VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use

Yonatan Bitton*, Hritik Bansal*, Jack Hessel*, Rulin Shao, Wanrong Zhu,
Anas Awadalla, Josh Gardner, Rohan Taori, Ludwig Schimdt

1. Why VisIT-Bench?

- Diverse tasks and human-chatbot interactions
- Converts 25 datasets into chatbot-friendly formats
- Reflect the dynamic demands of modern chatbots:

👉'Wish-list' instructions

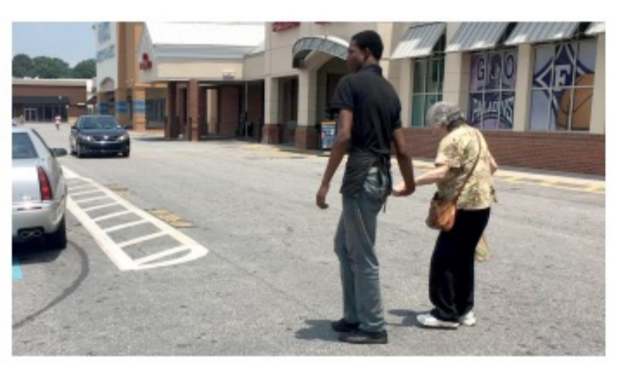
👉70 tested skills

👉Repurposing of existing datasets (+multi-image tasks)

Contextual Knowledge of Events

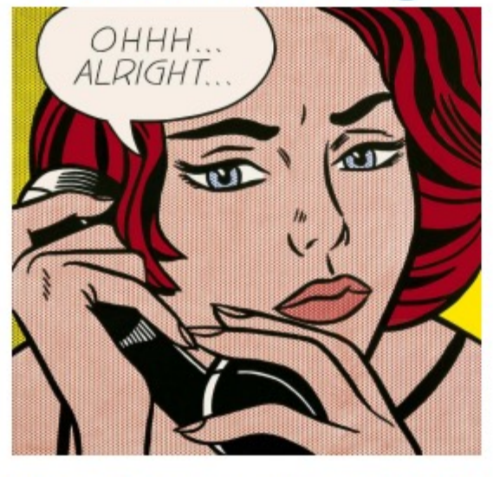
Tell me what is notable or important about the event in this photo.

Figurative Speech Explanation



Someone said that this man is an angel. Why?

Art Knowledge

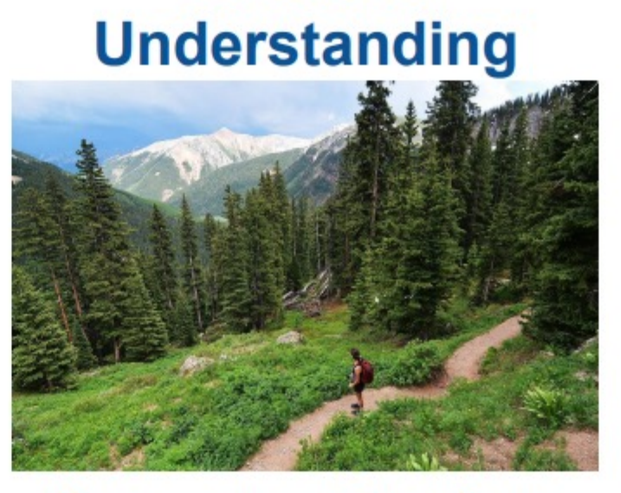


Teach me about this painting.

Recognition

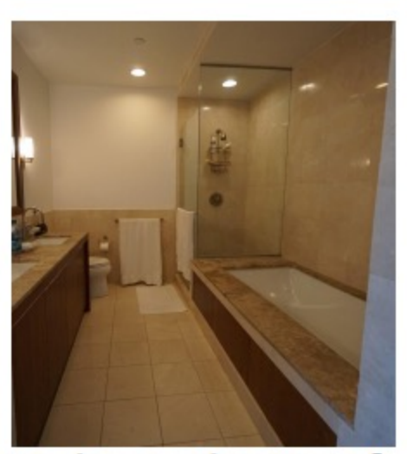
Where is this?

Location Understanding



If you are going for a picnic at this location, what items should you carry with you?

