A. Appendix

Magic icon in the title created by Freepik - Flaticon https://www.flaticon.com/free-icons/magic.

The following sections are relevant for **reproducibility** and providing more details and examples about the models evaluation, training and data collection.

A.1. Model versions

We take the official implementations for CLIP: https://github.com/openai/CLIP, OFA: https://github.com/OFA-Sys/OFA, CoCa: colab.research.google.com/github/mlfoundations/open_clip/blob/master/docs/Interacting_with_open_coca.ipynb, BLIP: https://colab.research.google.com/github/salesforce/BLIP/blob/main/demo.ipynb, BLIP2: https://github.com/salesforce/LAVIS/tree/main/projects/blip2. We take the versions specified in the paper without changing the default hyper-parameters.

A.2. Supervised Data Details

We format all supervised tasks in a sequence-to-sequence fashion. All training examples have the format $\langle x_{image}, x_{text}, y_{text} \rangle$. Where BLIP-2 Flan-T5 is trained to maximize the probability of the textual target given the image and prompt inputs, i.e., $P(y_{text}|x_{image}, x_{text})$. For each cross-validation split, we train a single multitask model for all tasks by setting the prompt in a strategic fashion. Example input/output targets are given in Table 4. We provide the training splits for completeness, although we recommend using WHOOPS! primarily as a test set.

The models were trained using 8xA6000 GPUs. A single training run for 15 epochs takes around 2 hours. To train all 5 splits, over two learning rates and three models, it requires approximately 72 hours of compute time on a machine with 8xA6000 GPUs, which is equivalent to about 576 GPU hours.

A.3. Image Generation Designers Guidelines

The task is to create an image that depicts something "weird" that will be intuitive for humans to understand and challenging to artificial intelligence models. The images should be relatively realistic, with one weird thing that requires the use of logic and general knowledge. The goal is to compare the explanations provided by the models and people to determine whether the models struggle more than humans.

To ensure the task is challenging for AI models, the criteria for "weird" should be conceptual and not directly related to the rest of the picture. For example, there should be no other illogical things in the picture, such as distorted objects or more than ten fingers.

To create prompts, participants should replace X_1 with some X_2 in situations where X and Y appear together in a normal way in the real world. The similarity between X_1 and X_2 should mislead the models when they ask "What's weird in the picture?" The prompts should include cultural, general knowledge, times, and behavioral elements that make the picture illogical, but not directly related to what is happening in the picture.

Participants will receive a link to their own shared directory where they can upload their high-quality images and prompts that follow the naming conventions provided by Midjourney.[8] The prompts that create the images should be recorded, and formats with seed should be used to allow the images to be restored later.

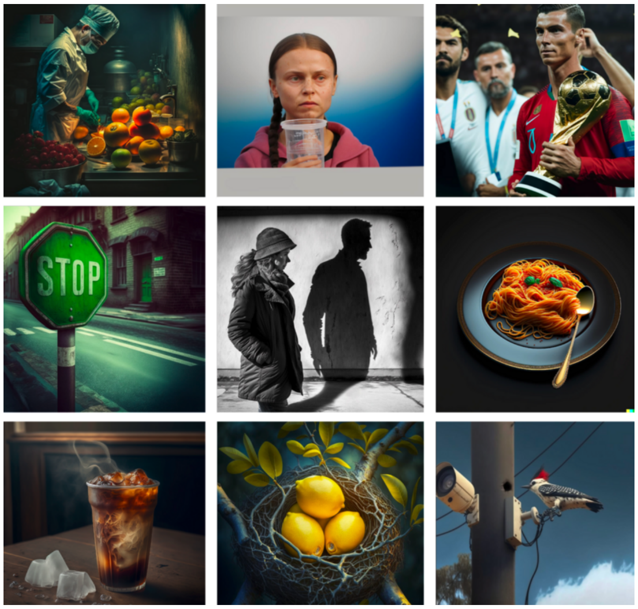


Figure 9: Designer Guidelines Examples.

