Appendix

In this appendix, we present Image Collection and Curation (Appendix A), Key Visual Features of DOCCI (Appendix B), Annotation Details of DOCCI (Appendix C), Object Statistics in DOCCI Images (Appendix D), Error Analysis for Text-to-Image Generation models (Appendix E), and Datasheet for DOCCI (Appendix H).

A Image Collection and Curation

Section 2.1 gives an overview of the motivations, time periods and locations of many of the photographs. Here, Jason Baldridge gives additional detail (in first person) about further background, motivations and choices regarding DOCCI image collection and curation.

Collection. While writing up the Parti paper [64], I put a lot of time and thought into conveying both the process of coming up with great images (growing cherry trees, section 6.2) and the limitations and failure modes of the model (section 6.3). This was before the time of broadly available text-to-image models and I felt this was essential to making sure that a broader view of the model was available than if we only included our favorite cherry picked outputs in the paper and elsewhere. Even though Parti came out in June 2022, we had early exciting models that I had been working with in summer 2021, and I was already obsessed with exploring the boundaries of what they could and could not do.

My son Nash plays competitive junior tennis, and there was a rain delay during one of his matches in August 2021. I was looking at the tennis court, which happened to have a basketball hoop in the back corner. With the rain falling on the court, it occurred to me that there were three interesting elements to the scene – and perhaps by just taking pictures of scenes like that, I could build up a library of interesting settings and controlled variations that we could describe and try to reproduce with our models. I started tentatively, thinking to take a couple hundred pictures. However, it takes time to build out a team and annotation process and this ended up growing substantially, to 15,000+ images. This was fueled in part by the fact that it started as we were coming out of the COVID pandemic lockdown and my family made up for lost travel opportunities, including vacations and many trips for my son's tennis tournaments, plus I had work travel to California and New York. This gave me opportunities to capture animals in the Everglades, the scenes and simulacra of Las Vegas, statues and buildings in NYC, cacti in Arizona, and many more, in addition to places and things all around Texas (my home state).

The nature of my choices was thus fueled in part by where I ended up, the activities I and my family took part in, what I found inherently interesting, and my goal of finding tricky or useful scenes for text-to-image models – combined with the *exclusion* of faces. As such, there are some clear biases in the image collection as a whole, including many images related to tennis, cars (I used to

restored old cheap sports cars in high school), farm and wild animals found in Texas, my own pets, graffiti, views from Google's high rise office building in Austin, and so on. I hope that my own act of taking and releasing images for research purposes will spur other researchers to release similar collections – and thus not only add quantity but also reduce the bias in the available images we collectively have for research with Creative Commons (or similar) licensing.

There are some notable aspects of themes and nature of the images:

- In addition to taking new pictures, I also dug back through images in my iPhotos collection to identify earlier images that were DOCCI-able.
- While the majority of the images were taken of existing moments or scenes, a fair number were set up to explicitly test categories like text rendering, spatial relations, counting, attribute binding and mixed media. Often, I used my kids' toys or random found objects to create images that were generally clean (not a lot of background detail) but had two, three or even four or more distinct objects in precise spatial configurations.
- Despite the exclusion of faces, I tried to find some creative ways to indirectly include people in it, via statues, shadows (e.g. a shadow that points to a specific object or letter) or hands holding things.
- We got both our labradoodle Ivy and my daughter's cat Yoshi during the collection, and they both have pictures from various stages of their development. With their entity annotations, this could provide for some interesting Dreambooth-style [50] explorations of the same entity at different ages.
- I tried to obtain groups of images covering multiple forms of the same basic objects, such as real horses, horse statues, carvings of horses, toys and figurines of horses, and so on, the same kind of car in both toy and real form, or orcas and dolphins in real life and toy form doing similar actions.
- There are many photos taken on US highways (covering regions from Texas to California, to Michigan, to North Carolina, and more). These were taken either while others were driving, or by my wife or son while I was.
- I captured many images of clocks, and annotators were instructed to include the indicated time in their descriptions. We have not directly tested generation of clocks with the correct time in this paper, DOCCI's data can support this precise and easily measured task.
- Traveling to many tennis tournaments meant staying at many hotels and bed and breakfasts, allowing for a diverse range of home and hotel scenes.
- The fact the collection spanned over a year means that aspects of all seasons and holidays in the US (primarily Texas) are represented.

