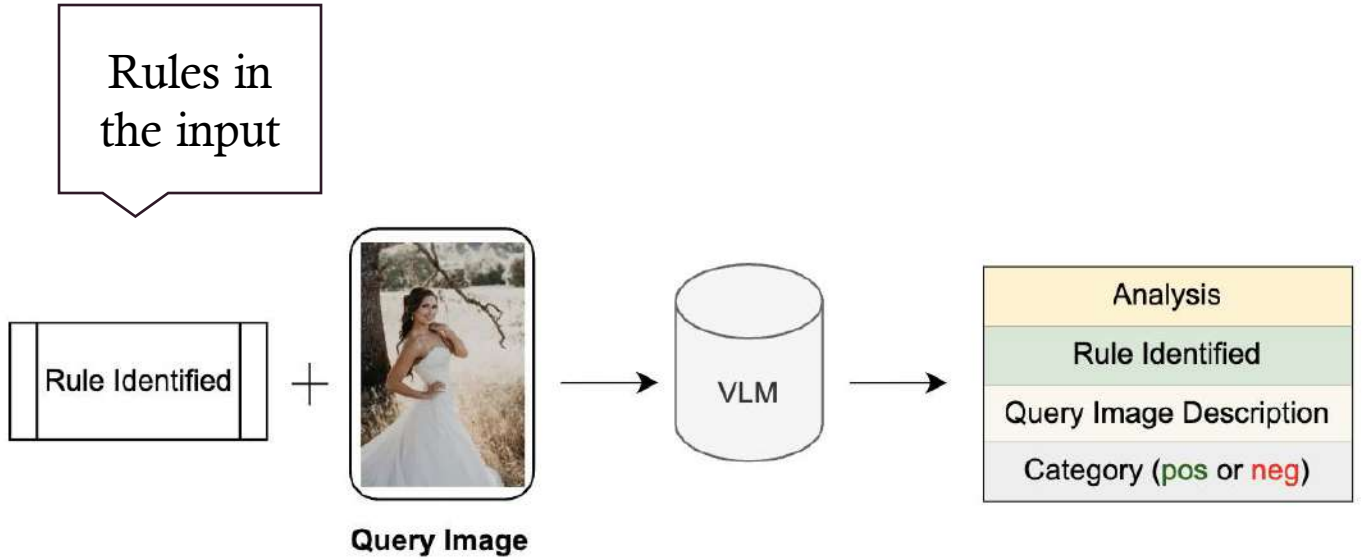
1. RULE BASED EVALUATION



Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

1. RULE BASED EVALUATION

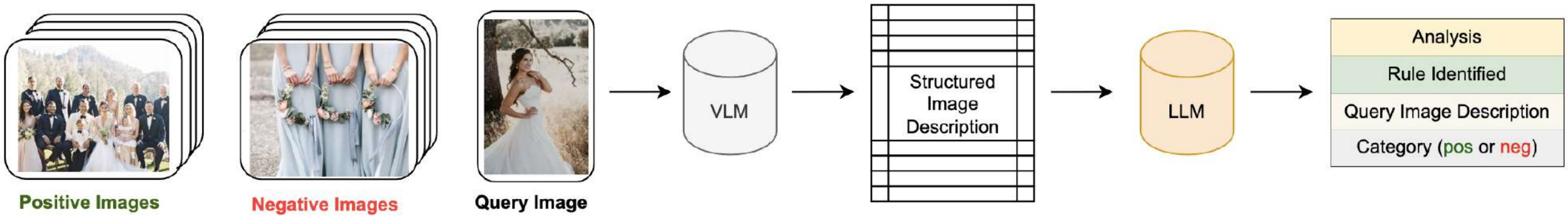
Model	Acc
LLaVA-7B + **DRL**	72.0
Llama-vision-11B + **DRL**	68.2
Pixtral-12B + **DRL**	83.8
LLaVA-13B + **DRL**	70.0
LLaVA-34B + **DRL**	74.8
Llama-vision-90B + **DRL**	74.2



Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

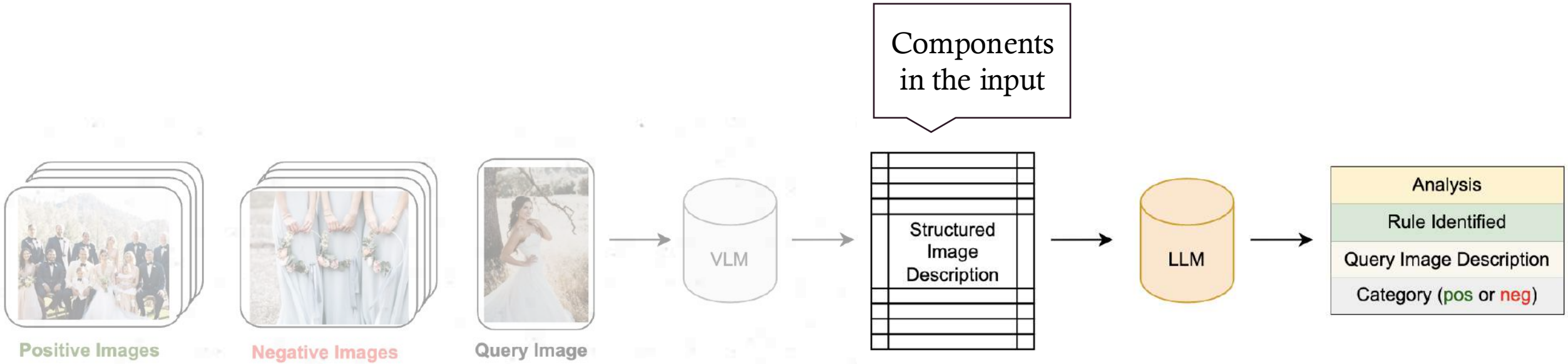
2. EVALUATING DESCRIPTION

Componential Analysis



Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

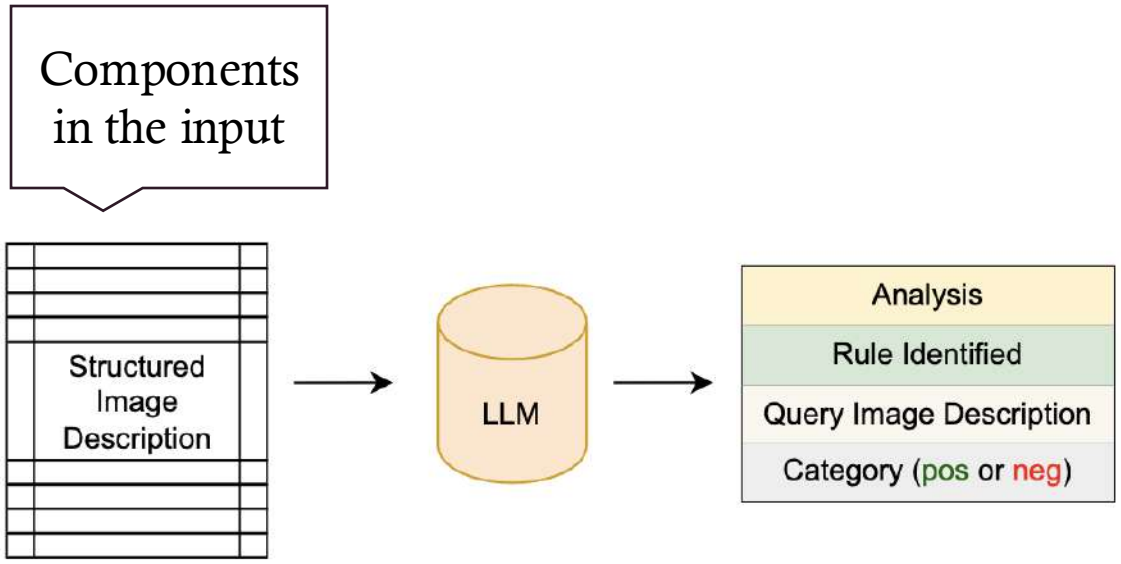
2. EVALUATING DESCRIPTION



Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

3. EVALUATING DESCRIPTION

Model	Acc
Llava:7b + **CA**	80.56
Llava:34b + **CA**	81.56
Llama-vision:11b + **CA**	84.17
Llama-vision:90b + **CA**	90.98
Deepseek-r1:14b + **CA**	87.98
Gemma2:27b + **CA**	88.98
Qwen2.5:7b + **CA**	90.38
Phi4:14b + **CA**	91.98
Qwen2.5:32b + **CA**	92.79
Qwen2.5:14b + **CA**	92.99



