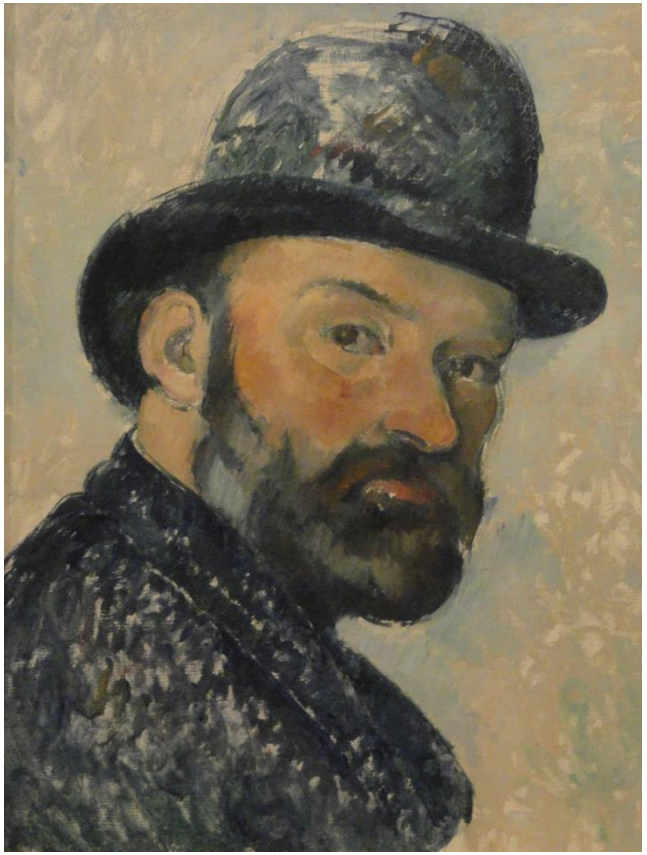
University of Nebraska Lincoln

ICML International Conference On Machine Learning

Seoul, South Korea, July 6-11, 2026

What Cézanne Knew About Visual Intelligence That Vision-Language Models Miss



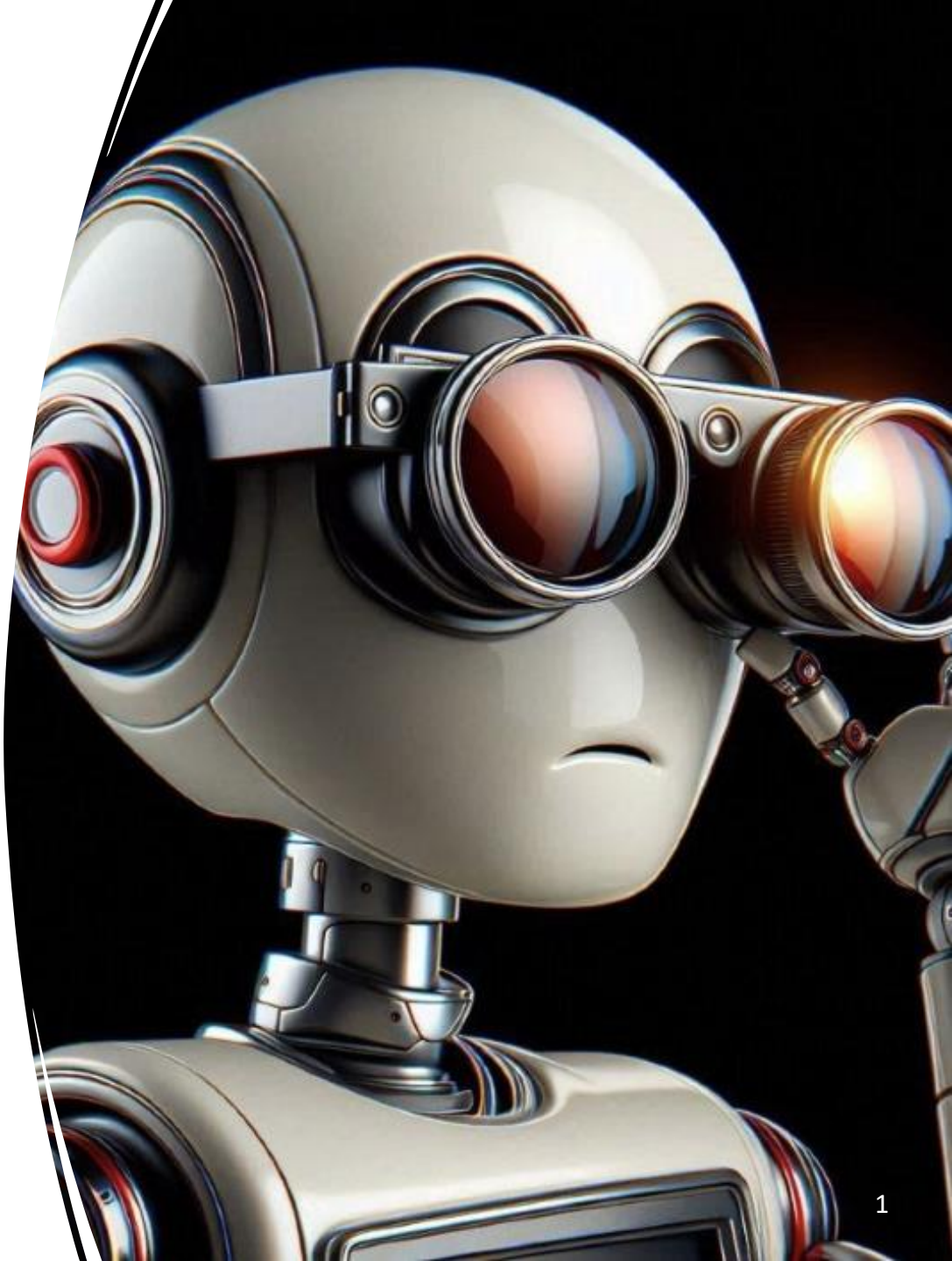