I feel incredibly fortunate that I had the opportunity and means to embark on this photo quest, and work with an amazing team to create DOCCI from them. It also opened my eyes to see the world differently and find new details every place I went. That said, the photographs themselves are generally not beautiful, high quality ones that a trained photographer would have been able to capture; mostly I just snapped something quickly so that it would be describable and useful as a reference for images later generated from those descriptions.

Curation. Throughout the whole period of taking pictures, I kept a rough internal mental model of things that would be novel or interestingly different from those I had already gotten pictures of. After queuing up a set up pictures, I would go through them (often during down times at tennis tournaments) to select which to keep and crop them to reduce visual clutter (to focus on the main reason for having taken a specific image). For cropping, I never changed aspect ratio – only zoomed in and selected a sub-part of the image. Every few months, I transferred the images for annotation (culling many in the process). Finally, when reviewing clusters before annotation, I selected some for deletion if they were not sufficiently distinctive (e.g. images of clouds, caves, fields and such).

DOCCI AAR Images. We capped the collection in September 2022 for the main DOCCI collection. However, I had the opportunity to make further trips, including to Canada and Europe, and ended up taking more to build up a further, more diverse, set of images that could be released for research. In many ways, these are nicer images than the DOCCI core images, because of the subject matter (so many incredible locations and interesting or beautiful objects), my increased use of compositional aspects like rule of thirds, photographic techniques such as bokeh, and the freedom to select the best crop rather than being constrained to only standard landscape and portrait. I also rotated and straightened many of these to improve their perspective and alignment. Though we are doing new work with these, we release them now along with DOCCI rather than holding on to them. We do this so that others who use DOCCI will be able to immediately take advantage of this temporally and spatially displaced set of images that I also took, e.g. for things like iterative caption-and-image generation.

B Key Visual Features

Objects All primary and secondary objects, either animate (e.g., cats and dogs) or inanimate (e.g., statues), that play key roles in the images.

Attributes Each object possesses important attributes, including shape, size, color, material, texture, pose, action, and state.

Spatial Relationships The orientation refers to the locations of objects in the image (e.g., center, top right). The direction indicates the way objects are facing based on the point of view (i.e., the camera view). When multiple objects are in the image, the relative position determine the locations of two or more objects.

Text Alphabets, numbers, and other characters can be found on different surfaces and materials (e.g., paper, sign boards, and concrete walls) in various forms (e.g., print, handwriting, chalk, and carving). In addition, text can be written in different styles (e.g., fonts and colors).

Counts The counts of primary and secondary objects appearing in the image. We focus on numbers up to approximately twenty, as tracking attributes for too many objects becomes difficult.

World Knowledge Objects in the image may be named entities, potentially requiring background knowledge (e.g., One World Trade Center in the NYC cityscape).

Scenes Images could have been taken indoors or outdoors, and either during the daytime or at nighttime.

Views The view types and camera angles define the overall frame. A view type is a combination of the horizontal position (e.g., front, back, side, three-quarter), the vertical position (e.g., bird's-eye, eye-level, worm's-eye), and the depth (e.g., close-up, medium, long).

Optical Effects Lighting is one of the salient features of the image. Shiny outdoor objects can reflect sunlight and cast shadows on the ground during the day. Images may become obscured or less distinct due to weather or lighting conditions.

