

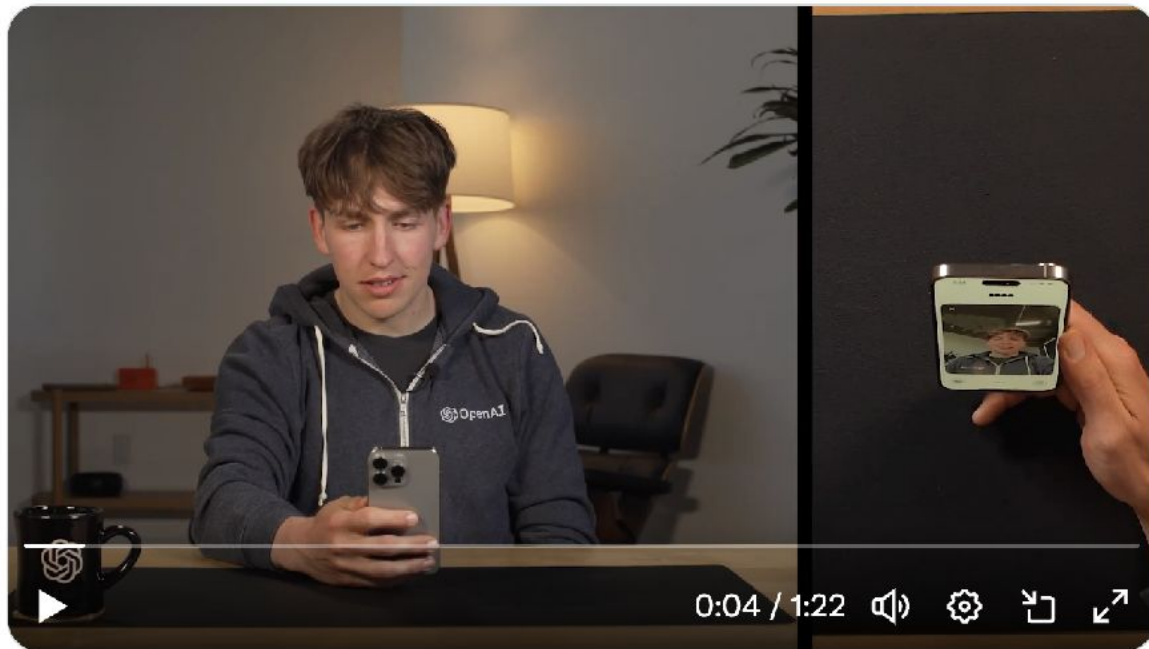
BLINK : Multimodal Large Language Models Can See but Not Perceive

Xingyu Fu



Multimodal LLMs are Powerful!

- They can talk to people



Multimodal LLMs are Powerful!

- They can think (?)

Did Aristotle Use a Laptop?



No, Aristotle did not use a laptop. Aristotle was a Greek philosopher who lived from 384 to 322 BCE, long before the invention of laptops or any modern technology. His work was done using the tools available in his time, such as scrolls, tablets, and the spoken word for teaching and communicating his ideas. The concept of laptops and computers would not emerge until many centuries later.

Multimodal LLMs are Powerful!

- They can see the world



What can you see?



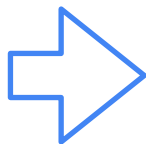
In the image, I see a cat with a light brown and white fur coat sitting on a blue chair. The cat appears to be wrapped in or resting on a dark gray fabric, possibly a piece of clothing or a small blanket. The cat's expression seems slightly grumpy or annoyed. The background includes part of the chair and a light-colored cushion.



Can they do this...?

- Can multimodal LLMs understand 3D space?

Question: Is the camera moving towards left or right?