Example of prompts for the images in Fig. 9:

1. Ronaldo (X_2) with the World Cup trophy (Y) instead of Messi (X_1) with the World Cup trophy (Y)

2. Greta Thunberg (Y) holds a disposable cup (X_2) instead of a reusable cup (X_1)

3. A surgeon (Y) in the kitchen with fruit (X_2) instead of an operating room (X_1)

4. Spaghetti plate (Y) with spoon (X_2) instead of fork (X_1)

5. Image of a girl (Y) with a shadow of a man (X_2) instead of a girl (Y) with her own shadow (X_1)

[8] www.midjourney.com

Task	Input (x_{text})	Target (y_{text})
VQA	Question: Through what is the man drinking tea looking at the Earth?	a porthole
Captioning	Describe the image.	A snow plow is plowing sand in the desert.
Matching	Which is better? A: Boys are being rained on. B: A group of children are wearing raincoats in a classroom.	B
Weird Id	Is this normal or weird?	weird
Weird Explain (crowd)	Why is it weird?	Walking in the road is dangerous, especially for a child who should be on a sidewalk instead.
Weird Explain (designer)	Why is it weird, in detail?	For an indoor fire to be safe, it has to be adequately ventilated and contained within a fireproof environment like a fireplace or a modern stove, which is why you don't see a campfire indoors because the fire would quickly spread and destroy everything and the carbon monoxide would suffocate any living creatures.

Table 4: Illustrations of sequence-to-sequence formatting of each task (images omitted, but are also included as inputs). Our supervised models are trained on a concatenation of all training data for all tasks.

6. Green stop sign (X_2) on the street (Y) instead of a red stop sign (X_1)

7. Woodpecker (Y) makes a hole in a metal electric pole (X_2) instead of a tree (X_1)

8. Lemons (X_2) in nest (Y) instead of bird eggs (X_1)

9. Cup of cold coffee (X_2) with steam (Y) instead of a cup of hot coffee (X_1) with steam

A.4. Commonsense Categories

WHOOPS! contains 26 different commonsense categories. Fig. 10 presents examples for 24 of them, the rest can be found in the paper. In Fig. 1 the image of Albert Einstein is categorized as Temporal discrepancy and the candle as Physics rules. The concept of inability to execute refers to a scenario where an object is unable to fulfill its intended purpose due to a change or situation depicted in the image. For instance, in the image shown in Fig. 10, the presence of trees in the forest blocks the wind from reaching the wind turbine, resulting in its inability to generate electricity. The Unnatural Environment category pertains to instances where objects, particularly animals, are depicted in settings that are not their natural habitats, such as a moose found on a tropical beach. On the other hand, the Unsuitable Environment category refers to situations where the object is placed in a location that is not suitable for fulfilling its intended function, as seen in the example of car racing in the Colosseum.

A.5. Human Annotation

Fig. 11 shows an example of the Mechanical Turk user-interface. Fig. 12 shows the instructions given to the annotators.

The basic requirements for our annotation task is percentage of approved assignments above 98%, more than 5,000 approved HITs, the location from the US, UK, Australia or New Zealand. We selected 5 examples from our dataset as qualification test and screen the annotators results.

Fig. 13 shows an example from the VQA verification part, where the annotators are asked to determine whether a visual question answering instance generated by an automatic process is correct.