C Annotation Details

C.1 Annotation Pipeline

Stage 1: Extracting Key Information First, we instruct annotators to extract key visual features and write brief descriptions of them, relating to the aspects listed in to the aspects listed in Appendix B. These descriptions may not always form complete sentences or phrases. Annotators may leave certain aspects blank if they do not find the corresponding information in the images. The goal of this stage is to extract salient information from images quickly.

Stage 2: Writing Descriptions In this stage, annotators are asked to write complete descriptions based on the key information extracted in the previous stage. Annotators can view the image and a brief description to capture the key information directly within the annotation UI, enabling them to concentrate on their writing. The descriptions generated at this stage will serve as the first draft, which will be refined in the next stage.

Stage 3: Elaborating Descriptions The first draft often misses key details in the image; therefore, we conduct the revising stage to address these omissions. Based on the descriptions from the previous stage, we request annotators to create more detailed and elaborated descriptions. Specifically, we ask them to include the key aspects listed in Appendix B. The goal of this phase is to refine the descriptions to be as detailed and specific as possible, ensuring they uniquely correspond to the images they describe.

C.2 Quality Control

Annotation Workflow For all stages, we provided comprehensive annotation guidelines and conducted pilot studies. We then proceeded to the full annotation process once the annotators had become familiar with the tasks. For the

stage 3, annotators who had passed our qualification tests participated in the full annotation process. We grouped images into 149 clusters based on image similarity.[4] We deployed those clusters as batches, maintaining small batch sizes of no more than 200 images, and provided feedback daily. This approach allowed us to provide batch specific guidelines with ease, ensuring that mistakes and misunderstandings were not carried over to later batches. In this stage, we collaborated with US-based annotators, who are familiar with the background knowledge of the DOCCI images, to ensure accurate interpretation and analysis. The curator also provided textual guidance for many images prior to stage 3 to clarify what was depicted in difficult situations (such as dinosaur tracks), to provide specific world knowledge that was either easy to state or which would be hard to verify for annotators on their own, or to provide specific cues about what was interesting about the photo so that their resulting description would reflect the challenge behind the curator's intention in taking the photo.

Images We manually reviewed all images and removed any personally identifiable information (PII), such as people's faces, phone numbers, URLs, and account names of SNS (Social Network Services). Additionally, we ran a safe search detection tool[5] on the images to identify potentially harmful content. 97.6% of images were judged to be unlikely harmful. We manually reviewed the remaining 2.4% of images and confirmed that they are false positives.

Text Descriptions We primarily focused on two types of errors included in the annotated descriptions: **precision** and **recall** errors. Precision governs incorrect information, while recall concerns the omission of information. For example, using a wrong object name will be penalized with precision, and failing to include key attributes will be treated as a recall error. For the precision errors, we investigated the results of a text-image alignment metric such as VQ^2 [62], which a VQA model provides confidence scores to the questions derived from a description. For example, in the statement, "The car in faded baby blue is parked on a field of dry grass," a corresponding question-answer pair would be: Q: "What color is the car?" A: "Faded baby blue." The VQA model then calculates the likelihood of the answer being accurate. Answer pairs with low probabilities indicate potential inaccuracies in the description. To mitigate precision errors, we reviewed descriptions with low confidence scores to ensure the accuracy of the highlighted information. To mitigate the recall errors, we inspect descriptions that fall below the 10 percentile in length. Short and brief descriptions often omit key details, making the length of the description a reliable indicator. Finally, we asked annotators to rewrite the disqualified descriptions.

C.3 Annotator Qualification Tests

Creating detailed descriptions for images requires a variety of skills, including comprehensive knowledge about the subjects depicted in the images and profi-

[4] We used in-house image embeddings to compute similarity.