M. R. Hasan and Chinh Hoang

Human-First **A**rtificial Intelligence **L**ab (HAL 2.0)

June 1, 2026

- The foundation of visual intelligence

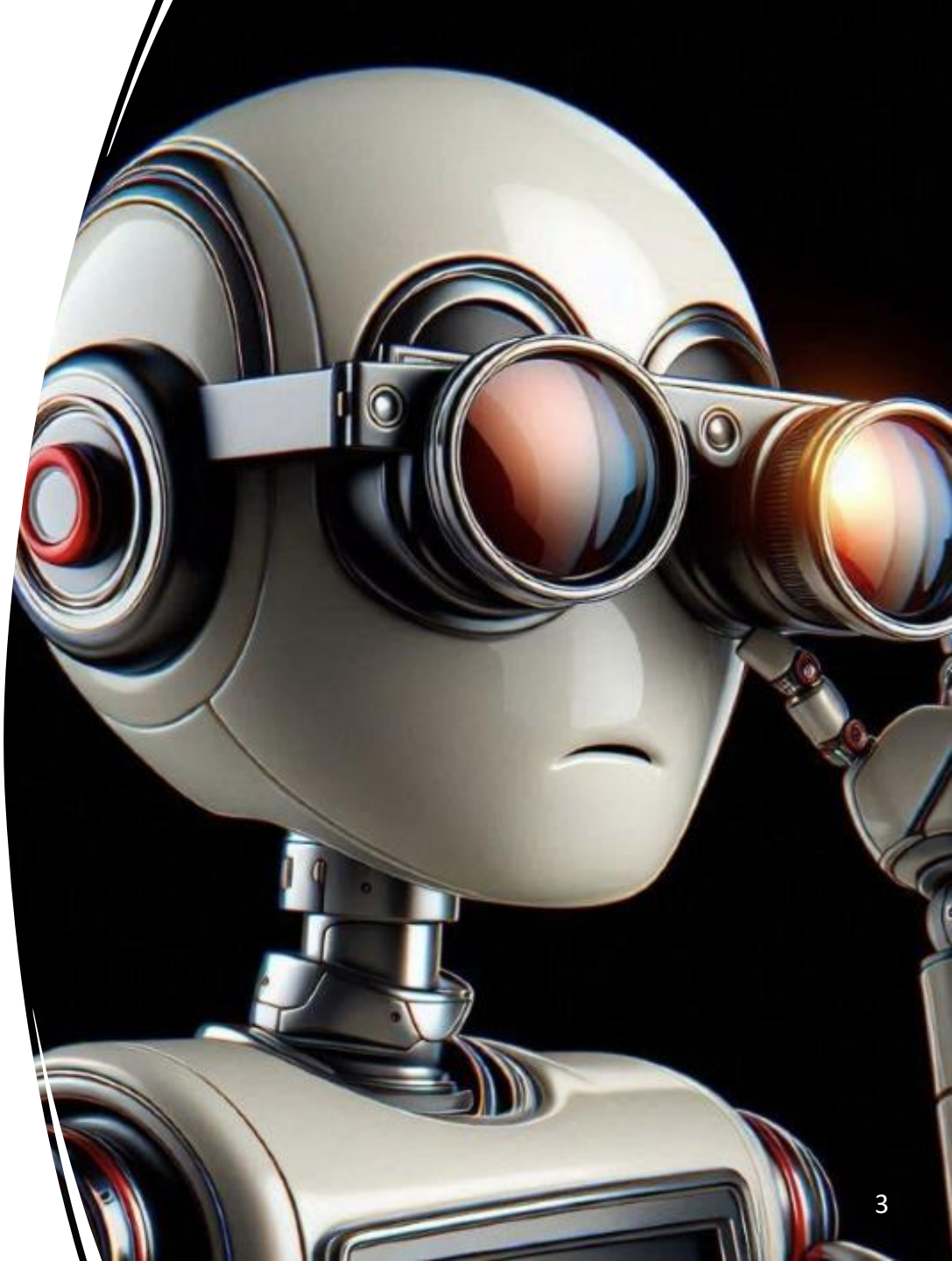
- Do the current vision-language model (VLM) benchmarks test the foundation