(A) Left



(B) Right



GPT 4V



Gemini:
I don't know

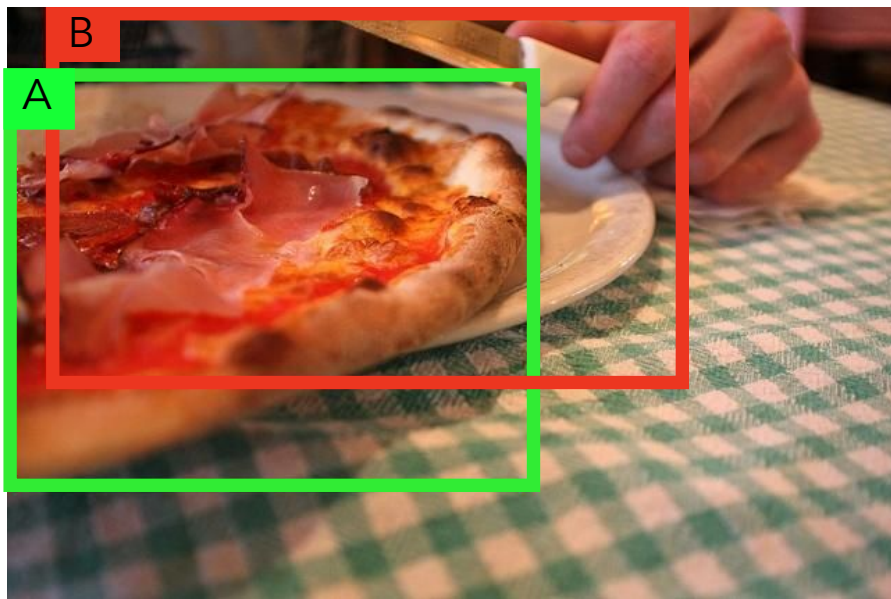




Let's try a simpler example

- Can multimodal LLMs identify an object?

Question: Which bounding box localizes the pizza better?



(A)



(B)



GPT 4V



Gemini Pro





More simpler example!

- Can multimodal LLMs play jigsaw?

Question: Which image fits the missing part?



(B)



GPT 4V



Gemini Pro



- Can models really see as humans do?
- Do models perceive the world as humans do?
- What's the difference between models' visual world and humans'?



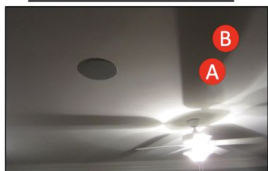
BLINK Benchmark

Relative depth



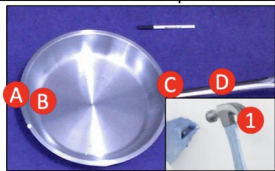
Which point is closer?

Relative reflectance



Which point is darker?

Functional correspondence



Which points have similar affordance when pulling out a nail?

Jigsaw



Which image fits here?

Multi-view reasoning



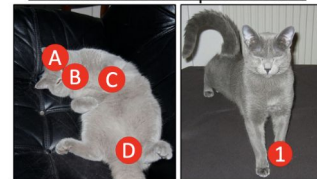
Is camera moving right?

Visual correspondence



Which point is the same?

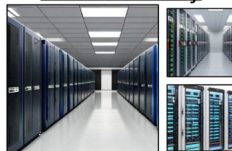
Semantic correspondence



Which points have similar semantics?

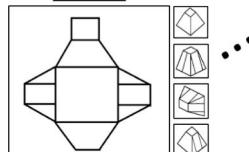


Visual similarity



Which image is more similar to the left?

IQ Test



Which object does it fold into?

Forensics detection



Which image is real?