[5] We used Google Cloud Vision API: `https://cloud.google.com/vision/docs/detecting-safe-search`

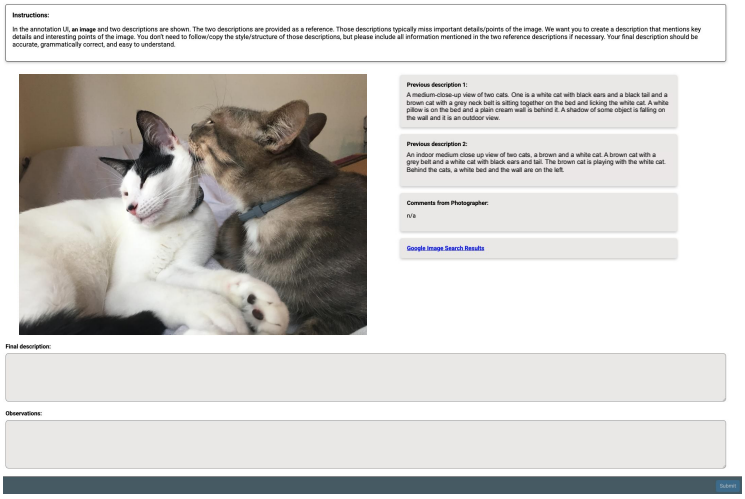


Fig. 9: The annotation UI for Stage 3: Elaborating Descriptions.

cient writing abilities. Given that the majority of these images are captured in the United States, we prefer to assign our annotation tasks to US-based annotators in Stage 3. We initially explain our annotation guidelines and standards to candidates through documents and training sessions. Then, we ask the candidates to annotate ten images and evaluate the quality of their descriptions. Candidates who achieve the minimum score (4 out of 5) are invited to participate in full-scale annotation. Candidates who receive lower scores may retake the qualification test up to three times. Those who fail the exam three times are not allowed to advance to full-scale annotation.

C.4 Annotation UI

Annotating Text Description In Stage 3 of annotation pipeline, we ask annotators to expand upon and refine the descriptions provided in Stage 2. In the user interface (UI), as illustrated in Figure 9, an annotator is shown one image along with two descriptions from Stage 2. We encourage annotators to employ Google Image Search to identify any objects in the images that they might not recognize. Additionally, we instruct annotators to report any personally identifiable information (PII) or inappropriate content they find in the images in the "Observations" box.

Image Elimination In the human evaluation described in Section 3.3, we ask human annotators to identify the correct *pivot* image from a set of four similar images (i.e., distractors), based on the pivot's description. For this, we sample 1k pivots images from the test set (**DOCCI-Test-Pivots**). Then, we collect other images as distractor candidates from the test set, based on their

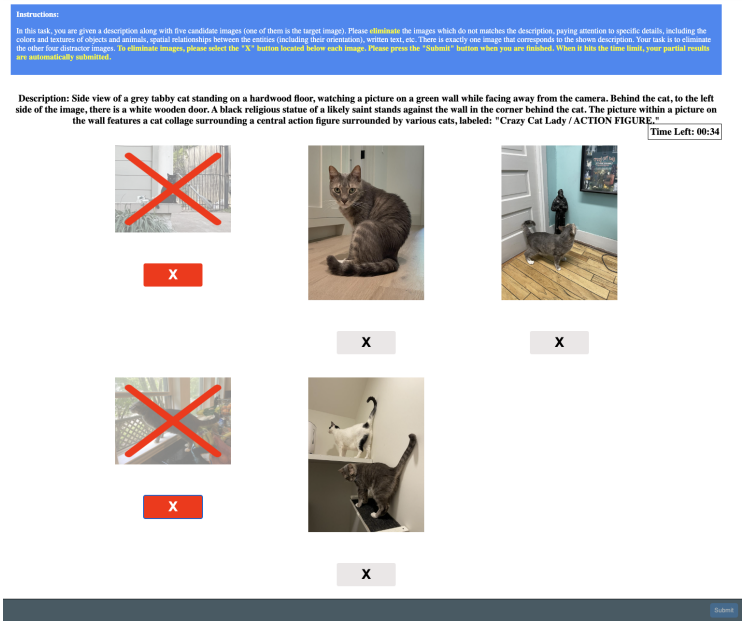


Fig. 10: The annotation UI for the human evaluation discussed in Section 3.3.

