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

Nitzan Bitton-Guetta[‡*]   **Yonatan Bitton**[†*]   **Jack Hessel**[∓]
Ludwig Schmidt[±]   **Yuval Elovici**[‡]   **Gabriel Stanovsky**[†,∓]   **Roy Schwartz**[†]

[†]The Hebrew University of Jerusalem   [‡]Ben Gurion University of the Negev
[∓]Allen Institute for Artificial Intelligence   [±]University of Washington
{yonatan.bitton,gabriel.stanovsky,roy.schwartz1}@mail.huji.ac.il;
nitzangu,elovici@bgu.ac.il; jackh@allenai.org; schmidt@cs.washington.edu

Abstract

Weird, unusual, and uncanny images pique the curiosity of observers because they challenge commonsense. For example, an image released during the 2022 world cup depicts the famous soccer stars Lionel Messi and Cristiano Ronaldo playing chess, which playfully violates our expectation that their competition should occur on the football field.[1] Humans can easily recognize and interpret these unconventional images, but can AI models do the same? We introduce WHOOPS!*, a new dataset and benchmark for visual commonsense. The dataset is comprised of purposefully commonsense-defying images created by designers using publicly-available image generation tools like Midjourney. We consider several tasks posed over the dataset. In addition to image captioning, cross-modal matching, and visual question answering, we introduce a difficult explanation generation task, where models must identify and explain why a given image is unusual. Our results show that state-of-the-art models such as GPT3 and BLIP2 still lag behind human performance on* WHOOPS!*. We hope our dataset will inspire the development of AI models with stronger visual commonsense reasoning abilities.* [2]

1. Introduction

Upon viewing an unusual image, humans can readily recognize odd, unusual, and incongruent factors. Consider the examples in Fig. 1: smartphones did not exist when Einstein was alive (left), and an oxygen-starved candle would not stay lit for long in a sealed bottle (right). While the images consist of "normal" constituent objects, *compositions* make them unusual. Although it's relatively easy

---

[*]Equal contribution.

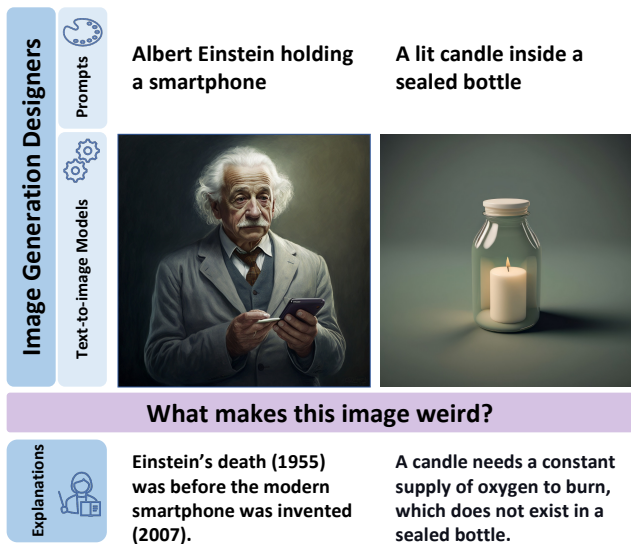[2]Data, models and code are available at the project website: `whoops-benchmark.github.io/`.

Figure 1: We introduce WHOOPS!: a dataset of commonsense-violating images. Designers create interesting, unusual images using prompt-based image-generation tools like Midjourney. We pose several tasks over WHOOPS!, including an explanation generation task. While humans easily identify the weird elements in each image, we show that state-of-the-art AI models struggle.

for humans to identify/explain why an image is unusual, the multi-step reasoning is sophisticated. Connecting visual cues to knowledge about the world goes beyond object recognition, and requires commonsense derived from everyday experiences, physical/social knowledge, and cultural norms [35, 44, 20, 36].

In this work, we introduce WHOOPS!,[3] a dataset of 500 synthetic images and 10,874 annotations designed to chal-

---

[3]*Weird and Heterogene**O**us **O**bjects, **P**henomena, and **S**ituations.*
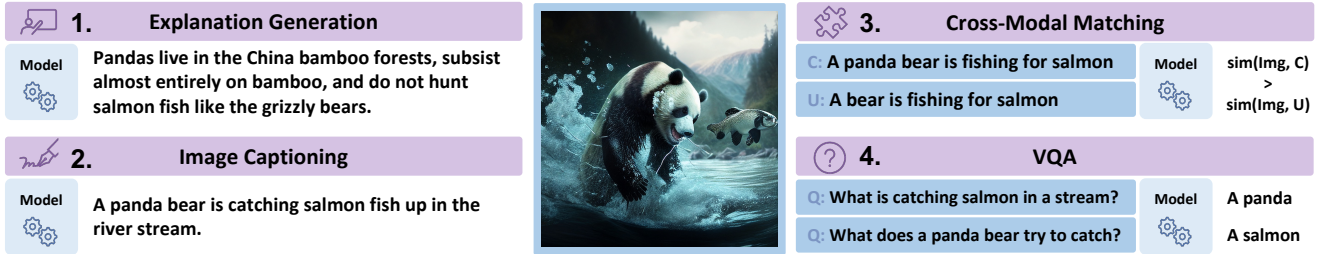
Figure 2: The WHOOPS! benchmark includes four tasks: 1. generating a detailed explanation for what makes the image weird, 2. generating a literal caption, 3. distinguishing between detailed and underspecified captions, and 4. answering questions that test compositional understanding. Inputs to the models are indicated in dark blue.

lenge AI models' ability to reason about commonsense and compositionality. To construct WHOOPS!, we collaborate with designers who use text-to-image models such as Midjourney, DALL-E [31] and Stable-Diffusion [32] to generate images that would be challenging (or even impossible) to collect otherwise. First, prompts that contain two plausibly co-occurring elements are constructed, and then, a modification to one of them is made to create an implausible combination that violates commonsense. Fig. 1 (left), for example, was created by our designers thinking of a plausible scene of Albert Einstein holding a notebook, and then replacing the notebook with a smartphone, which did not exist at the time. We annotate our images with textual information, including both descriptive captions and explanations for what makes each image weird.