Home Renovation



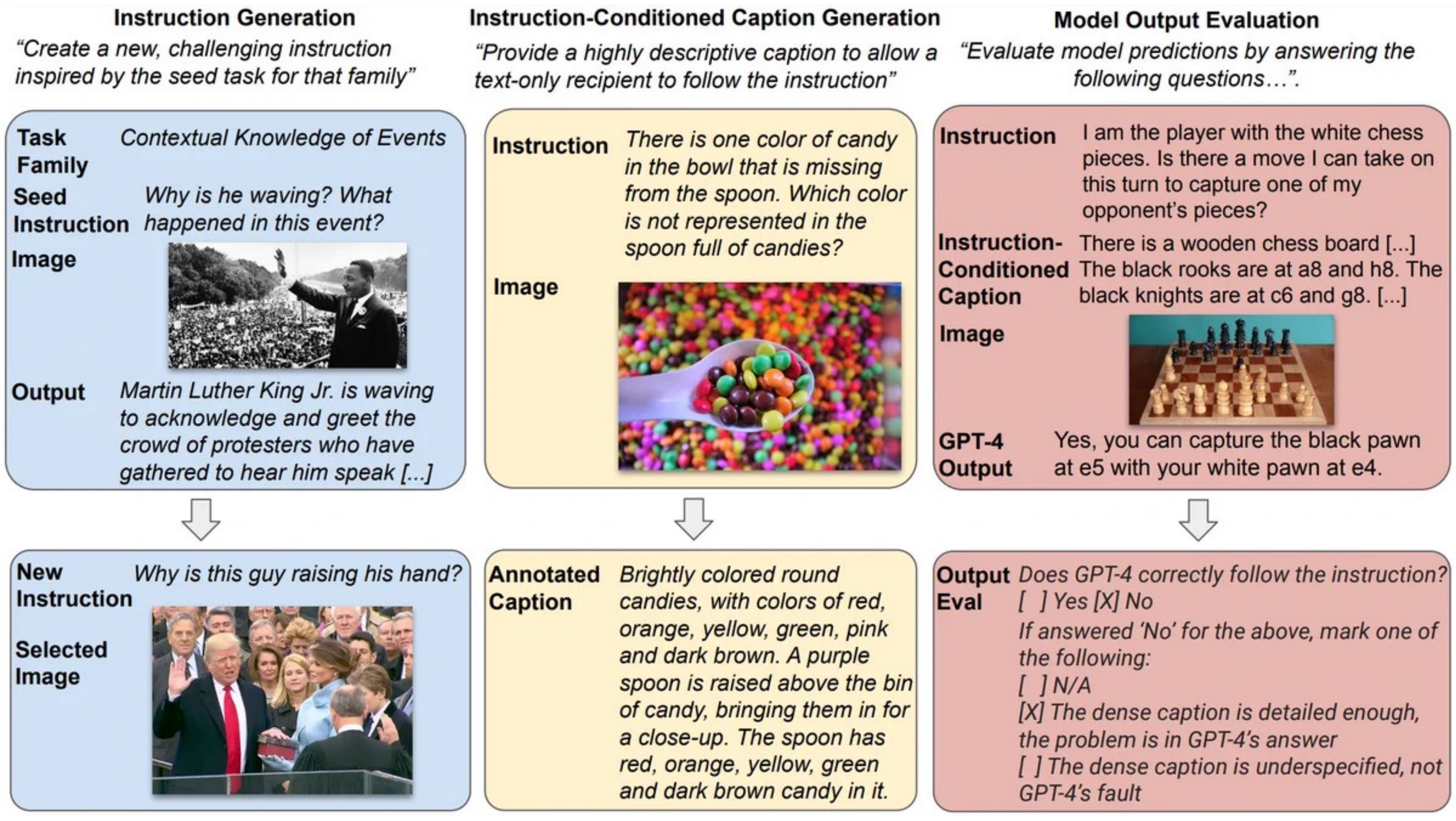
Here is a photo of my bathroom. How can I design it nicer?