What is BLINK

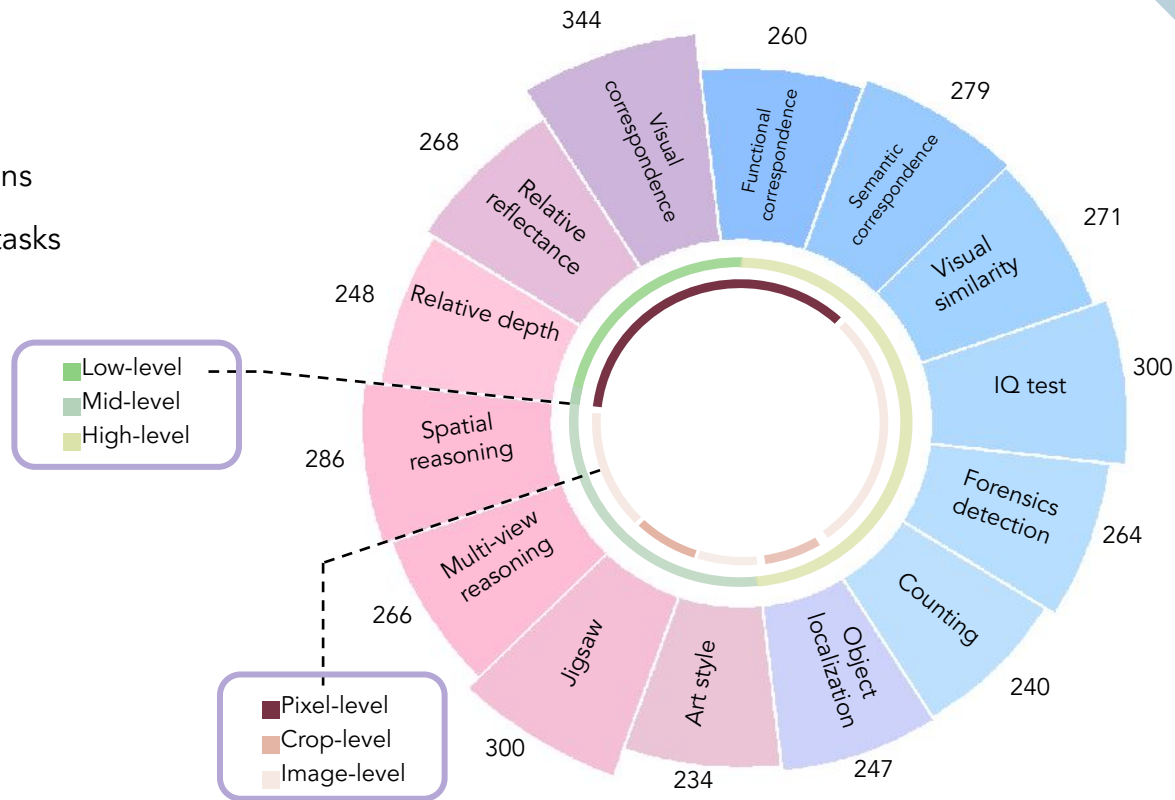
A new benchmark to test multimodal models' visual perception ability, on problems that humans can solve with a “blink” of an eye.

BLINK Benchmark contains...

7,358 unique images

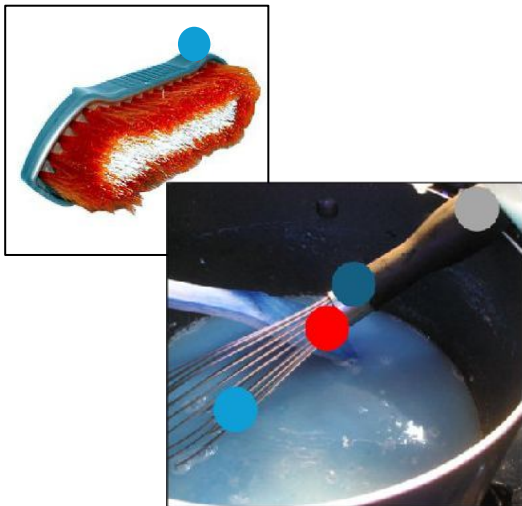
3,807 unique multiple-choice questions

Over 14 visual perception-demanding tasks

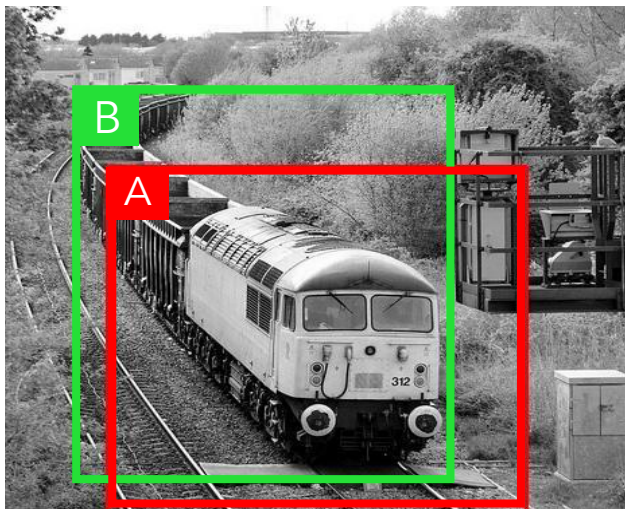


BLINK unique features

 Diverse visual prompting **VS** Text prompt only



Q: Which points have similar affordance?