Next, we pose four visual commonsense reasoning tasks over the WHOOPS! corpus: (1) explanation generation, where models provide detailed explanations of what makes an image weird; (2) image captioning, where models summarize the content of the images; (3) cross-modal matching, where models should score a detailed caption higher than a correct but underspecified one, and (4) visual question answering, where models answer questions that test their comprehension of the weird images (Fig. 2). Our evaluation covers both zero-shot and supervised experimental settings.

Experiments on WHOOPS! show that state-of-the-art vision-and-language models (e.g., OFA [41], BLIP [26], CoCa [43]) lag behind human performance for all tasks. For instance, a human evaluation reveals that a fine-tuned version of BLIP2-XXL [25] achieves a performance of 27% acceptability, and a "pipeline" approach of feeding a predicted image description to the latest version of GPT3 (davinci-003) [9] reaches 33%. However, both these models fail to generate explanations as well as humans, who achieve 95% on the same task.

To support fully automated evaluations, we present a model-based metric for the explanation-of-violation task. This involves a GPT4 model on ground-truth explanation and predicted explanation. Achieving an accuracy of over

81%, this metric aligns well with human ratings. We make human annotated data and the complete automatic evaluation code publicly accessible. Researchers can evaluate their models and submit the results to the leaderboard on the project website.

Finally, we show that the difficulty WHOOPS! goes beyond recognition; even providing a ground-truth oracle image description instead of the predicted caption in the "pipelined" setting, models still struggle to effectively explain the incongruity of the scene, with an accuracy rate of only 68%. Overall, our results show that WHOOPS! is a challenging benchmark, even for state-of-the-art vision-and-language models. This result highlights the need for continued development in commonsense reasoning, compositionality, and explanation generation. We release our models, code, and data.

2. Collecting *Weird* Images

WHOOPS! is designed to challenge vision-and-language models with images that require commonsense reasoning and understanding beyond simple object co-occurrence. The term "weird" is ultimately subjective, ambiguous, and culture-dependent. Because our goal is to create a benchmark, we aim to generate images that are unusual for a diverse set of reasons, including temporal, biological, cultural, physical and others. We start by describing how we generate the images, and then present an analysis of the different reasons for the images, which shows that our dataset is indeed diverse in this respect.

2.1. Human Generated Synthetic Images via Text-to-Image Models

We recruit a group of 30 image designers who using Midjourney, DALL-E [31], or Stable-Diffusion [32] as text-to-image models. They are requested to generate weird images by first coming up with "weird" prompts, and editing them until a desired image is generated.

These prompts should adhere to the following guideline: first generate a prompt of an image that depicts two ele-
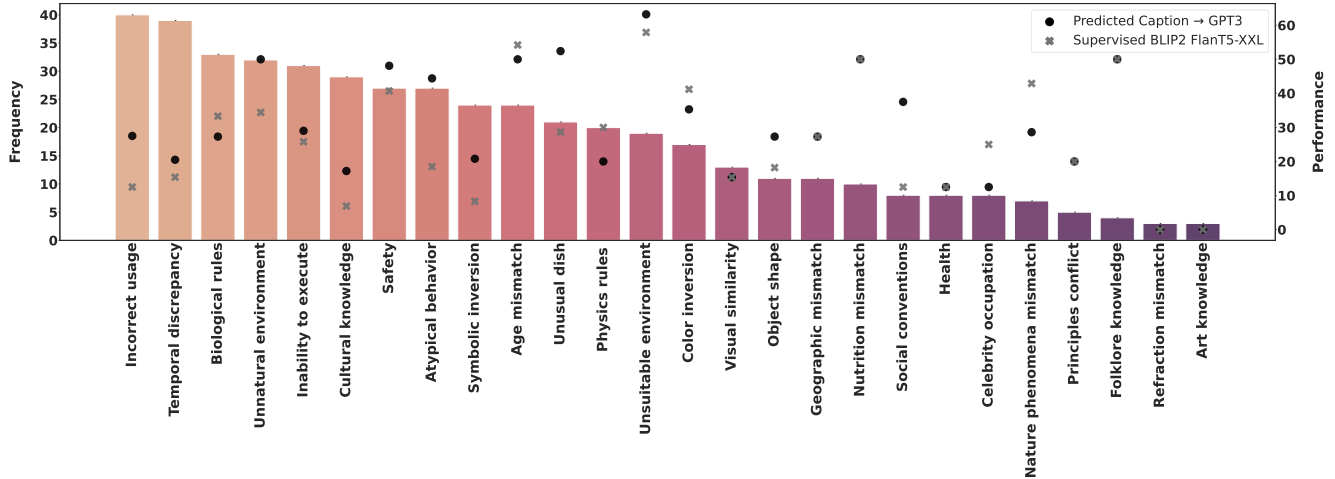
Figure 3: A histogram of the annotated commonsense reasons for WHOOPS! images to be weird. The reasons include a wide range of deviations from expected social norms and everyday knowledge. We also present the explanation generation performance of the two top models in our experiments (Section 5). Left axis is the frequency for each commonsense category, and right is the performance of both models.
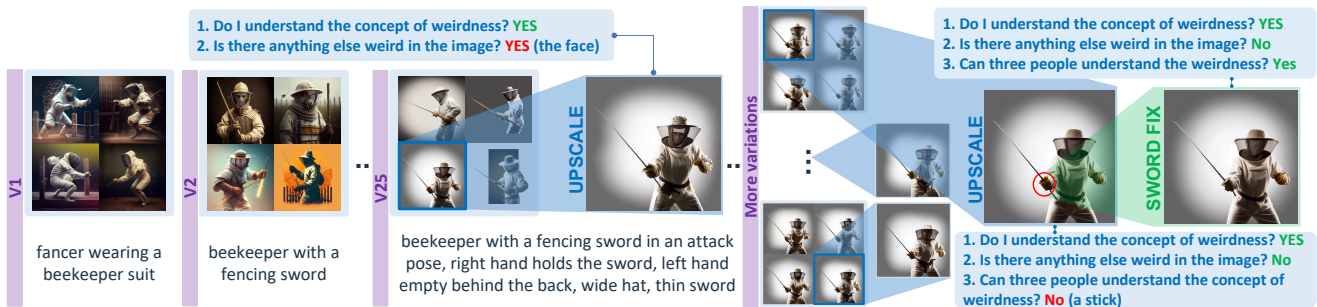