687 Examples

Instruction	Instruction-Conditioned Caption	Human-Verified GPT4 Response
Create a catchy title for a country song based upon the advice printed on the wheel cover.	There is a jeep parked on the street with the top down. There is a wheel cover on the spare tire in the back with an inscription that says, "Quit your Job, Buy a Ticket, Get a Tan, Fall in Love, Never Return."	"Tan Lines & Ticket Stubs: A Love Story Unbound" - A country song about leaving it all behind, finding love, and embracing the freedom of the open road.

2. Data Collection

- Create "wish-list" instructions for desired V&L chatbot capabilities
- Use these as inspiration for instructions annotation
- Collect instruction-conditioned dense captions
- Generate human-verified chatbot responses from GPT-4 outputs



Instruction Generation

"Create a new, challenging instruction inspired by the seed task for that family"

Task Family *Contextual Knowledge of Events*

Seed Instruction *Why is he waving? What happened in this event?*

Image

Output *Martin Luther King Jr. is waving to acknowledge and greet the crowd of protesters who have gathered to hear him speak [...]*

New Instruction *Why is this guy raising his hand?*

Selected Image

Instruction-Conditioned Caption Generation

"Provide a highly descriptive caption to allow a text-only recipient to follow the instruction"

Instruction *There is one color of candy in the bowl that is missing from the spoon. Which color is not represented in the spoon full of candies?*

Image

Annotated Caption *Brightly colored round candies, with colors of red, orange, yellow, green, pink and dark brown. A purple spoon is raised above the bin of candy, bringing them in for a close-up. The spoon has red, orange, yellow, green and dark brown candy in it.*

Model Output Evaluation

"Evaluate model predictions by answering the following questions…".

Instruction I am the player with the white chess pieces. Is there a move I can take on this turn to capture one of my opponent's pieces?

Instruction-Conditioned Caption There is a wooden chess board [...] The black rooks are at a8 and h8. The black knights are at c6 and g8. [...]

Image

GPT-4 Output Yes, you can capture the black pawn at e5 with your white pawn at e4.

Output Eval *Does GPT-4 correctly follow the instruction?*

[ ] Yes [X] No

If answered 'No' for the above, mark one of the following:

