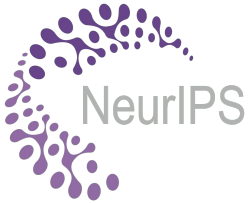




# Visual Structures Help Visual Reasoning: Addressing the Binding Problem in LVLMs

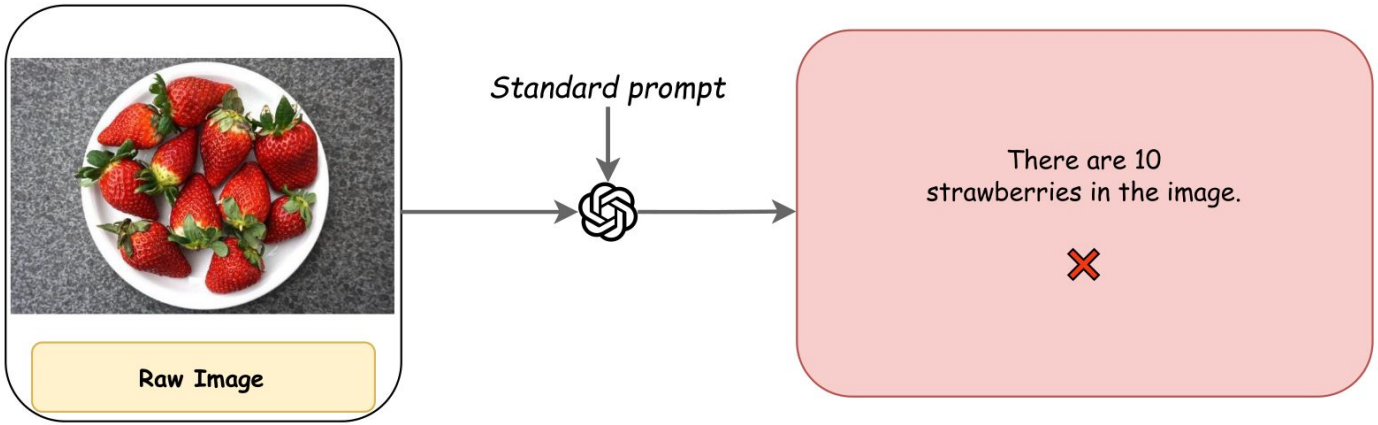
Amirmohammad Izadi, Mohammad Ali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah

Department of Computer Engineering
Sharif University of Technology

# Motivation

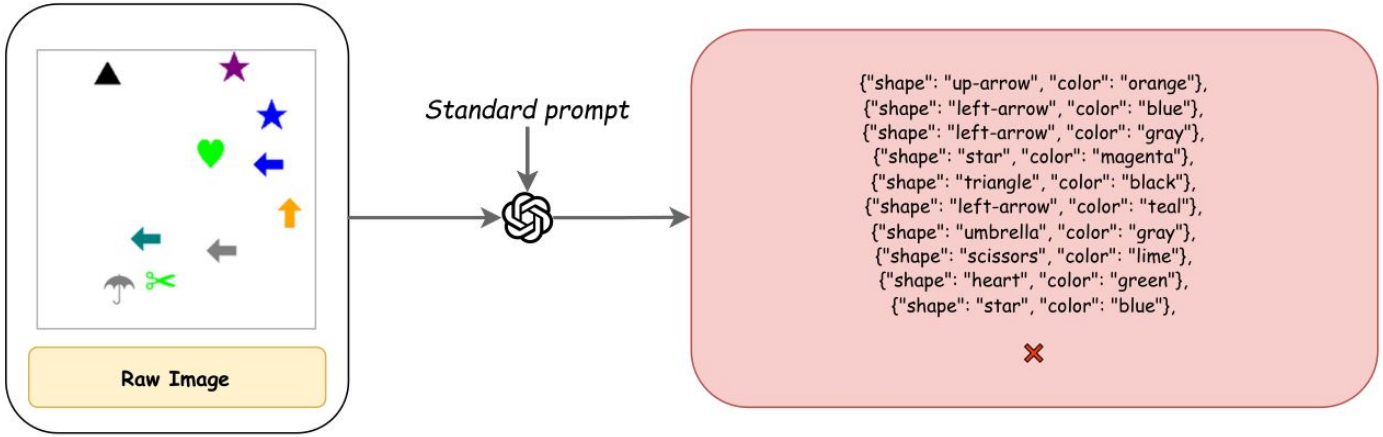
- LVLMs still fail at **counting**, **visual search**, **spatial reasoning**, and **scene description**.
- Root cause: **The Binding Problem**
  - Models detect features but fail to bind them to the correct objects.
  - Errors increase in cluttered or multi-object scenes.