Figure 4: WHOOPS! image generation example: This image, produced through more than 25 iterations, demonstrates the process from initial prompt to finalized 'weird' image produced by the text-to-image model. In each iteration the designer verifies (1) their understanding of the 'weirdness' concept, (2) absence of any additional weird elements in the image, and (3) the 'weirdness' clarify to three independent individuals. Only images meeting these criteria are included in the dataset.

ments that are *likely* to co-occur, and then replace one of them with a different element to create a new prompt that describes an image that is *unlikely* to exist in reality. For instance, taking a prompt of Albert Einstein holding a notebook and replacing the notebook with a smartphone, resulting in a prompt for an unlikely image, as smartphones did not exist during Einstein's time. Each image is required to be synthetic, rather than edited from existing images. This results in a total of 500 weird images. See Appendix A.3 for full guidelines and examples.

The generation of each image, as depicted in Fig. 4, goes beyond a simple use of a text-to-image model. It is a meticulous process, managed by expert designers, involving around 25 iterations per image. The primary objectives are to ensure the image clearly portrays its 'weirdness' concept and to eliminate any extraneous elements that could be

misunderstood as the main source of 'weirdness.'

For every image, the designers also provide a concise one-sentence explanation that encapsulates the unique reason for its 'weirdness.' The authors review and refine these explanations for factual accuracy and specificity, incorporating relevant details such as names, dates, and other pertinent information.

This process aims to create images with unmistakable 'weirdness,' understandable to a wide audience. To verify this, each image is presented to a small control group before its inclusion in the dataset. Images that fail to pass this test are returned to the designers for refinement or to explore alternative concepts. To mitigate the concern that weirdness is limited to a specific culture or region, the designers who created the images and the annotators who labeled the data (Section 4) are from different countries and continents.

2.2. Commonsense Categorization of *Weird* Images

Fig. 3 provides a histogram of the different types of commonsense reasoning that underlie the weirdness of images in WHOOPS!. To create it, we manually annotate each image with the main reason that contributes to its overall sense of "weirdness". Our annotation includes 26 different categories. The reasons cover a broad range of domains, including but not limited to temporal discrepancy (Fig. 1 left), physical rules (Fig. 1 right), nutrition mismatch (Fig. 2), unsuitable environment (Fig. 5), atypical activity (Fig. 6), symbolic inversion, folklore knowledge and more. This analysis shows the diversity and complexity of the reasoning skills that vision-and-language models must possess in order to perform well on our benchmark. Further elaboration on commonsense categories is available in Appendix A.4, including examples.

3. Related Work

The field of commonsense reasoning has recently gained significant attention, with various tasks proposed both in natural language processing (NLP) [33, 45, 46, 34, 4, 16] and computer vision [40, 8]. In the field of vision-and-language, models are being developed to solve complex visual reasoning tasks. These include visual understanding tasks, such as VCR [44], as well as tasks that evaluate commonsense reasoning in association and analogy tasks, like WinoGAViL [5] and VASR [8]. Other tasks evaluate compositionality (e.g., Winoground; [37]), visual abductive reasoning (e.g., Sherlock; [19]) and comprehension and explanation of multi-modal humor [20]. Recent progress in large language models is making way for models that can solve these tasks using instructions like BLIP2 [25] and in-context learning, or zero-shot learning, like Flamingo [2] and MLLM [22]. These recent advances pave the way for our work, which provides a challenging resource for commonsense and compositionality.

WHOOPS!, is distinct from prior work that focuses on reasoning with pre-existing images. Instead, it contains synthetic images that are specifically designed to challenge AI models' abilities to reason about commonsense and compositionality, with an emphasis on images that violate expectations. Our approach uses image generation models to create unique and complex images that would be difficult or impossible to obtain otherwise, providing an opportunity to evaluate critical aspects of visual reasoning, including compositionality and commonsense reasoning.

4. Using Weird Images to Create V&L Tasks

We pose four vision-and-language tasks over the WHOOPS! images to form a benchmark. We first consider a novel task—*explanation-of-violation generation*—which evaluates a model's ability to identify the commonsense rule

that an image violates and reason about the relationships between different elements in the image. We then consider three other well-established tasks posed over the corpus: image captioning, cross-modal matching, and VQA. Fig. 2 shows example annotations for the four tasks over a single image. Previous works have shown that these tasks can be prone to relying on language priors [23, 48, 18, 1, 7, 6, 14]. However, the images in WHOOPS! were purposefully designed to include uncommon combinations, which may make it more challenging for models to exploit language priors. We also report the (relatively low) performance of a text-only baseline in Section 5.2. The end result is a challenging test set for evaluating the performance of vision-and-language models on complex reasoning tasks.