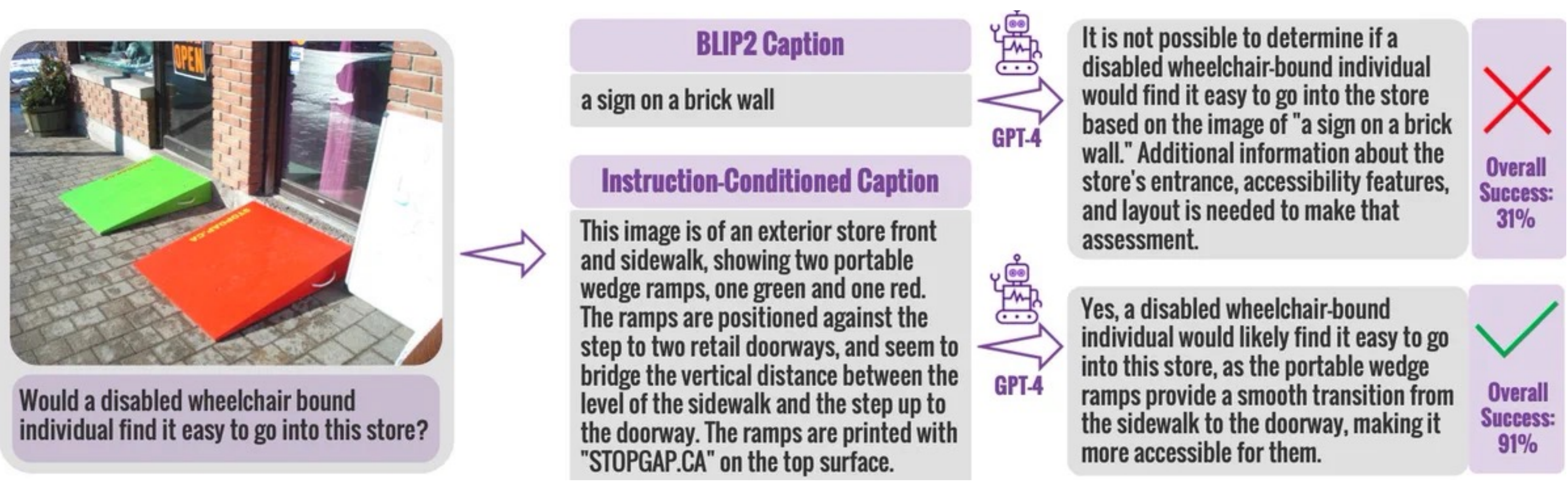
[ ] N/A

[X] The dense caption is detailed enough, the problem is in GPT-4's answer

[ ] The dense caption is underspecified, not GPT-4's fault

3. Instruction-Conditioned Captions

- 91.5% success rate in single-image scenarios
- Data collection demonstrates the effectiveness of instruction-conditioned dense captions.
- Showcases the necessity of VisIT-Bench dense captions over generated captions from a SoTA BLIP2 captioning model.



Would a disabled wheelchair bound individual find it easy to go into this store?

BLIP2 Caption

a sign on a brick wall

It is not possible to determine if a disabled wheelchair-bound individual would find it easy to go into the store based on the image of "a sign on a brick wall." Additional information about the store's entrance, accessibility features, and layout is needed to make that assessment.

Overall Success: 31%

Instruction-Conditioned Caption

This image is of an exterior store front and sidewalk, showing two portable wedge ramps, one green and one red. The ramps are positioned against the step to two retail doorways, and seem to bridge the vertical distance between the level of the sidewalk and the step up to the doorway. The ramps are printed with "STOPGAP.CA" on the top surface.

Yes, a disabled wheelchair-bound individual would likely find it easy to go into this store, as the portable wedge ramps provide a smooth transition from the sidewalk to the doorway, making it more accessible for them.