Finding:

1. Dramatic improvement for weaker VLMs.
2. Text-only LLMs achieve high performance

Mohit, and Tanel Tammet. "A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs." *arXiv rXiv:2501.13620* (2025).

DISCUSSION

CA's Success: Task-Agnostic Descriptions + Powerful Reasoning

- **Stage 1: Rich, Task-Agnostic Perception:** Generate comprehensive textual "world models" of images, independent of the specific downstream task.
- **Stage 2: Robust Linguistic Reasoning:** Leverage powerful LLMs (VLM backbones or text-only) to reason over these symbolic representations.

DISCUSSION

What Our Framework Reveals About VLM "Cognition

- **Benefit of Structure:** Trend of DVRL < DRL < CA
- **Perception as a Foundation:** Human reasoning relies on robust perception; failures can stem from misperception. Our findings show a similar pattern in VLMs – the perception bottleneck is critical.
- **Language as a Symbolic Interface:** The success of CA and LLMs in Stage 2 highlights the power of language as an intermediate representation for complex visual reasoning.

DISCUSSION

A Framework for Diagnosis and Understanding

- Our framework provides tools to:
  - Pinpoint specific weaknesses (e.g., perception vs. reasoning vs. rule application).
  - Compare different reasoning strategies systematically.
  - Reveal latent reasoning potential masked by perceptual failures.
  - Assess suitability of models for tasks requiring different cognitive demands.

LIMITATION

- **Benchmark Scope & Subsets:**
  - Other reasoning types (physical, causal) needed.
  - BP-OW/HOI results on subsets – full dataset validation for unqualified SOTA.
- **CA Dependencies:** Performance depends on description generator.
- **Prompting:** Used fixed prompts; optimization was out of scope.
- **Cognitive Parallels:** Interpretive, not mechanistic equivalence.

SYMBOLIC GROUNDING REVEALS REPRESENTATIONAL BOTTLENECKS IN ABSTRACT VISUAL REASONING

Mohit Vaishnav

"Is abstract visual reasoning failing because models can't reason — or because they're grounded in the wrong representations?"

THE PROBLEM: VLMS ARE "STRUCTURALLY BLIND"

Modern Vision-Language Models (VLMs) are powerful but fragile

- Models like GPT-4o and Gemini excel at semantic tasks (e.g., "find the picture of a cat").
- However, they fail at tasks requiring **compositional, spatial, or abstract reasoning**. They struggle to count, identify relationships, or understand transformations.
- Their internal representations are often a **"bag-of-objects,"** where they recognize what's in an image but not how things are structured or related.

THE ULTIMATE TEST

Bongard Problems (BPs) test "pure" intelligence.

What is a BP?

- You are given 6 "positive" examples and 6 "negative" examples that follow a hidden, abstract rule.
- The task: Classify a new query image based on that rule.