To create task instances, we crowdsource annotations for each of the 500 images in our dataset, including captions, various explanations. We pay each annotator 12–15$ per hour for providing a caption and a explanation; annotation details/instructions are available in the Appendix A.5. We also use auto-generation techniques to create VQA data based on these captions [11, 21], and then validate it using human verification. We describe each task and its evaluation below.

4.1. Explanation-of-violation Generation

The task of the explanation generation is to provide a single-sentence detailed explanation of what makes an image weird. The goal is to test a model's ability to identify the commonsense rule that the image violates and reason about the relationships between different elements in an image. For instance, in Fig. 2, the explanation should provide information such as, *"Pandas usually reside in Chinese bamboo forests, eat almost exclusively bamboo, and do not hunt salmon fish like grizzly bears do"*. We break the task down into two components: *identifying* whether an image is weird and *explaining* what makes it weird.

Identifying weird images. We select a subset of 100 weird images and use a similar protocol to the one described in Section 2 to collect the corresponding "normal" images for them (e.g., for the image of Einstein holding a smartphone, generate an image of him holding a notebook). This task is evaluated using binary accuracy over this paired set, where random chance is 50%. To assess human performance on this task, we ask three human annotators to classify each image as either "weird" or "normal". We determine the final classification through a majority vote. Human performance is 92%, and 3/3 agreement is achieved in 70% of the cases. These results suggest that, while "weirdness" is subjective, on average, humans readily agree on what is weird and what is not in the context of WHOOPS!.

Explanation-of-violation. We ask annotators to provide a detailed single-sentence explanation of what makes the image strange and include the reason why two elements are unlikely to co-exist in the scene. We collect five explanation per image, a total of 2,500. The metric to evaluate model predictions on this task is human judgment. We compare model generations to references using three crowdworker judgments: full details and examples in Appendix A.5.

4.2. Established Tasks

Image captioning. This task requires generating a single-sentence description of an image that includes both elements whose combination makes the image weird. Unlike the explanation task, the captioning task does not demand any reasoning about incongruities. i.e., for the example in Fig. 2, it suffices to just generate *A panda bear fishing for a salmon in the river*. This identification task, however, could be helpful for the explanation task presented in Section 5. We crowdsource five textual descriptions per image, for a total of 2,500 captions; evaluation is using the standard automatic captioning metrics CIDEr [39] and BLEU-4 [29] compared to crowd-authored references.

Cross-modal matching. In this task, a model is given an image and a set of captions, all of which accurately describe the scene, but some of which leave out important details. The evaluation setup challenges models to rank the detailed captions more highly than the underspecified ones. This task tests the model's ability to match the correct caption to the image and overcome its language priors, e.g., a text-only model may rate "A panda hunting for salmon" less likely than "A bear hunting for fish".[4] Performance is measured as the proportion of correct rankings. We collect 500 underspecified captions per image, a total of 2,500 captions.

Visual question answering (VQA). To create question-answer pairs for WHOOPS!, we follow the Q^2 pipeline for automatic VQA generation [11, 21]. This process: 1) derives answers from captions; 2) uses a question generation model to generate a question for each answer; 3) filter the generated questions with the Q^2 NLI model. Fig. 5 presents generated questions and answers for each of the answer candidates in the image caption. We then [21] to ensure that the questions are answerable. We filter out instances solvable by a text-only model performs well so that models must focus on visual-textual interactions. Specifically, we use a language model, FlanT5 XL [12], to answer the questions and filter out instances where the BEM metric is above 0.1. This filtering removes approximately 30% of the questions

Figure 5: We obtain five caption from human annotations, for example: "Two walruses are swimming in the jungle". We then automatically generate question-answering pairs:
(1) Where are the two walruses swimming? *in the jungle*
(2) How many walruses are swimming in the jungle? *two*
(3) What is swimming in the jungle? *two walruses*

and results in 3,374 VQA samples. In 5.2, we show a text-only finetuned model performs poorly on the resulting set. We evaluate using two metrics: (1) strict exact match; and (2) BERT Matching (BEM) [10], which approximates a reference answer to a candidate answer given a question [15] using a language model score.

We manually verify a sample 300 (image, question, answer) triplet by asking three crowdworkers to classify whether the answer is correct. For a baseline for human verification, we mix in randomly sampled 25% of the "negative" answers. The majority vote is selected as the final answer. Humans reach full agreement in 94% of the cases, and the majority vote agrees with the automatic VQA label in 97% of the cases, which provides strong evidence that the generation process generates high-quality QA instances.

4.3. Toxic Content Filtering

Finally, we take two steps to filter toxic and harmful images. First, four of the paper authors manually verify all images and remove those that could be potentially offensive for some groups. Second, we use the Perspectives API[5] to detect and filter out toxic language from our annotated data. We find that the vast majority of our data is non-toxic. Only a very small percentage (0.4%, 0.1%, and 0.4% for captions, explanations, and underspecified captions, respectively) contains toxic language, which we have removed and replaced with new data.

5. Experiments

We evaluate models on our tasks in a fully zero-shot setting and a 5-fold cross-validation supervised configuration.

---

[4]We confirm this point with a FlanT5 XL language model [12] by asking it to determine which caption is more likely, and it rates the underspecified one as more likely in 85% of cases.