Overall Success: 91%

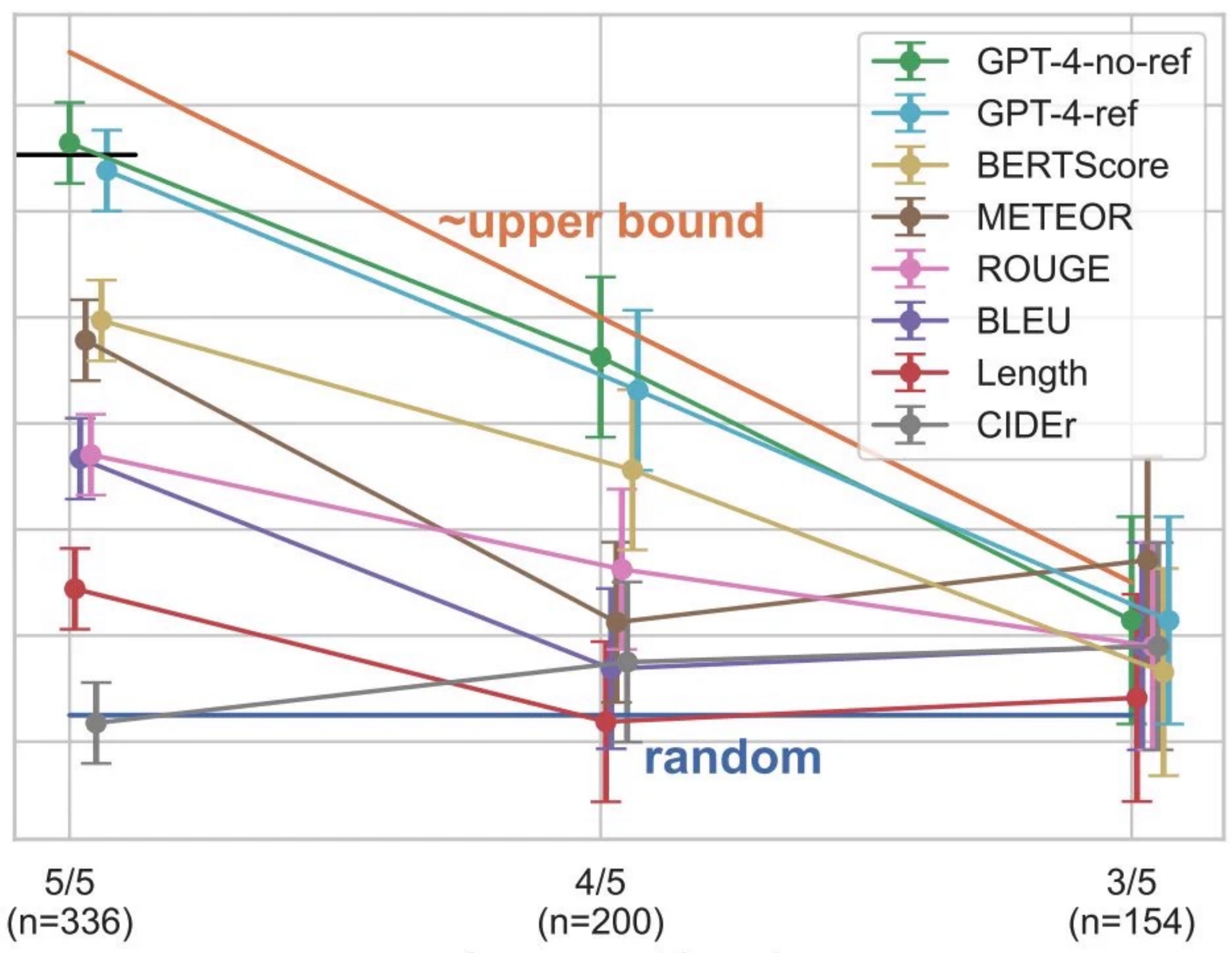
4. Results

- VisIT-Bench facilitates the comparison of different V&L models.
- **How good is our automatic metric?**

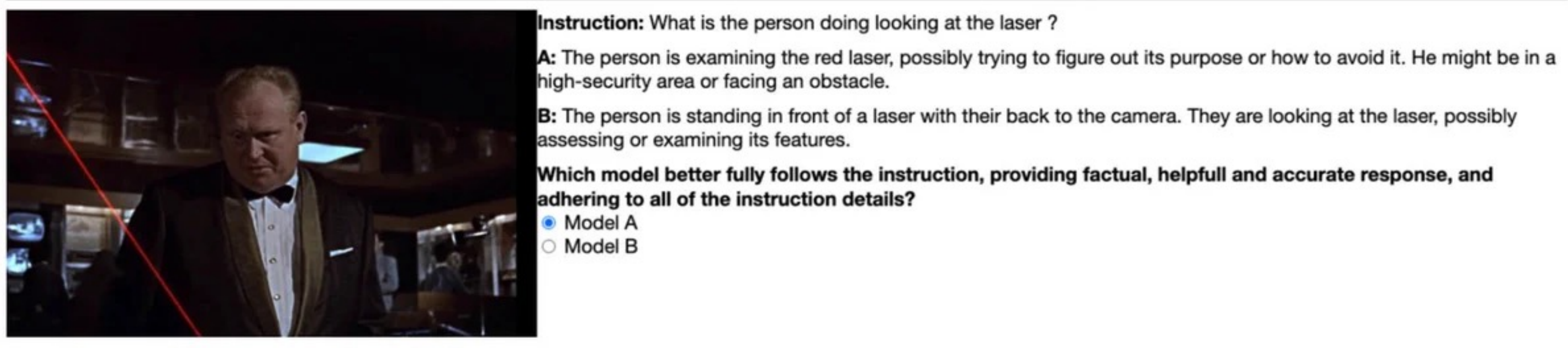
➢Correlations of automatic metrics vs. human preferences, with reference free (GPT-4-no-ref)

➢Showing strong alignment

top orange line - upper bound, bottom blue line - random chance (50%)



Human-preferences based Leaderboard



Instruction: What is the person doing looking at the laser ?

A: The person is examining the red laser, possibly trying to figure out its purpose or how to avoid it. He might be in a high-security area or facing an obstacle.

