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
‡ 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 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-

---

1. Equation contribution.
2. Data, models and code are available at the project website: whoops-benchmark.github.io.
3. Weird and Heterogeneous Objects, Phenomena, and Situations.
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.

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—explained-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 an explanation; annotation details/instructions are available in 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 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 API5 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.

For zero-shot evaluations, we use the officially published implementations of CLIP ViT L/14 [30].

4We 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.

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

<table>
<thead>
<tr>
<th>Task</th>
<th>Identify</th>
<th>Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary Accuracy (↑)</td>
<td>Human Rating (↑)</td>
</tr>
<tr>
<td>End-to-end</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLIP2 FlanT5-XXL (Zero-shot)</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>BLIP2 FlanT5-XL (Fine-tuned)</td>
<td>60</td>
<td>15</td>
</tr>
<tr>
<td>BLIP2 FlanT5-XXL (Fine-tuned)</td>
<td>73</td>
<td>27</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>InstructBLIP</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>mPLUG-Owl</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LLaVA</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pipeline (Zero-shot)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted Caption → GPT3</td>
<td>59</td>
<td>33</td>
</tr>
<tr>
<td>Ground-truth Caption → GPT3 (Oracle)</td>
<td>74</td>
<td>68</td>
</tr>
<tr>
<td>Predicted Caption → GPT4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Predicted Caption → Llama-2-7b</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Predicted Caption → Llama-2-13b</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ground-truth Caption → GPT4 (Oracle)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ground-truth Caption → Llama-2-7b (Oracle)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ground-truth Caption → Llama-2-13b (Oracle)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Humans</td>
<td>92</td>
<td>95</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Captioning</th>
<th>VQA</th>
<th>Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-4 (↑)</td>
<td>CIDEr (↑)</td>
<td>ExactM (↑)</td>
</tr>
<tr>
<td>Zero-shot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP ViT-L/14</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OFA Large</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>CoCa ViT-L-14 MSCOCO</td>
<td>25</td>
<td>102</td>
<td>-</td>
</tr>
<tr>
<td>BLIP Large</td>
<td>13</td>
<td>65</td>
<td>6</td>
</tr>
<tr>
<td>BLIP2 FlanT5-XXL</td>
<td>31</td>
<td>120</td>
<td>15</td>
</tr>
<tr>
<td>Fine-tuned</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLIP2 FlanT5-XL</td>
<td>41</td>
<td>174</td>
<td>20</td>
</tr>
<tr>
<td>BLIP2 FlanT5-XXL</td>
<td><strong>42</strong></td>
<td><strong>177</strong></td>
<td><strong>21</strong></td>
</tr>
<tr>
<td>Text only FT</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

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 WHOPS! into 5 cross-validation splits. For these 5 splits independently, we train supervised models using 60% of the data as training, 20% as validation, and 20% for test. We fine-tune just the Q-
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.

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 language model (GPT3), which is then tasked with reasoning about incongruity. In the “end-to-end” approach, the image is passed directly to the language model. 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 evaluations 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.

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 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 synthetic, we collect a set of “normal” and “natural” images. Our question collection process filters out questions that could be answered using text alone in a zero-shot setting (see Section 4.2). The current analysis validates that this filter indeed removed shortcuts, even for supervised models. In contrast, the text-only model performs well in the matching task with a performance of 94%; likely, the model learns to prefer more detailed captions, even without seeing the image. We thus advocate for matching to be used only as a zero-shot evaluation.

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.

<table>
<thead>
<tr>
<th>Weird</th>
<th>Normal</th>
<th>Natural</th>
<th>% Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>V V V</td>
<td>V V V</td>
<td>V V X</td>
<td>45</td>
</tr>
<tr>
<td>X V V</td>
<td>X V V</td>
<td>X X X</td>
<td>40</td>
</tr>
<tr>
<td>X X X</td>
<td>X X X</td>
<td>X X V</td>
<td>6</td>
</tr>
<tr>
<td>X V X</td>
<td>X V X</td>
<td>X V X</td>
<td>3</td>
</tr>
<tr>
<td>V X V</td>
<td>V X V</td>
<td>V X V</td>
<td>2</td>
</tr>
<tr>
<td>V V X</td>
<td>V V X</td>
<td>V V X</td>
<td>2</td>
</tr>
<tr>
<td>X X V</td>
<td>X X V</td>
<td>X X V</td>
<td>2</td>
</tr>
<tr>
<td>V X X</td>
<td>V X X</td>
<td>V X X</td>
<td>0</td>
</tr>
</tbody>
</table>

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

(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.

(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.

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 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


[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 Chen, Johan Bjorck, Vishrav Chaudhary, Subhijot Som, Xia Song, and Furu Wei. Language is not all you need: Aligning vision and pattern recognition with language models, 2023.