[5]https://www.perspectiveapi.com/

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	-	-

Table 1: Test results for Explanation-of-violation encompass two main tasks: *identifying* unusual images and *explaining* their anomalies. While humans consistently outperform models across tasks, providing an oracle image description narrows the performance gap. The *explaining* subtask incorporates metrics from human evaluations and GPT4 ratings. These metrics quantify the fraction of correctly classified explanations, either by human judgment or the GPT4 model, with the latter's accuracy detailed in the GPT4 Rating Accuracy column. Models without human evaluation, namely InstructBLIP, GPT4, Llama-2, mPLUG-Owl, LLaVA, were added on August 9, 2023, with the executing of GPT4 auto-evaluation.

		Image Captioning		VQA		Matching
		B-4 (↑)	CIDEr (↑)	ExactM (↑)	BEM (↑)	Specificity (↑)
Zero-shot	CLIP ViT-L/14	–	–	–	–	70
	OFA Large	0	0	8	38	
	CoCa ViT-L-14 MSCOCO	25	102	–	–	72
	BLIP Large	13	65	6	39	77
	BLIP2 FlanT5-XXL	31	120	15	55	71
Fine-tuned	BLIP2 FlanT5-XL	41	174	20	55	81
	BLIP2 FlanT5-XXL	**42**	**177**	**21**	**57**	84
Text only FT	BLIP2 FlanT5-XXL	1	2	4	24	**94**

Table 2: Test results for image captioning, cross-modal matching and visual question answering. A fine-tuned version of BLIP2 FlanT5-XXL generally performs best but there's significant headroom.

For zero-shot evaluations, we use the officially published implementations of CLIP ViT L/14 [30], OFA Large [41], BLIP Large [26], CoCa ViT-L-14 MSCOCO [43], and BLIP2 FlanT5-XXL [25]. Additional details can be found in Appendix A.1. Some models can be used to tackle all tasks (BLIP2), and some only a subset of the tasks: OFA for image captioning and VQA; CoCa for image captioning and cross-modal matching; and CLIP only for cross-modal matching. For CLIP and CoCa, we evaluate all available model versions (four for each model), and for readability report the best performing ones.

For supervised evaluations, we fine-tune the BLIP2. To report over the same instances as in the zero-shot evaluations, we split the images in WHOOPS! into 5 cross-validation splits. For these 5 splits independently, we train

Figure 6: We explore two approaches for explanation-of-violation generation. The first approach involved using an end-to-end model that receives an image as input and generates an explanation as output, evaluating both a zero-shot and fine-tuned versions of the BLIP2 model. The second approach was a pipeline that predicted a caption for the image, or used a ground-truth caption (oracle), and then used a language reasoning model (GPT3) to generate an explanation based on the caption.

supervised models using 60% of the data as training, 20% as validation, and 20% for test. We fine-tune just the Q-former parameters of BLIP2 using Adam [24]. We train for 15 epochs, and use the validation set for early stopping and to select learning rate between {1e-5, 5e-5}. We concatenate training instances in a sequence-to-sequence format for all tasks jointly such that a single supervised model can address all tasks; see Appendix A.2 for details.

We also consider "pipelined," methods [47] for the explanation-of-violation tasks. These methods decouple recognition of objects from reasoning about incongruity. In the "pipelined" approach, an image caption is passed to a large language model (LLM), which is then tasked with the two explanation-of-violation subtasks. We use GPT3 text-davinci-003 as the LLM [9] and experiment with two textual descriptions: a predicted image caption by the BLIP2 model and an oracle version, which includes the ground-truth caption collected by annotators and verified by the authors.

As a baseline, we train a text-only BLIP2 FlanT5-XXL using the same cross-validation/hyperparameter setup as for the full supervised models, except we set all pixels of the image to mean so that image content cannot be used at training or testing time.

Addition of new models for explanation-of-Generation task. on August 9, 2023, we expanded our zero-shot eval-

uations by incorporating new models. For the "pipelined" models, we integrated Llama2 [38] and GPT4 [28]. Additionally, for the end-to-end vision-language models, we introduced InstructBLIP [13], LLaVA [27] and mPlug-Owl [42].

5.1. Automatic Evaluation for Explanation-of-violation

In the explanation-of-violation task, we supplement human judgments with an automatic evaluation metric, demonstrating its significant alignment with human assessments. The objective of introducing this automatic evaluation method is to provide a reproducible standard for result reporting on WHOOPS!.

We utilize the human-annotated data as described in Section 4 and shown in Fig. 8. This data comprises 5,000 pairs of ground-truth explanations, candidate explanations, and an associated label indicating whether humans rated the given explanation as correct or not.

We report two results: (1) The proportion of positive and negative labels. This forms the *model rating* that can be compared with the *human rating*. (2) The accuracy of the automatic evaluation metric as compared to the human judgement. This helps in understanding the alignment between human and machine-based assessments.

We used a GPT4 model with the prompt: *Evaluate the equivalence of the following explanations for the question "What is unusual in this image?" Answer with True or False: A: $sentence_A$ B: $sentence_B$. True if A and B have the same meaning, False if they do not.* The parameters $sentence_A$, $sentence_B$ contains the ground truth explanation and the candidate explanation respectively. [6]

5.2. Results

Explanation-of-violation. The results for the two identification and explanation subtasks, are presented in Table 1. For both cases, models significantly lag behind human performance. For example, on identification, the best end-to-end fine-tuned BLIP2 FlanT5-XXL model achieves at best 73%. For explanation, even the oracle model (which is given access to a ground-truth, human-authored description of the image) only achieves a performance of 68%, falling substantially short of human performance (95%). These results indicate that our dataset provides a challenging benchmark for the development of next-generation vision-and-language models. We provide an example of model predictions in Fig. 6 and an example of the evaluation task to rate both model predictions and human explanations in Fig. 8.