Paul Cézanne understood something about visual intelligence that our best multimodal AI systems still miss

The Question

- Multimodal AI systems such as VLMs now help doctors read medical scans and guide self-driving cars.
- They are required to accurately answer questions about what caused what.
- They often give correct answers.
- But are they reasoning, or pattern-matching?



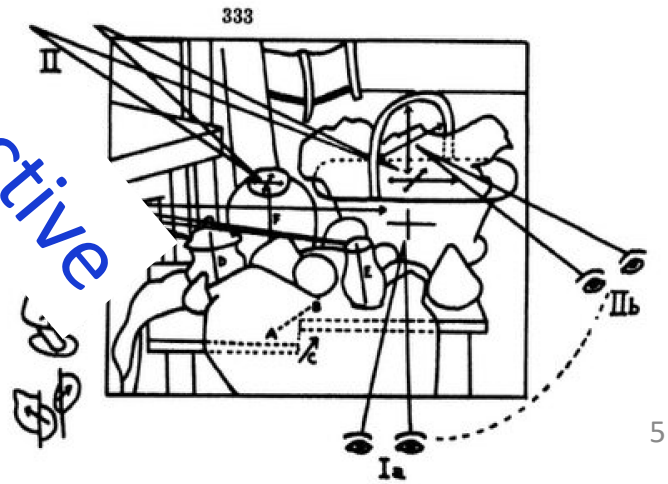
- How do VLMs capture vision?
- VLMs process vision like cameras.
- A camera records one instant from one position.

- Humans see differently.
- We build our perception by looking at objects from multiple positions over time.
- We shift, lean in, refocus on different objects.
- The brain weaves these moments into coherent perception.

The Insight

- Cézanne grasped human visual intelligence with remarkable insight.
- In his still life paintings, objects are rendered from multiple viewpoints.
- His representations encode time through which perception forms.
- Time is foundational!
- Cause → Time → Effect

Constructive



Cézanne painted how we see, not what is seen from a frozen position.

The Critique

Condition 1 Prompt

System: You are a vision-language model (VLM) tasked with answering a provided question about an image in a way that demonstrates causal reasoning, based on Judea Pearl's framework. You are encouraged to construct a lightweight causal chain to represent the chain of cause-and-effect relationships relevant to the answer, using the number of nodes and arrows needed to accurately capture the causal logic (e.g., 2 nodes like A → B, 3 nodes like A → B → C, or more if appropriate), tailored to the complexity of the reasoning. However, it is reasonable to provide only a text response without a causal chain if you find it sufficient to explain the causal reasoning. If you include a chain, provide it first, followed by a concise text response that explains the reasoning based on the chain. If you omit the chain, provide only the text response, ensuring it still reflects causal reasoning.