similarity scores and sample four as distractors, ensuring that all images appear as a distractor for at least one pivot. This produces 1,000 groups of five images, and each is evaluated by three annotators. We designed an elimination-based UI, as depicted in Figure 10, because eliminating unmatched images is substantially quicker and simpler than choosing a pivot image from a set of five images. This approach allows annotators to make judgments without needing to read the entire description, consequently reducing the average response time to less than a minute.

Side-by-Side Human Evaluation In this side-by-side (SxS) human evaluation framework, annotators are asked to provide a 5-point Likert scale score for both precision and recall. They are also instructed to highlight text spans with incorrect information in red and to mark text containing information that is missing but present in another description in blue. Additionally, annotators must provide justifications in text form. See Figure 5 for an illustrative example.

D Object Statistics in DOCCI Images

Figure 12 plots the counts of popular object types detected by an object detection tool that was run on the images. Since images have been taken in everyday scenes, the object coverage is remarkably diverse, capturing a wide range of subjects from both indoor and outdoor settings.

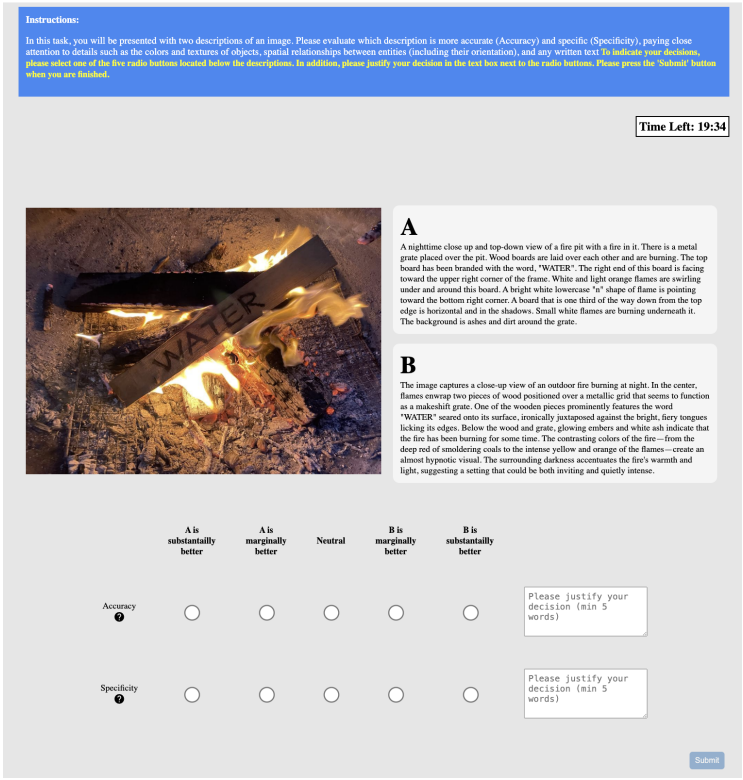


Fig. 11: The annotation UI for the side-by-side human evaluation discussed in Section 4.

E Error Analysis for Text-to-Image Generation models

We discuss the common error modes exhibited by the three SOTA models on the DOCCI qualification test set. We compare the generated images along with their descriptions and reference images in Figure 13. Note that each model has different maximum input lengths: 128 for Imagen, 4,000 characters for DALL-E 3, and 77 tokens for SDXL. Any prompt exceeding these limits will be truncated, potentially limiting the ability of models, especially Imagen and SDXL, to fully incorporate all information from the descriptions into the generated images.