In the automated explanation of violation task, the maximum deviation between the automatic rating and human rating is 9% for the Predicted Caption → GPT3 task. All

---

[6]We publish the annotated data and code for automatic evaluation on WHOOPS! in this notebook.

Weird	Normal	Natural	% Proportion
V	V	V	45
X	V	V	40
X	X	X	6
X	V	X	3
V	X	V	2
V	V	X	2
X	X	V	2
V	X	X	0

Table 3: Analysis of caption errors, human rate of correct caption: 40% of the errors are "commonsense errors" where the incorrect caption is for the "weird" image only. Only 3% are "naturalness errors" where the natural image caption is better than the synthetic image captions. 5% of the cases had synthetic image captions better than natural ones.

the automated explanation-of-violation models achieve an accuracy higher than 81%. These results suggest that automatic evaluation yields ratings correlated to human ratings.

Captioning, VQA, + Matching. The results are presented in Table 2. The zero-shot results highlight the strengths and weaknesses of each model. OFA achieves the lowest results, particularly in image captioning, where it frequently predicts the pattern "digital art selected for the #". Zero-shot BLIP2 demonstrates a substantial improvement over the other models. But even the supervised models have significant room for improvement, especially in VQA (maximum BEM score is 57%) and image captioning; In section 6, we conducted an analysis in which humans rate the BLIP2 zero-shot predictions. Despite the relatively high CIDEr score, the model failed to capture important information, resulting in a human rating of 49%. We also report results for a text-only supervised baseline. The results show that it performs poorly on captioning and VQA. [7]
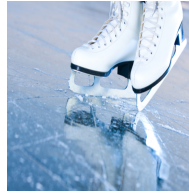
6. Analysis

In this section, we analyze if the challenges in WHOOPS! come from the syntheticity or weirdness of images and evaluate the performance on different commonsense categories.
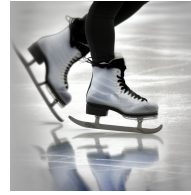
6.1. Main Challenge: Weirdness, not Synthesis

To discern whether the difficulties models face in WHOOPS! arise from the images being "weird" or syn-

---

(a) a pair of white ice skates on an ice rink

(b) a close up of a person's skates on an ice rink

(c) a person is skating on an ice rink

(1) Wrong caption only for the *weird* image (caption *c*). The flooring is made of wooden parquet, and not an ice rink.

(d) an old abandoned house in the middle of a field with a lightning bolt

(e) a house in the middle of a field

(f) an old house in the desert with a lightning bolt

(2) Wrong captions for both synthetic images, weird and normal (captions *e* and *f*). The middle one misses the lightning, the right one misses the clear sky.

(g) a man in a top hat playing an electric guitar

(h) slash plays a custom electric guitar in smoke

(i) slash playing a saxophone in a band

(3) Wrong caption only for the natural image (caption *g*). The left caption misses the famous guitarist name (*Slash*)

Figure 7: Examples of caption errors by the BLIP2 model. The images from left to right are the *weird* (synthetic) images, *normal* (synthetic, without weirdness) and *natural*.

thetic, we collect a set of "normal" and "natural" images. The "normal" images are created by replacing the unconventional element with a conventional element, resulting in minimal changes between the pairs of $(normal, weird)$ images, following the idea of contrast sets [17, 7]. To obtain the "natural" images, we search for similar non-synthetic images using Clip Retrieval [3], which finds close images through CLIP [30] embedding similarity. High quality images with an "aesthetic score" above 7 are chosen, and the
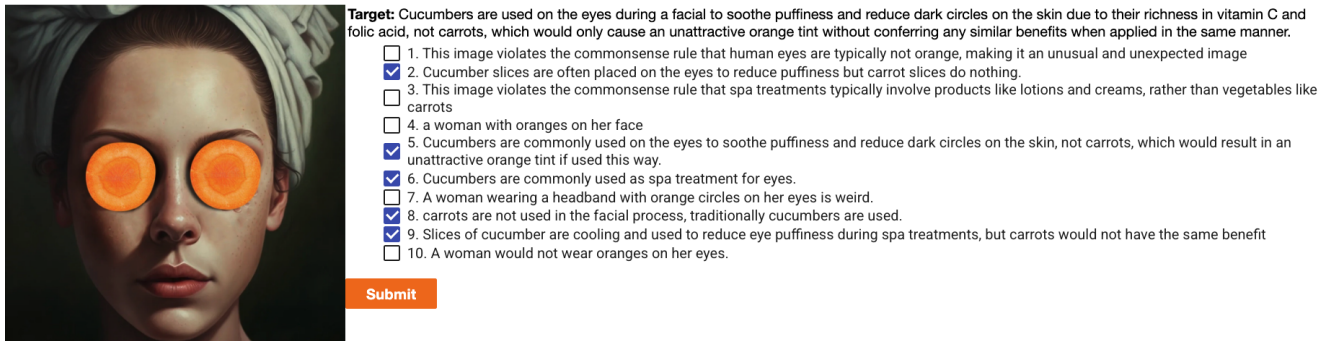
Figure 8: Amazon Mechanical Turk user interface for the task of explanation selection. The annotators receive an image and number of explanations, both human curated and models predictions, and need to mark the correct explanations.

top similar images are selected. Fig. 7 shows examples of the collected images, including an image of ice skates in a "natural" photograph, and two synthetic images, one on a conventional element (an ice rink) and the other on an unconventional surface (a basketball parquet).