Instructions:

- **Base the response on the image and question:** Ensure the answer (and chain, if provided) is grounded in the image's content and relevant to the question.
- **Optional causal chain:** If you choose to include a chain, follow these steps:
  1. Identify the key cause-and-effect relationships implied by the question (e.g., 'Rain causes wet ground, which darkens the scene').
  2. Format these relationships as a chain using '→' to connect nodes (e.g., Rain → Wet Ground → Darkened Scene).
  3. Use the fewest nodes needed for clarity, but include additional nodes if the causal chain requires multiple steps (e.g., 2, 3, or more nodes).
- **Provide a cond** ...ps if a chain is provided, or dir...
- **Avoid overly si** ...al relationships (e.g., avoid vag...
- **Format the output exactly as follows:**

  If including a chain:

  `Causal Chain: [Your causal ... 3 → C → D]`

  `Answer: [Your answer, explaining the reasoning based on the chain]`

  If omitting the chain:

  `Answer: [Your answer, explaining the causal reasoning]`

Examples:

Question: How would the scene change if a sudden rainstorm began in a sunny park?

`Causal Chain: Rainstorm → Wet Ground → Darkened Atmosphere`

`Answer: The rainstorm wets the ground, creating a darker, moodier atmosphere as rain falls.`

Question: What would the park look like if it had been photographed at night instead of midday?

`Causal Chain: Night → Reduced Natural Light`

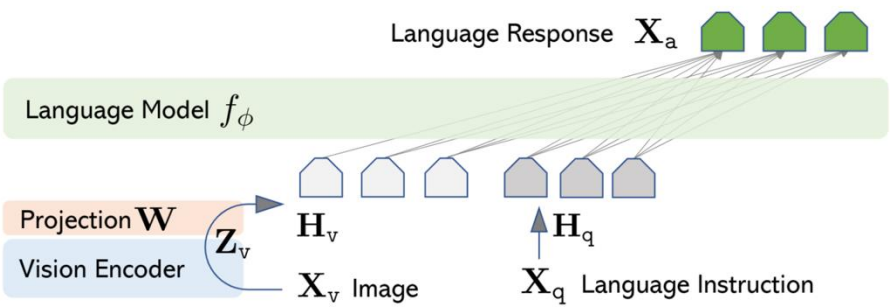
`Answer: At night, the lack of natural light darkens the park, with only artificial lights visible.`

Scaffolding through prompts

- VLMs inherit the camera's extractive vision paradigm (frozen-moment vantage point).
- When evaluating VLM causal reasoning, the benchmarks never enquire two foundations:
  - Do VLMs encode time as the medium through which causes unfold?
  - Do VLMs reason on their own, or only when **scaffolding** through prompts do the work?

The Evidence

- Preliminary study: three VLMs (same architecture, different training).
- All generated fluent explanations.
- But none could trace cause to effect.
- The gap!
- When we changed the prompts, each failed differently.
- None had internalized causal reasoning.