(a) The three models show different types of errors. Imagen misunderstands a mural, resulting in the generation of a photorealistic image of a dog and cat (**style**). The image generated by DALL-E 3 combines a real car parked in the lot with a drawing of a cat and dog, which is an example of **media blending** [64]. SDXL completely misses the number of cats (**counting**), and a dog appears in a location where it is physically impossible (**common sense**). All models struggle with spatial relations and directions of the objects (e.g, the orientations of the cat and dog, the directions of their heads).

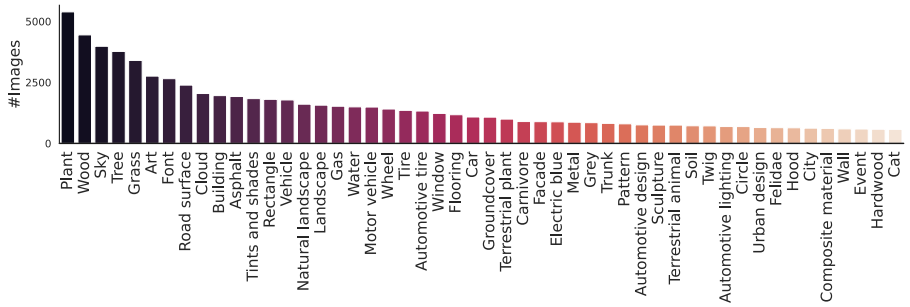


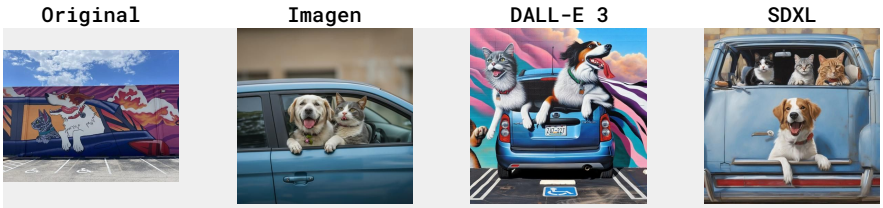
Fig. 12: 50 most common objects that appear in the DOCCI images. We use an off-the-shelf object detection tool and count object labels.

(b) DALL-E 3 impressively captures the key idea of this prompt, the statue is in a serving position and is trying to hit the sun. However, its **viewpoint** (a high angle view instead of a front view) and **style** are inaccurate, and two tennis balls are added (**hallucination**). Imagen and SDXL completely miss the key idea of this prompt, likely due to their limited input lengths.

(c) This example involves an uncommon variant of a familiar object, a toy clock with colorful blocks, and text rendering. Imagen demonstrates relatively better rendering of letters but misspells "watch." More critically, it depicts a standard clock instead of the desired toy clock with hour numbers on small, colorful blocks (**strong linguistic prior**). DALL-E 3 creates an image of a clock with small blocks on its face; however, it confuses the blocks beneath the clock with those on the clock face (**feature blending**). SDXL fails to produce an accurate image of a toy clock (e.g., the second hand sticks out from the clock face).

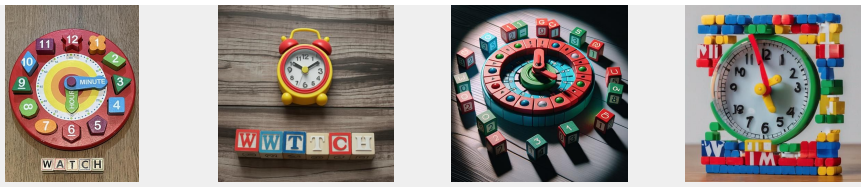
(d) All three models struggle to count objects correctly even in this simple case (**counting**). DALL-E 3 correctly depicts the horses in varying sizes (the leftmost horse is one-third the size of the others) and with the correct orientation (all the horses are facing right). However, it fails to capture the **spatial relation** accurately; the tile wall is not positioned directly behind the toy horses. Imagen and SDXL fail to capture the correct orientation because this information is provided later in the description, which might lead to it being truncated.