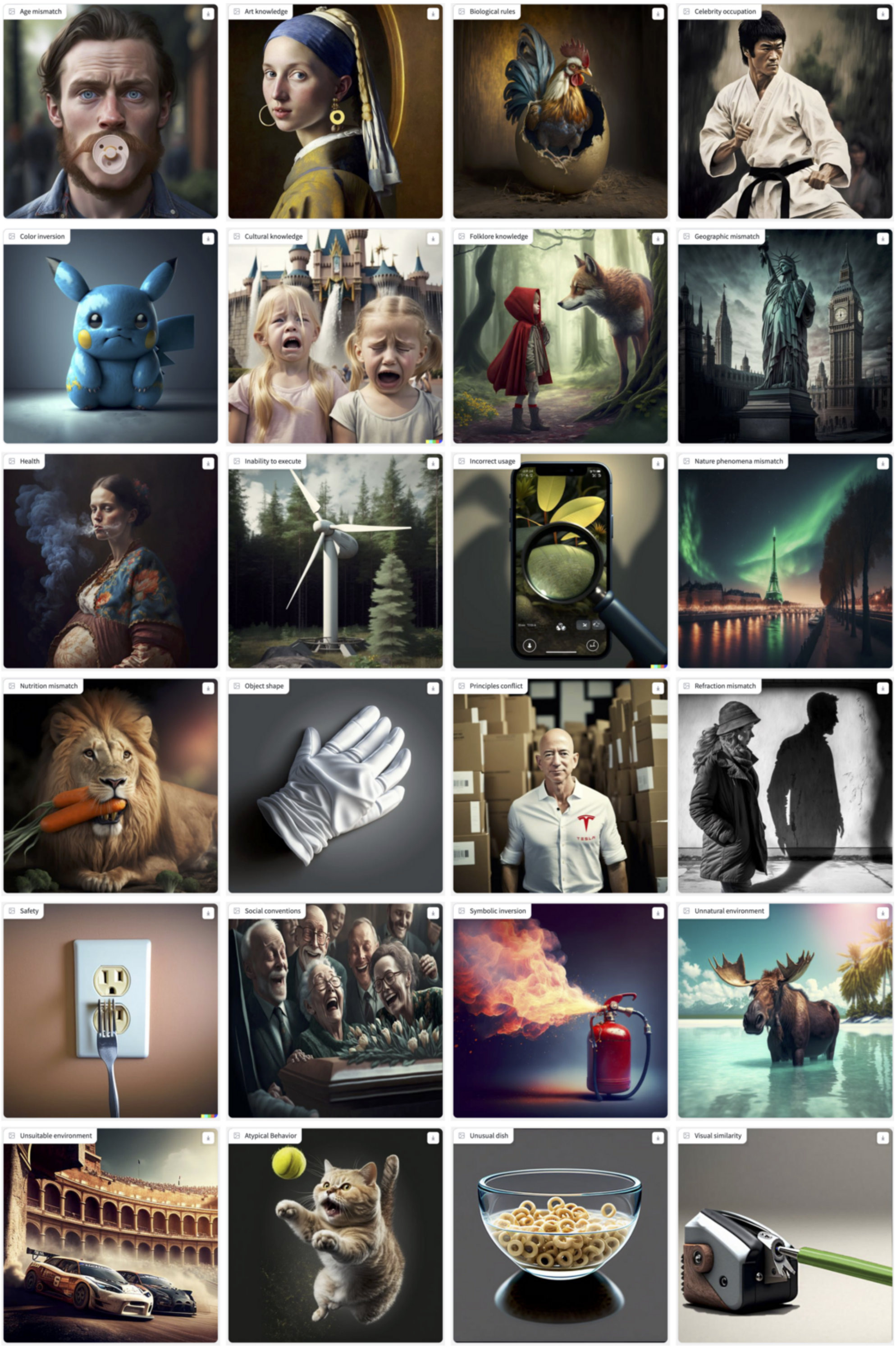


Figure 10: The WHOOPS! images span across various categories and are intended to test AI models in multiple areas of common sense.

Fig. 8 shows an example from the explanation selection part, where the annotators are asked to select correct explanations for the task. The options are both human selections and model predictions, and by aggregate the raters selec-

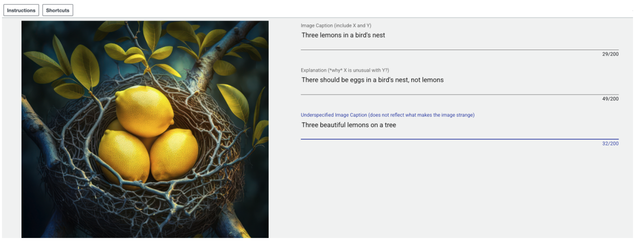


Image Caption (include X and Y)

Three lemons in a bird's nest

29/200

Explanation (*why* X is unusual with Y?)