Crucial advantage

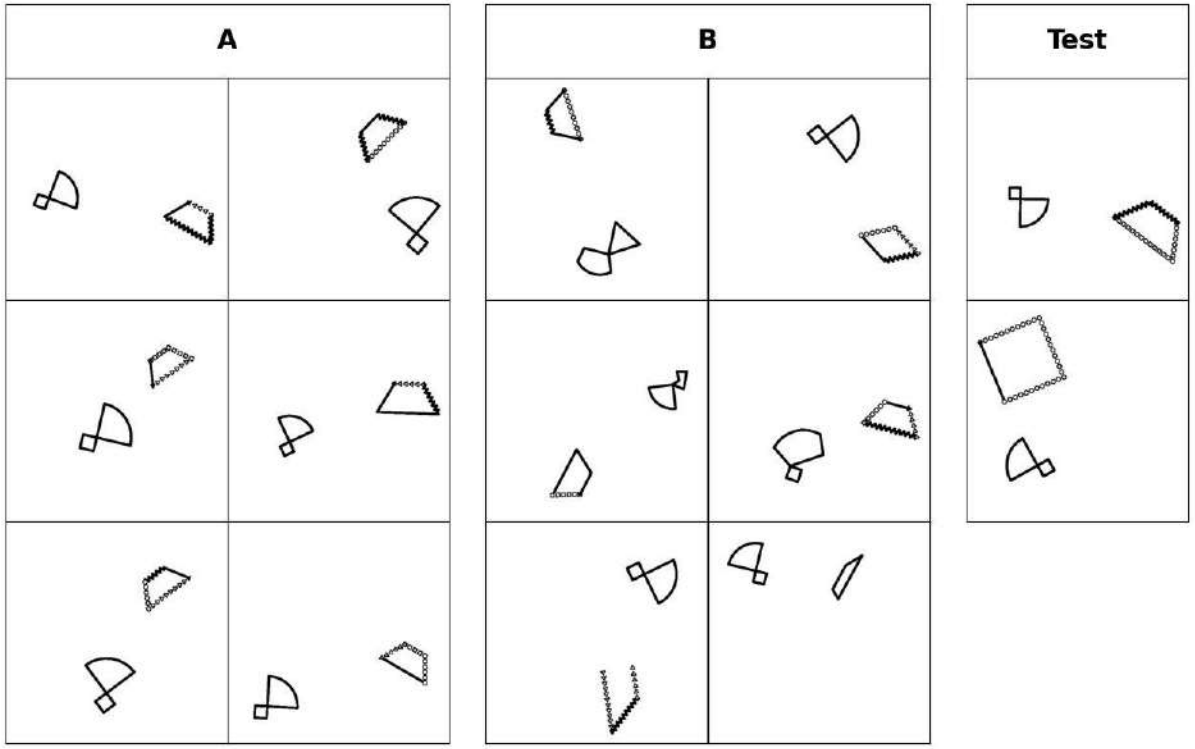
- Each image has a ground-truth LOGO action program

DATASET

Bongard-LOGO

- A large-scale benchmark with over 12,000 programmatically generated BPs.
- Crucially, every image has an underlying **ground-truth generative program**. We use this program directly, ignoring the pixels.
- Three problem types: **Free-form**, **Basic**, and **Abstract** shapes.

EXAMPLE BONGARD LOGO TASK



WHY BONGARD PROBLEMS?

Bongard-style reasoning

- Infer a **latent rule** separating positive vs negative examples
- Emphasizes: geometry, relations, t opology, compositional structure, many more …

Why Bongard-LOGO

- Synthetic, controlled
- Ground-truth **generative programs available**
- Ideal for isolating representation vs reasoning

THE ULTIMATE TEST

Why are they hard for AI?

- The rule is **abstract and conceptual** (e.g., "convexity," "symmetry," "containment").
- It's not about recognizing objects but about inducing a principle.
- State-of-the-art models perform **abysmally** on these tasks.