Next, we use BLIP2 to generate captions for 300 images in three categories: natural, normal, and weird. Human annotators evaluate the accuracy of the captions for each image category, and the results are presented in Table 3. The accuracy of image captioning is high for the natural and normal categories (89%), but low for the weird category (49%). We find BLIP2 generates correct captions for all three categories for 45% of the image triplets. For 40% of the triplets, incorrect captions are generated only for the weird images (Fig. 7 (a)). In 5% of the cases, synthetic images have better captions than natural ones, while only 3% of cases have errors related to naturalness, with incorrect captions for both synthetic images (Fig. 7 (b)). These results suggest that the BLIP2 model can generate captions for synthetic images as well as natural images with high accuracy, but the primary challenge lies in commonsense reasoning, underscoring the need for improvement in state-of-the-art models. The full experiment results are available in the project website.

6.2. Performance by Commonsense Categories

In Fig. 3, we included the performance of the top two models, demonstrating that WHOOPS! provides insights into their strengths and weaknesses. Specifically, we observe that the Predicted Caption → GPT3 pipeline approach outperforms the Supervised BLIP2-XXL end-to-end model in 46% of the categories, such as in cases of Incorrect usage (e.g., A bowl of ice cream is inside the microwave), and performs worse in 23% of the categories, such as in Biological rules (e.g., A mouse hatches from an egg), and similarly in 31% of the categories. Notably, both models perform poorly in identifying temporal discrepancy (e.g., women in ornate Renaissance clothing take a selfie with a smartphone) and art knowledge (e.g., The Girl with a Pearl Earring wears

a golden hoop earring), while performing well in identifying an Unsuitable Environment (e.g., A snowman sits on the beach on a sunny day).

7. Conclusions

We introduced WHOOPS!, a novel dataset of synthetic images challenging AI models to reason about commonsense and compositionality. Using text-to-image models, we generated difficult or impossible to obtain images and annotated them with explanations, captions, underspecified captions, and visual question answering pairs. We proposed a benchmark of four challenging tasks and evaluated state-of-the-art models, which struggled, especially in the new task of explanation generation, where a significant gap between human and model performance remains. Our dataset and benchmark tasks are a valuable resource for advancing research in these areas. We provide an evaluation code and a leaderboard for methodical tracking and replication of results across our four benchmark tasks.

8. Limitations

We took measures to filter out potentially harmful or offensive images and texts in WHOOPS! (Section 4.3), but it is still possible that some individuals may find certain content objectionable. Any harmful cases can be reported in the project website and removed from the dataset.

While WHOOPS! has fewer images than other benchmarks, we intentionally created unique and challenging images to provide diverse commonsense challenges. The smaller size allowed for efficient manual annotation and evaluation, ensuring data quality and reliability. We plan to expand the dataset in the future to enhance its usefulness.

We have made significant efforts to develop reliable and advanced models for this task, our focus is not on achieving the ultimate upper bound on model performance, but on providing a challenging resource for commonsense and compositionality using image generation models.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9687–9695. IEEE, 2020.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[3] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022.

[4] Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press, 2020.

[5] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. WinoGAViL: Gamified association benchmark to challenge vision-and-language models. *arXiv preprint arXiv:2207.12576*, 2022.

[6] Yonatan Bitton, Gabriel Stanovsky, Michael Elhadad, and Roy Schwartz. Data efficient masked language modeling for vision and language. *arXiv preprint arXiv:2109.02040*, 2021.

[7] Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. *arXiv preprint arXiv:2103.09591*, 2021.

[8] Yonatan Bitton, Ron Yosef, Eli Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. VASR: Visual analogies of situation recognition. *arXiv preprint arXiv:2212.04542*, 2022.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022.

[11] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*, 2022.

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[14] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *arXiv preprint arXiv:2104.03149*, 2021.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *ArXiv preprint*, abs/1908.02899, 2019.

[17] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[19] Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 558–575. Springer, 2022.

[20] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

[21] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021.

[22] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models, 2023.

[23] Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[28] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[33] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[34] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press, 2019.

[35] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[36] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. Com-

monsenseqa 2.0: Exposing the limits of ai through gamification. *ArXiv preprint*, abs/2201.05320, 2022.

[37] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *ArXiv preprint*, abs/2204.03162, 2022.

[38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[40] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2542–2550. IEEE Computer Society, 2015.

[41] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[42] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[44] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE, 2019.

[45] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, 2018. Association for Computational Linguistics.

[46] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics.

[47] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo,

Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[48] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5014–5022. IEEE Computer Society, 2016.

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.
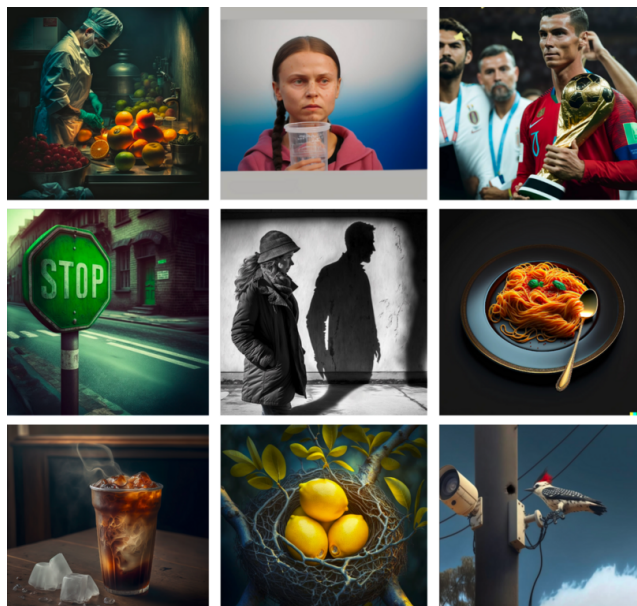

Figure 9: Designer Guidelines Examples.

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)

---

[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.