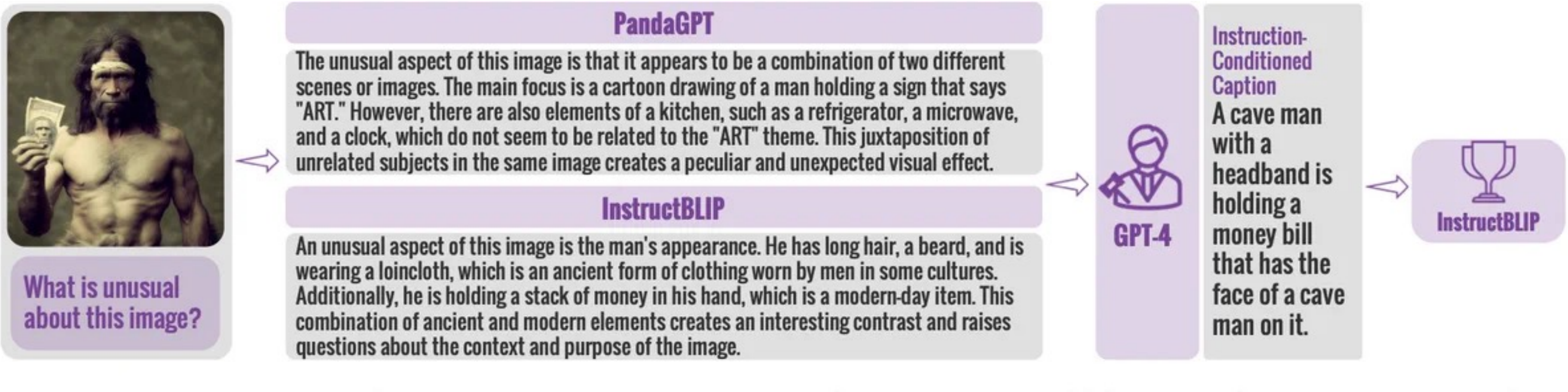
B: The person is standing in front of a laser with their back to the camera. They are looking at the laser, possibly assessing or examining its features.

Which model better fully follows the instruction, providing factual, helpfull and accurate response, and adhering to all of the instruction details?

● Model A
○ Model B

	Model	Elo	matches	Win-rate vs. reference (w/ # ratings)
Single Image	Human Verified GPT-4 Reference	1223	1439	–
	LLaVA (13B)	1085	1462	26.23% (n=244)
	LlamaAdapter-v2 (7B)	1061	1507	**27.41%** (n=259)
	mPLUG-Owl (7B)	995	1345	14.95% (n=214)
	InstructBLIP (13B)	957	1315	12.37% (n=194)
	MiniGPT-4 (7B)	893	1513	14.72% (n=299)
	PandaGPT (13B)	786	1441	10.48% (n=229)
Multiple Images	Human Verified GPT-4 Reference	1193	210	–
	mPLUG-Owl	997	190	15.38% (n=78)
	Otter	917	147	3.17% (n=63)
	OpenFlamingo	893	171	4.35% (n=69)

Auto-rater based Leaderboard



What is unusual about this image?

PandaGPT

The unusual aspect of this image is that it appears to be a combination of two different scenes or images. The main focus is a cartoon drawing of a man holding a sign that says "ART." However, there are also elements of a kitchen, such as a refrigerator, a microwave, and a clock, which do not seem to be related to the "ART" theme. This juxtaposition of unrelated subjects in the same image creates a peculiar and unexpected visual effect.