These examples suggest that T2I models have difficulty adhering to detailed descriptions precisely, primarily because of their limited understanding of textural input (e.g., orientations/directions in words) and architectural constraints, (e.g., the maximum input length). We urge further model development for future work.

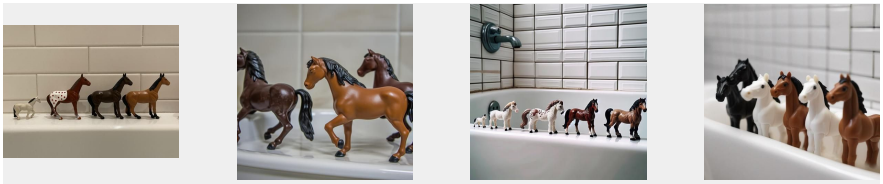


(a) An outdoor wide angle shot large wall art mural of a white and brown dog and a black and gray cat leaning out of the rear driver side window have a blue compact car with both of their front paws hanging out. The dog's mouth is open, lower teeth exposed, and its red tongue flailing backwards in the wind. Both animals look forward to the left, with the cat's ears straight up and the dog's ears also flailing in the wind. The cat is wearing a green color and the dog with a red collar, each with a round tag angled toward the back from the wind. The background of the mural has a pink and purple zebra stripe pattern. Orange, red, and grey clouds are painted behind the car along the upper edge of the mural, adding depth to the image. White lines are painted on the asphalt Crown in front of the mural, creating four parking spots. The two parking spots to the left contain a white handicap symbol. A shadow of the wall is being cast down onto the parking blocks aligned in front of the wall from the sun high and behind. The viewable blue sky above consist of a large cluster of altocumulus clouds.

(b) A front view of a statue on cement in a park. The statue is of a person with a tennis racket in their right hand. They are facing to the left, and the left arm is extended out to the sky. The statue is dark and covered in a shadow. It is on a cement square with dirt around it. The sun is shining on it. The extended arm with the hand open and up is directly under the sun that is shining in the blue sky. It looks like the sun is the ball the statue is releasing to hit for a serve. Stones are around the dirt and cement square that the statue is on. A shadow is on the cement and dirt in the front. Some green plants are around the square. There are 2 black spotlights on the ground aimed at the statue. Grass is around the area, and bushes can be seen on the right in a circle with stones around them. Trees are in the distance on the far right. Past the grass are parked cars on cement. Some tall trees are on the far left. A tall tree is behind the statue. A flag pole is near a white parked car on the left with an American flag on it that is hanging down.



(c) An up-close view of a colorful clock toy is seen with small blocks below spelling 'WATCH'. The clock base is red, with shaped blocks around the edge in different colors with white numbers printed on them from one to twelve. Inside the clock are two arms to show the time, the short green one reads 'HOUR', while the long blue one reads 'MINUTE'. Intervals of five are seen next to small green dots inside the clock circle. The toy and blocks sit on a cool-tone faux wood tabletop with black striations. The light is coming from a slight angle above the clock, as evident from the shadow being cast down from the clock toy.



(d) An indoor, close up shot of the side of 4 small horse toy figures placed on the side of the bathtub, with a white tile wall directly behind the horses. The left most horse is one third of the size compared to the others. The left most horse is completely white with a black mane and tail. The horse second to the left is brown with a brown mane and tail, with its left half of its body covered in white with red dots. The third horse to the left is dark brown with a black mane and tail. The horse all the way on the right is light brown with a black mane and tail. All the horses are facing to the right.

Fig. 13: Images generated by SOTA T2I models using DOCCI descriptions as prompts. The leftmost image is the reference, and the remaining four images are generated by Imagen, DALL-E 3, and Stable Diffusion XL, respectively.

F Experimental Details

F.1 PaLI 5B

We used a version of PaLI 5B, not trained on captioning tasks, as a base model. We use the hyperparameters explained in the original paper (e.g., 812x812 as the input image size).