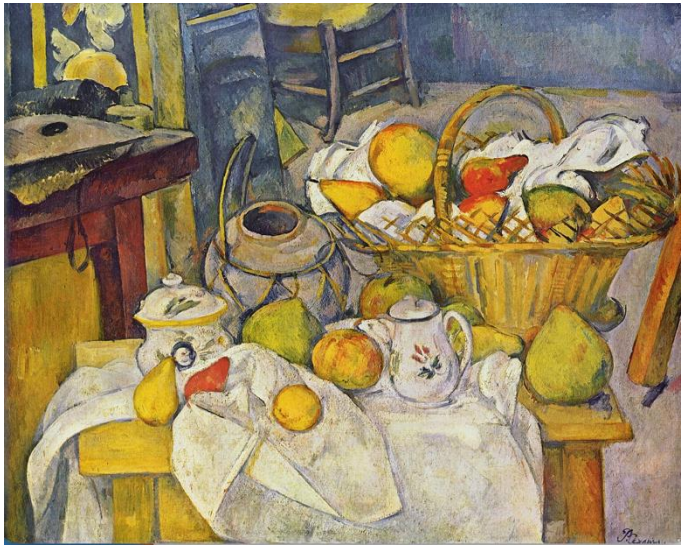
LLaVA-NeXT 13B

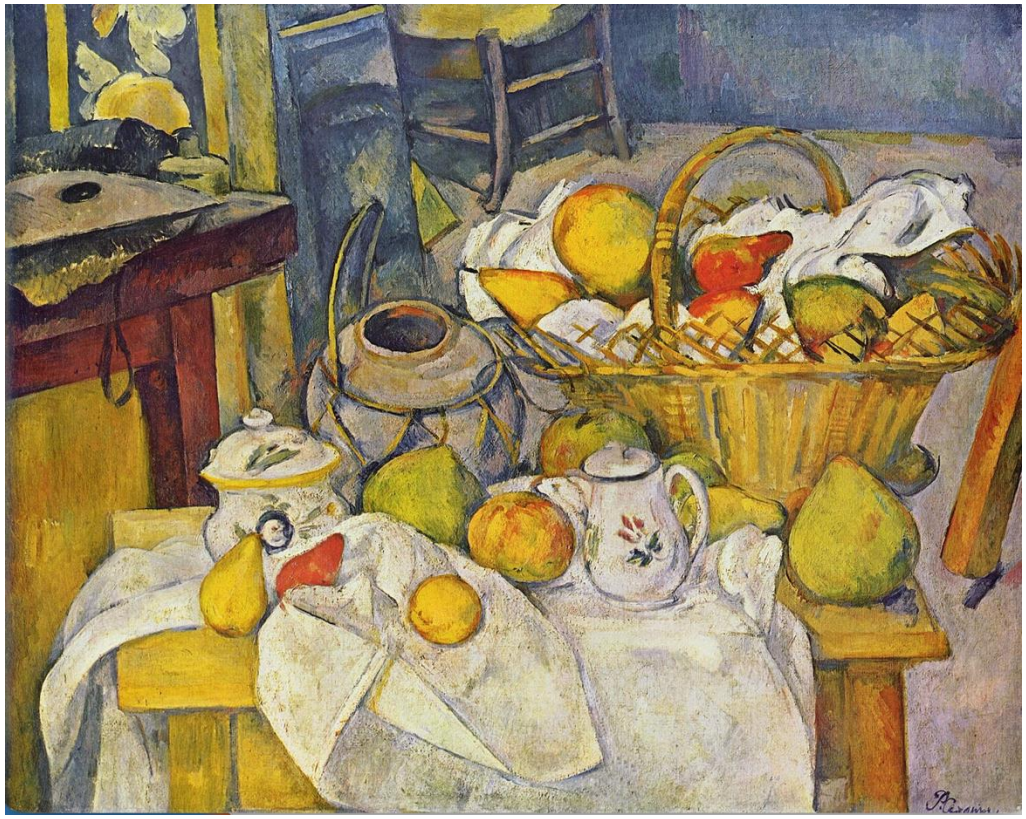
LLaVA-RLHF 13B

LLaVA-CoT 11B

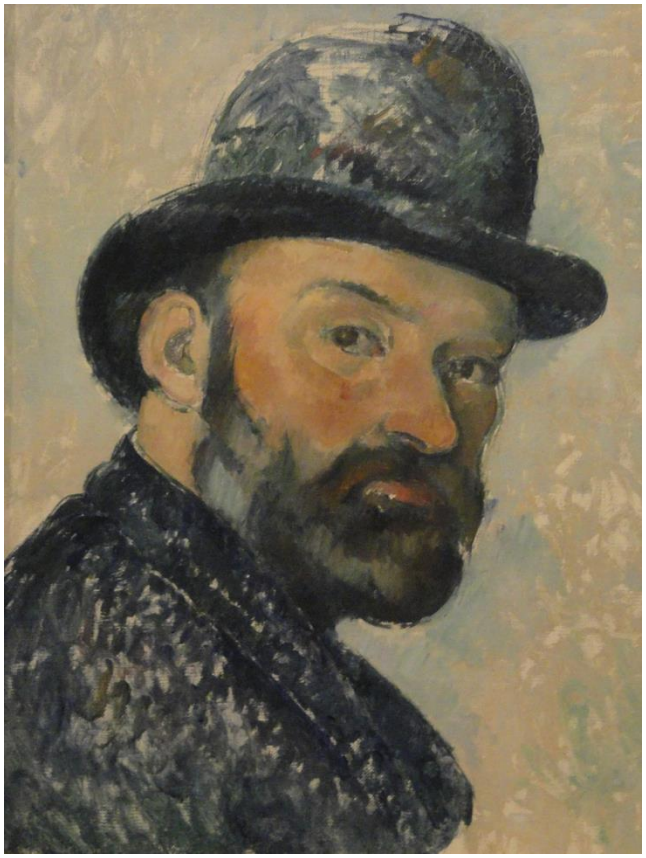
The Position

- VLM causal benchmarks test outputs, not foundations.
- Benchmark success may reflect scaffolding optimization instead of causal reasoning.
- The gap between linguistic fluency and genuine causal reasoning will persist until evaluation shifts from outputs to foundations.





Progress requires benchmarks that test temporal understanding and scaffolding-invariance.



AI has not yet grappled with what Cézanne discovered.