InstructBLIP

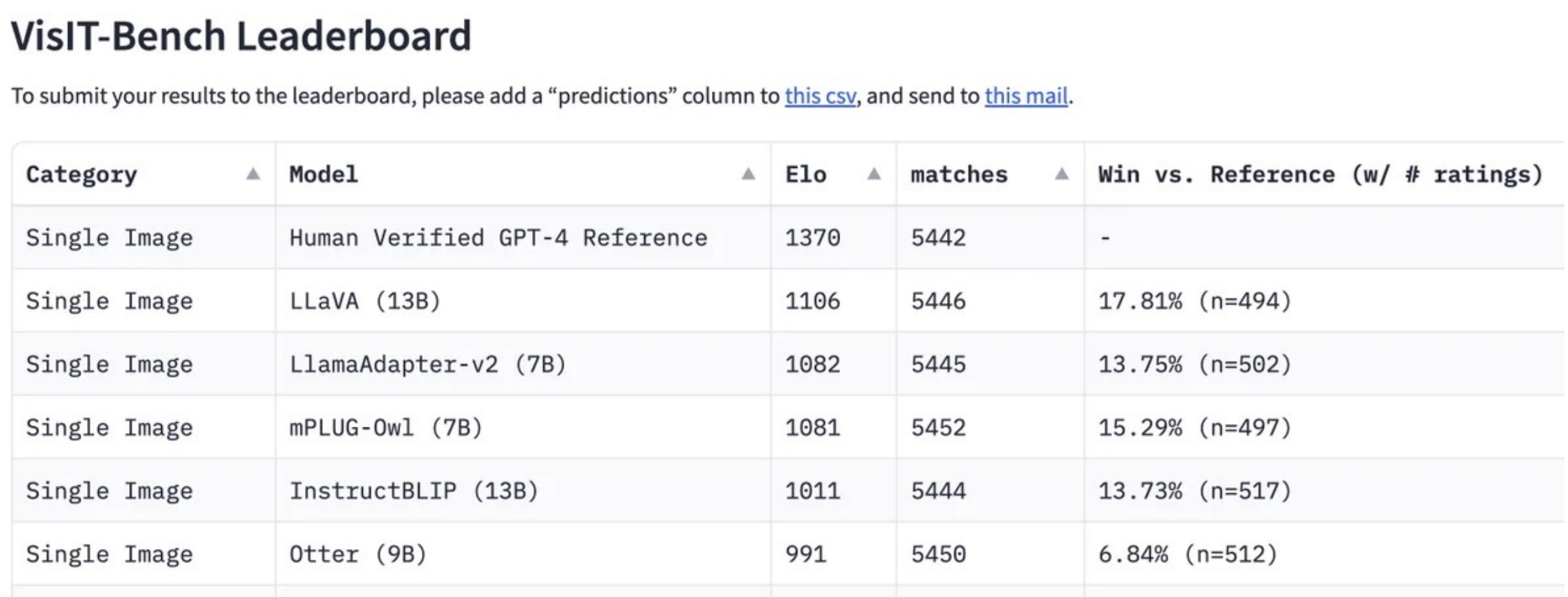
An unusual aspect of this image is the man's appearance. He has long hair, a beard, and is wearing a loincloth, which is an ancient form of clothing worn by men in some cultures. Additionally, he is holding a stack of money in his hand, which is a modern-day item. This combination of ancient and modern elements creates an interesting contrast and raises questions about the context and purpose of the image.

Instruction-Conditioned Caption

A cave man with a headband is holding a money bill that has the face of a cave man on it.

	Model	Elo	matches	Win vs. Reference (w/ # ratings)
Single Image	Human Verified GPT-4 Reference	1370	5442	-
	LLaVA (13B)	1106	5446	**17.81%** (n=494)
	LlamaAdapter-v2 (7B)	1082	5445	13.75% (n=502)
	mPLUG-Owl (7B)	1081	5452	15.29% (n=497)
	InstructBLIP (13B)	1011	5444	13.73% (n=517)
	Otter (9B)	991	5450	6.84% (n=512)
	VisualGPT (Da Vinci 003)	972	5445	1.52% (n=527)
	MiniGPT-4 (7B)	921	5442	3.26% (n=522)
	OpenFlamingo (9B)	877	5449	2.86% (n=524)
	PandaGPT (13B)	826	5441	2.63% (n=533)
	Multimodal GPT	763	5450	0.18% (n=544)
Multiple Images	Human Verified GPT-4 Reference	1192	180	-
	mPLUG-Owl	995	180	6.67% (n=60)
	Otter	911	180	1.69% (n=59)
	OpenFlamingo	902	180	1.67% (n=60)

👉Add your models to VisIT-Bench Leaderboard!



VisIT-Bench Leaderboard

To submit your results to the leaderboard, please add a "predictions" column to this csv, and send to this mail.

Category ▲	Model ▲	Elo ▲	matches ▲	Win vs. Reference (w/ # ratings)
Single Image	Human Verified GPT-4 Reference	1370	5442	-
Single Image	LLaVA (13B)	1106	5446	17.81% (n=494)
Single Image	LlamaAdapter-v2 (7B)	1082	5445	13.75% (n=502)
Single Image	mPLUG-Owl (7B)	1081	5452	15.29% (n=497)
Single Image	InstructBLIP (13B)	1011	5444	13.73% (n=517)
Single Image	Otter (9B)	991	5450	6.84% (n=512)