Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images

Nitzan Guetta Bitton, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, Roy Schwartz

What Makes These Images Weird?

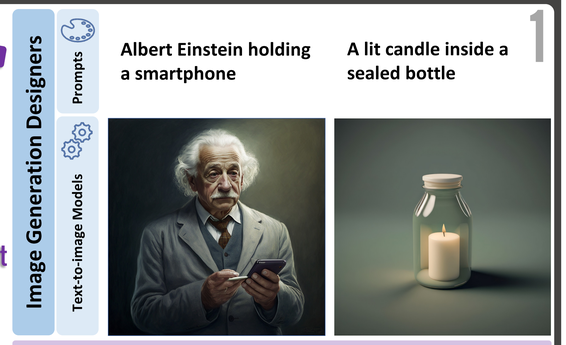
Recognition is NOT Enough to Solve WHOOPS!

- Commonsense-violating images
- Designers were hired via social media
- Text-to-Image generation models

MSCOCO / ImageNet

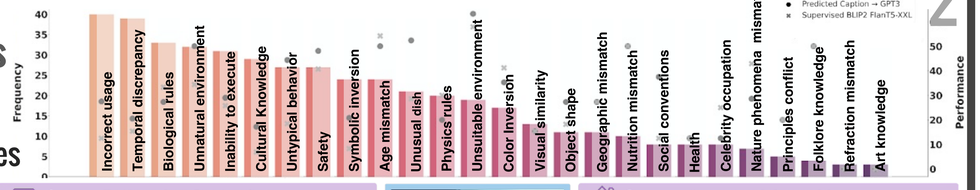
How to Create a Weird Image?

- Depict two likely to co-occur items X (=Slash) and Y (=Guitar)
- Switch Y to Y' (=Saxophone) that is unlikely to co-exist with X



WHOOPS! Consists of Four Tasks

- 500 synthetic images
- 10,874 annotations
- A novel task



1. Explanation Generation
   - Model: Pandas live in the China bamboo forests, subsist almost entirely on bamboo, and do not hunt salmon fish like the grizzly bears.
2. Image Captioning
   - Model: A panda bear is catching salmon fish up in the river stream.

3. Cross-Modal Matching
   - C: A panda bear is fishing for salmon
   - U: A bear is fishing for salmon
   - Model: sim(Img, C) > sim(Img, U)
4. VQA
   - Q: What is catching salmon in a stream? — Model: A panda
   - Q: What does a panda bear try to catch? — Model: A salmon

End-To-End

- Zero-shot BLIP2: "the wolf is howling at the sun."
- Fine-tuned BLIP2: "wolfs usually howls during the night, not the day."

Pipeline

Predicted Caption

- Caption BLIP2: "wolf howling on top of the rock at sunset"
- Explanation GPT3: "a wolf howling on top of a rock at sunset, which is not a typical behavior of a wild wolf"

Ground-Truth Caption (Oracle)

- Caption Human Curated: "a wolf howling , bright sunny day background"
- Explanation GPT3: "a wolf howling in the middle of a bright, sunny day, which is unusual because wolves are typically most active during the night."

Two Approaches for Explanation-of-Violation Generation

- End-to-end : Image -> model -> explanation
- Pipeline: predicted/ ground-truth image caption -> language reasoning model -> explanation

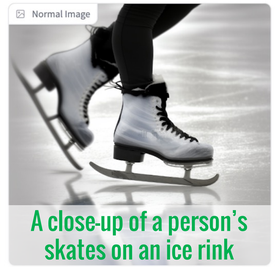
Results

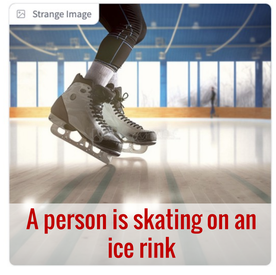
BLIP2 Flan T5-XXL (11B): The Statue of Liberty is in New York, not Sydney

BLIP2 Flan T5-XL (3B): The cookie monster eats cookies, not apples

BLIP2 Flan T5-XXL (11B): Babies are too young to read books

	Task	Identify	Explain		
		Binary Accuracy (↑)	Human Rating (↑)	GPT4 Rating (↑)	GPT4 Rating Accuracy (↑)
End-to-end	BLIP2 FlanT5-XXL (Zero-shot)	50	0	12	88
	BLIP2 FlanT5-XL (Fine-tuned)	60	15	18	87
	BLIP2 FlanT5-XXL (Fine-tuned)	73	27	27	81
	InstructBLIP	-	-	31	-
	mPLUG-Owl	-	-	24	-
	LLaVA	-	-	31	-
Pipeline (Zero-shot)	Predicted Caption → GPT3	59	33	36	87
	Ground-truth Caption → GPT3 (Oracle)	74	68	70	81
	Predicted Caption → GPT4	-	-	36	-
	Predicted Caption → Llama-2-7b	-	-	36	-
	Predicted Caption → Llama-2-13b	-	-	36	-
	Ground-truth Caption → GPT4 (Oracle)	-	-	69	-
	Ground-truth Caption → Llama-2-7b (Oracle)	-	-	71	-
	Ground-truth Caption → Llama-2-13b (Oracle)	-	-	70	-
	Humans	92	95	-	-





Main Challenge- Weirdness, not Synthesis

- Analyzing non-weird synthetic and natural images for each WHOOPS! image.

👉Weirdness is the primary challenge