Q: Which box localizes train better?



Q: Which image fits here better?

BLINK unique features



Beyond recognition

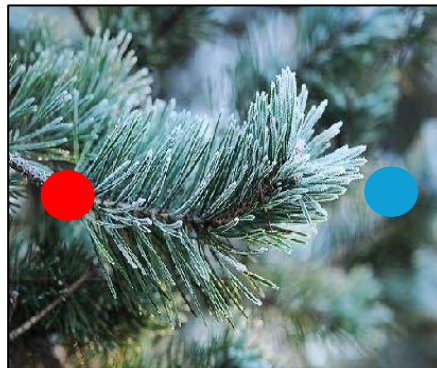
VS

Focus mainly on recognition

(e.g., 3D/reflectance estimation)



Q: Is the camera moving clockwise around the object?



Q: Which point is farther?



Q: Which point is darker?

BLINK unique features



“Visual” commonsense

VS

Require domain knowledge

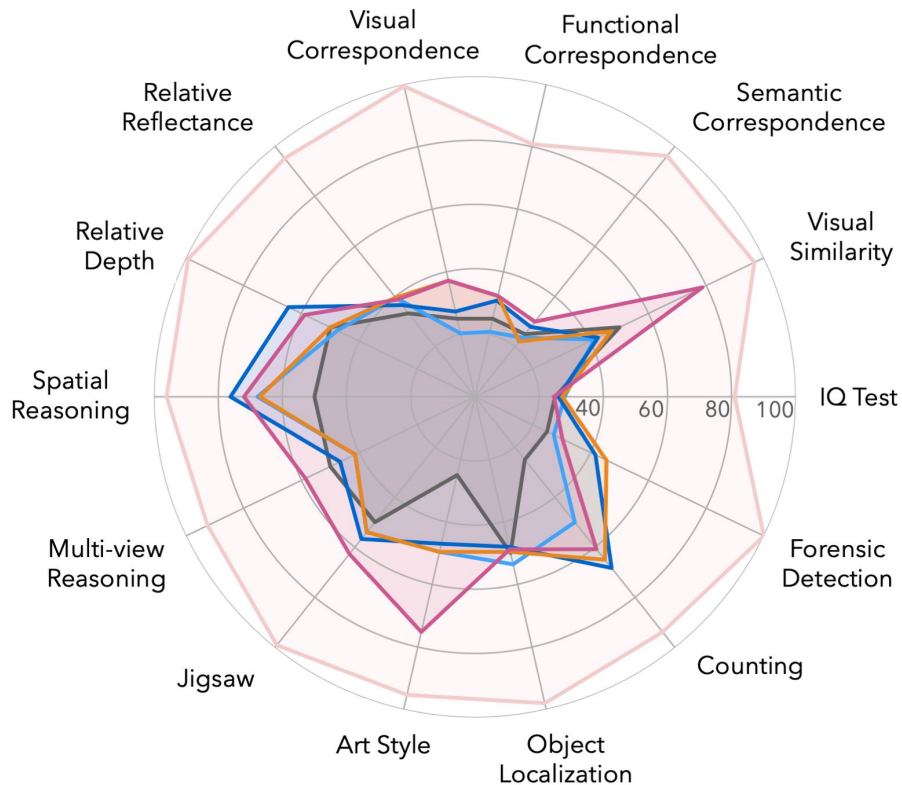


Q: Which image is real?



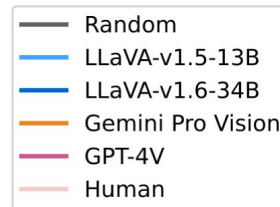
Q: Which image is more similar to the first one?

BLINK Benchmark Results

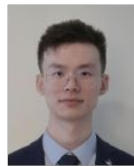
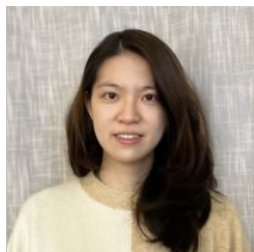


Takeway: Models are far worse than humans!

- ❑ What are the common errors?
- ❑ How do visual prompts affect model outputs?
- ❑ Can specialist models solve BLINK?
- ❑



BLINK Benchmark



Many thanks to our amazing collaborators! 🔥



Project page