*"We need a way to enforce **structured, sequential visual processing**."*



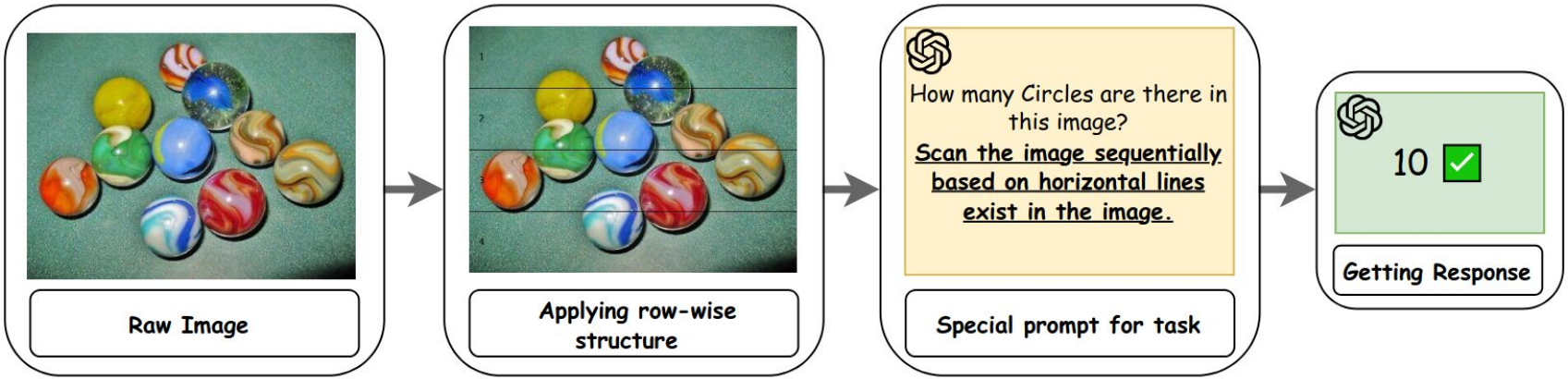
# What Is the Binding Problem?

- **From cognitive science:**
  - difficulty in linking **shape**, **color**, **location**, etc.
- **Leads to:**
  - Object confusion
  - Attribute mixing
- LVLMs process images **in parallel**, causing feature interference.



# VISER: Our Proposed Method

1. **Visual scaffolding:**
   - Add **3 horizontal lines** + optional row numbers.
   - Splits the image into structured regions.
2. **Sequential scanning prompt:**
   - *"Scan the image sequentially based on the horizontal lines in the image."*
   - Encourages **row-wise reasoning**, similar to visual Chain-of-Thought.



# Tasks Evaluated

We test VISER on 4 core visual reasoning **tasks**:

- Visual Search (target present/absent)
- Counting
- Scene Description (edit distance)
- Spatial Relationships (left/right/above/below)

**Datasets:**

- Synthetic 2D & 3D
- Real natural images (counting, spatial)

**Models:**

- GPT-4o
- Claude 3.5
- Qwen2.5-VL
- LLaMA-4
- Mulberry
- OpenVLThinker

# Key Results

**Counting:**

- GPT-4o (2D): 12% → 38.8%
- Qwen2.5-VL (2D): 5.8% → 40.8%
- Big gains also in 3D scenes.

**Visual Search (Harmonic Mean):**

- GPT-4o (2D): 0.48 → 0.73
- Claude 3.5 (2D): 0.34 → 0.66

**Spatial Relationships:**

- GPT-4o (Natural): 69.4% → 77.4%
- Gains across all models.

**Scene Description:**

- Edit distance improves in all settings
- Hardest scenes benefit the most
  - e.g., 2D: 1.94 → 1.62

# Beyond Core Tasks

**VISER vs Fine-Tuned Models:**

- Matches or exceeds Mulberry and OpenVLThinker on many tasks

**VISER vs. Chain-of-Thought**

- **CoT often hurts performance** on visual tasks.

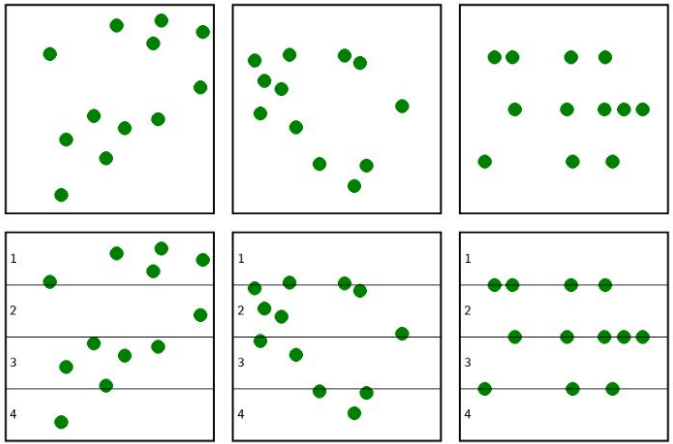
**Broader Benchmarks:**

- Improves performance on **MMBench**, **PhysBench**, **RAVEN**, and **Visual Analogy**
- Shows strong generalization beyond our main tasks

# Limitations & Next Steps

**Limitations:**

- Static lines may interfere with some images.
- Gains smaller on some natural image datasets.
- Not adaptive to content layout.




**Future Work:**

- Adaptive or learned scaffolding
- Multi-scaffold ensembles
- Architectural support for serial visual attention
- Applications in hallucination reduction

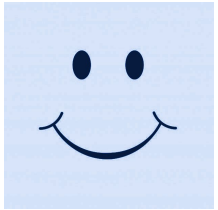
# Takeaway

**VISER:**

- Adds minimal visual structure
- Strong improvements in binding-heavy tasks
- Outperforms CoT and competes with fine-tuned models
- Zero compute, zero training

***"If we want LVLMs to reason visually, we must structure the visual input—not only the text."***

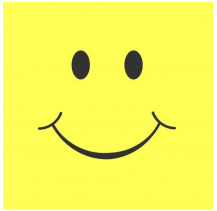
1

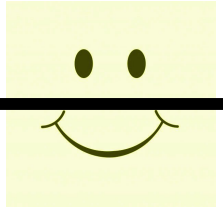


2

*Thank you!*

3





4