There should be eggs in a bird's nest, not lemons

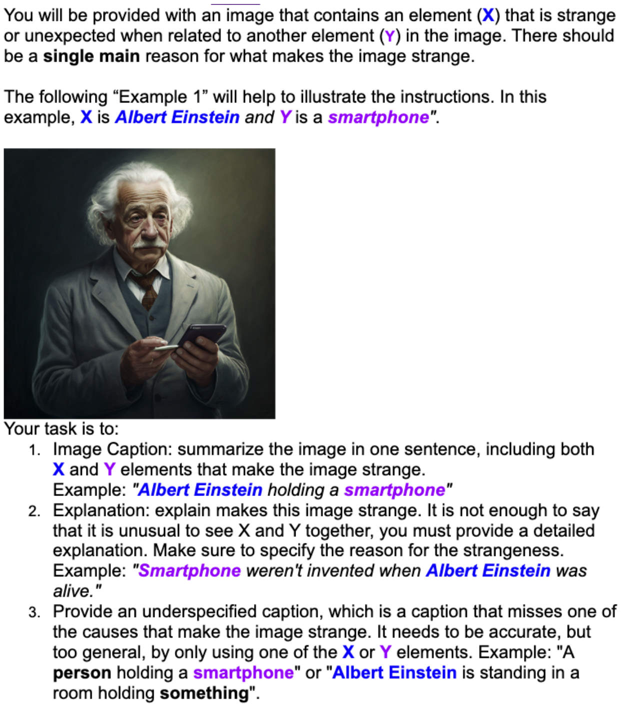
49/200

Underspecified Image Caption (does not reflect what makes the image strange)

Three beautiful lemons on a tree

32/200

Figure 11: Amazon Mechanical Turk Annotators user interface. The annotators receive an image and provide three types of annotations.



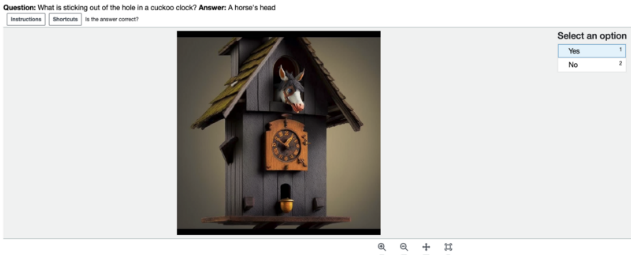
You will be provided with an image that contains an element (**X**) that is strange or unexpected when related to another element (**Y**) in the image. There should be a **single main** reason for what makes the image strange.

The following "Example 1" will help to illustrate the instructions. In this example, **X** is ***Albert Einstein*** *and* **Y** is a ***smartphone***".

Your task is to:

1. Image Caption: summarize the image in one sentence, including both **X** and **Y** elements that make the image strange.
   Example: *"**Albert Einstein** holding a **smartphone**"*
2. Explanation: explain makes this image strange. It is not enough to say that it is unusual to see X and Y together, you must provide a detailed explanation. Make sure to specify the reason for the strangeness.
   Example: *"**Smartphone** weren't invented when **Albert Einstein** was alive."*
3. Provide an underspecified caption, which is a caption that misses one of the causes that make the image strange. It needs to be accurate, but too general, by only using one of the **X** or **Y** elements. Example: "A **person** holding a **smartphone**" or "**Albert Einstein** is standing in a room holding **something**".

Figure 12: Amazon Mechanical Turk Annotators Instructions for Dataset Annotation.



Question: What is sticking out of the hole in a cuckoo clock? **Answer:** A horse's head

Is the answer correct?

Select an option

[✓ Yes] 1

No 2

Figure 13: Amazon Mechanical Turk interface for the task of VQA verification. The annotators receive an (image, question, answer) triplet, and need to determine whether the answer is correct or not.

tions, we can extract human metric performance for both the human annotators who solved the task, and both for the different models. The explanation selection process is critical for obtaining accurate evaluations of model performance. The annotators are presented with a set of options that include both human and model-generated explanations. The selected explanations are then aggregated to derive a human metric for both the annotators and the models. A good explanation should identify why X and Y are unusual together due to some reason Z. For example, "Thorns are sharp and will cut the brides arms" correctly identifies the reason why thorns and a bride are unusual together. However, if the explanation fails to identify reason Z, it is not acceptable, such as "An old man cannot skateboard." General statements or Wikipedia snippets that fail to explain what makes the image strange are also inadequate, e.g., "Brides usually hold a bouquet of flowers". Moreover, explanations that contain incorrect information are not considered correct. Finally, if an explanation requires verification with a search engine, it should be excluded from the selection process. Overall, the explanation selection process ensures that the chosen explanations accurately capture the reason for an image's weirdness and contribute to the accurate evaluation of model performance.