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)

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.
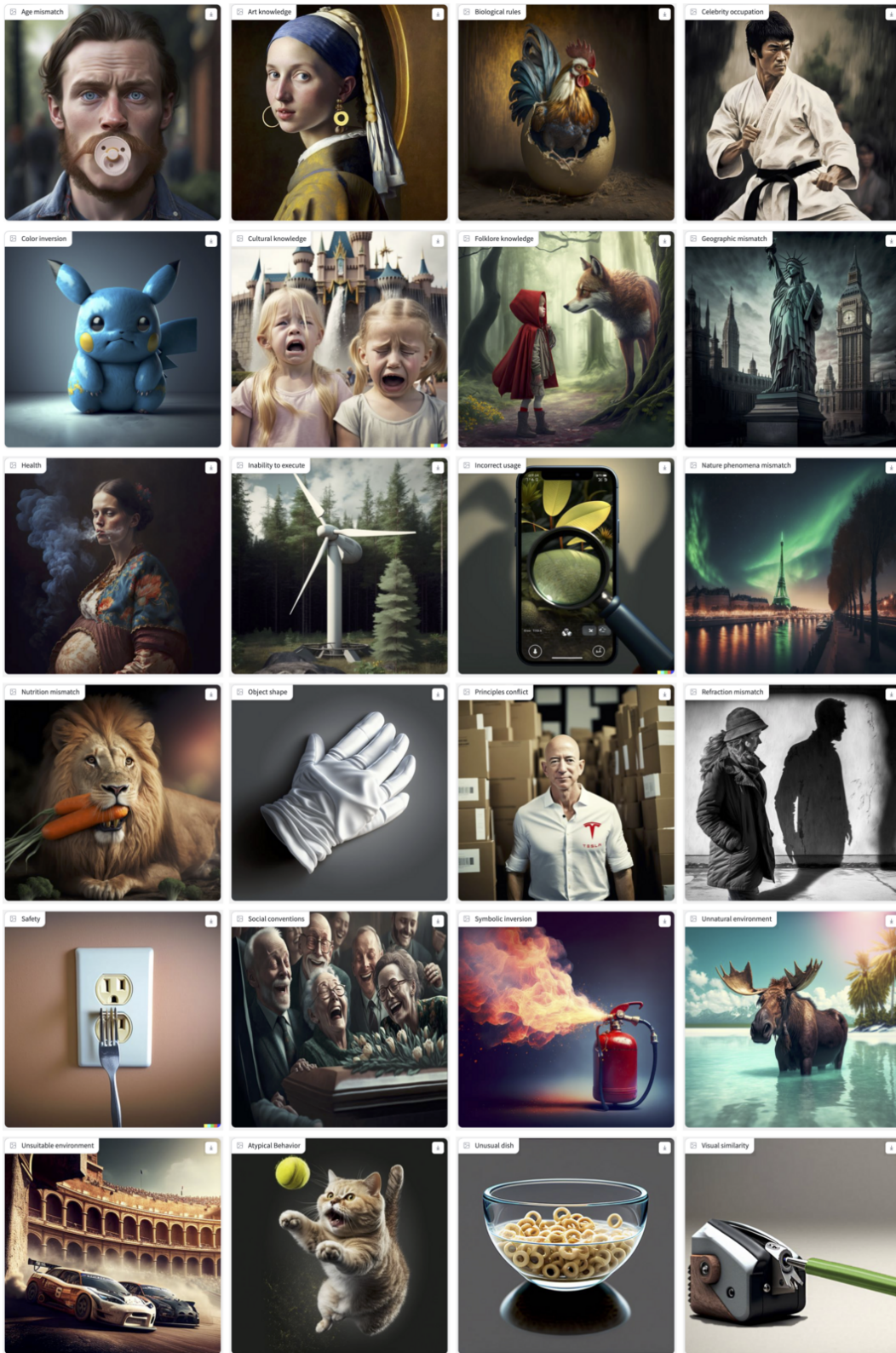
Figure 10: The WHOOPS! images span across various categories and are intended to test AI models in multiple areas of common sense.
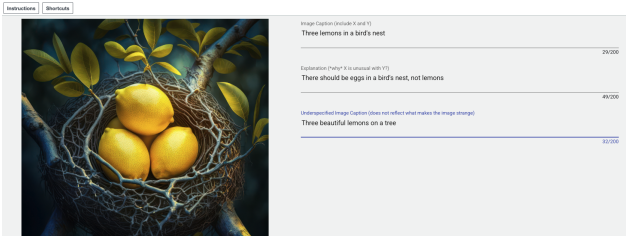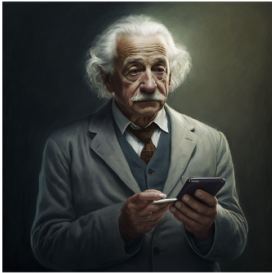
Figure 11: Amazon Mechanical Turk Annotators user interface. The annotators receive an image and provide three types of annotations.

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

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.
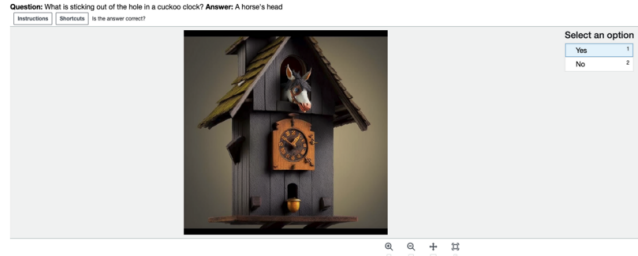

Figure 13: Amazon Mechanical Turk interface for the task of VQA verification. The annotators receive an (image, question, answer) triplet, and need to determine whether the answer is correct or not.

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 selections, 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.