let's us swap pixels for symbolic structure

COMPONENTIAL–GRAMMATICAL (C–G) PARADIGM

Key move

Decouple perception from reasoning

Two-stage pipeline

Symbolic grounding

- Use LOGO programs or structured descriptions

Textual reasoning

- LLM induces rule + classifies query
- Important framing

Symbolic input is an oracle probe, not a deployed system

COMPONENTIAL–GRAMMATICAL (C–G) PARADIGM

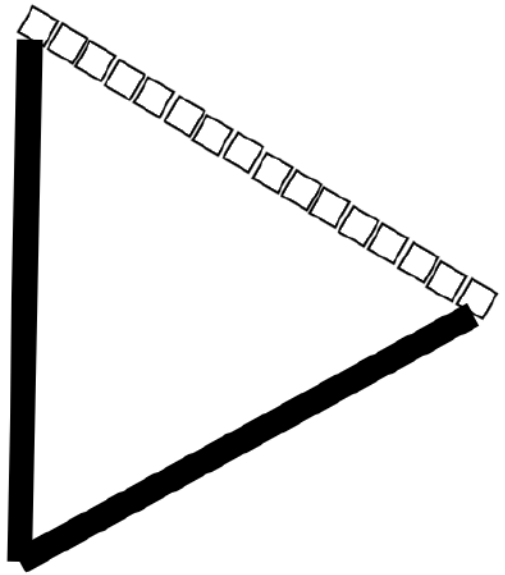
Example:

- **Instead of:** "a pointy shape with three equal sides"
- **We use:** `equilateral_triangle line_1.0, line_1.0, line_1.0 L0--L120.0--L120.0`
- This representation is **unambiguous, structured, and perfect**.
- **The Hypothesis:** An LLM given this perfect, symbolic input can easily solve the reasoning task.

THREE INTERFACES COMPARED

Interface	Input
Visual baseline	Raw images (VLM)
C–G (AP)	LOGO-style action programs
C–G (AD)	Natural-language action descriptions

Same task. Same reasoning engine.
Only representation changes.

Image	Action Program (AP)	Action Desription (AD)
	"line_normal_1.000-0.500", "line_square_1.000-0.833", "line_normal_1.000-0.833"	To draw figure 0, follow these steps: Step 1: draw a normal line of 1.000 units. Then, continue straight without turning. Step 2: draw a normal line of 1.000 units. After that, turn left by 119.9 degrees. Step 3: draw a normal line of 1.000 units. After that, turn left by 119.9 degrees. The figure is now complete.

Concept Closed shape

RESULT

Condition	Accuracy
Visual VLM	~50% (chance)
C–G (AP)	60–80%
Best model (Phi-4-Reasoning, FF)	**96.2%**

Models *can* reason — when grounded in explicit structure.

WHY THIS MATTERS

The same LLM:

- Fails on pixels
- Succeeds on symbolic programs

➞ *Strong evidence for a representational bottleneck*

Not primarily a reasoning failure.

AP VS AD: DOES FORMAT MATTER?

Findings

- Action Descriptions (AD): Better on Free-form and Basic
- Action Programs (AP): More stable on Human-designed tasks

Interpretation

- Linguistic bias helps simpler concepts
- Procedural structure helps complex abstraction

CONCEPT CONDITIONING: HELPFUL OR NOT?

Adding the target concept label:

- Small, model-dependent gains
- No universal improvement

Takeaway: Structure > supervision

ARE MODELS REALLY USING STRUCTURE?

Randomization controls

1. Category shuffle

Swap pos/neg support examples --> Accuracy drops ~7–9 pts

2. Sequence shuffle

Randomize stroke order in query --> FF drops ~20 pts

➞ Models rely on compositional structure, not token counts

OPEN CHALLENGES AND FUTURE DIRECTION

- **Robustness & Generalization:** Models that perform well on specific benchmarks often fail in others
- **Perception Bottleneck:** Improving fundamental visual understanding remains critical (as highlighted by my evaluation work).
- **Compositional Generalization:** Still largely unsolved; models struggle with novel combinations (CVR helps measure this).
- **Out-of-Distribution Performance:** Handling variations not seen during training.
- **Knowledge Integration:** Effectively incorporating common sense and structured world knowledge.

OPEN CHALLENGES AND FUTURE DIRECTION

- **Explainability & Interpretability:** Understanding why a model made a decision, especially complex ones.
- **Data Efficiency:** Reducing reliance on massive labeled datasets; exploring self-supervised, few-shot learning for VR.
- **Causal Reasoning:** Moving beyond correlations to understanding cause-and- effect in visual scenes.
- **Temporal Reasoning:** Extending VR concepts to video and dynamic events.
- **Future Directions:** Cognitively-inspired architectures, improved evaluation metrics, embodied VR (robotics), lifelong learning systems for VR.

THANK YOU