F.2 GPT-4v

For all GPT-4v experiments, we use this prompt to generate descriptions: "`Generate a detailed image description with around 120 words, but you may adjust the length if you want.`". The max output token length is set to 500.

F.3 FID, CMMD, $\text{FD}_{\text{DINOv2}}$

We used the public implementation of FID: https://github.com/mseitzer/pytorch-fid, CMMD: https://github.com/sayakpaul/cmmd-pytorch, and $\text{FD}_{\text{DINOv2}}$: https://github.com/layer6ai-labs/dgm-eval.

G Clarification on Densely Captioned Images (DCI)

The DCI dataset [59] provides captions that could contain over 1000 words (full caption). As mentioned in Section 2.2, we use only the extra captions and exclude captions of the submasks (i.e., image segments). During our manual inspection, we found that the full captions of DCI largely concatenate captions of submasks, and thus they are not coherent and concise paragraphs. For example, sentences might not be seamlessly connected since they are written independently for different submasks. Additionally, the order of captions of submasks does not necessarily align with the importance of the submasks (e.g., primary objects should be described earlier in the description). In DOCCI, we deeply care about linguistic structures and writing quality. Due to this discrepancy in annotation goals, the full captions of DCI are not directly comparable with DOCCI. Instead of using the full descriptions, we take the most comparable portion of DCI captions available in the "extra_caption" field. We do not concatenate the "short_caption" and "extra_caption" fields since, in many cases, their contents are redundant (e.g., the first sentence of "extra_caption" is very similar to "short_caption"). We use the "short_caption" field when the "extra_caption" field is empty.

H Datasheet for DOCCI

We provide our responses to the questions listed in the Datasheets for Datasets [20].

H.1 Motivation For Datasheet Creation

What tasks could the dataset be used for? DOCCI is directly usable for text-to-image and image-to-text generation tasks. Additionally, it can facilitate other vision-language tasks, such as image-to-text and text-to-image retrieval.

Who funded the creation dataset? Google Research

H.2 Datasheet Composition

What are the instances? Are there multiple types of instances? Images with text descriptions

How many instances are there in total? 14,847 annotated images (DOCCI) and 8,932 unannotated images (DOCCI-AAR)

What data does each instance consist of? A single instance consists of an image and a text description.

Is there a label or target associated with each instance? We provide entity tags for 15 distinct entities that occur in multiple images.

Is any information missing from individual instances? The entity tags mentioned above are only available for certain images, not for all images.

Are relationships between individual instances made explicit? We provide the cluster ID for each image. Please note that these clusters are identified using k-means, not by human annotators.

Does the dataset contain all possible instances or is it a sample of instances from a larger set? DOCCI is a newly created dataset and is not a subset of any existing dataset.

Are there recommended data splits? We split DOCCI into four sets: 9,647 train, 5,000 test, 100 qualification-dev, and 100 qualification-test. We split DOCCI-AAR into 3,932 trai and 5,000 test sets.

Are there any errors, sources of noise, or redundancies in the dataset? Annotation errors, such as precision and recall errors and typos, may be present in the dataset. DOCCI is designed to include similar images; however, we have removed images that are exactly the same, based on similarity scores.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? DOCCI is self-contained.

H.3 Collection Process

What mechanisms or procedures were used to collect the data? All images were taken by one of the authors and their family. All text descriptions were written by human annotators. We do not rely on any automated process in our data annotation pipeline.

How was the data associated with each instance acquired? We curated all images and annotated text descriptions. We do not use any existing datasets or other data sources.

If the dataset is a sample from a larger set, what was the sampling strategy? We did not sample anything from a larger set.

Who was involved in the data collection process and how were they compensated? We employ in-house annotators who are compensated on an hourly basis, at rates well above the legal minimum wage.

Over what timeframe was the data collected? The images were curated from August 2021 to September 2022. The text descriptions were annotated in 2023.

H.4 Data Preprocessing

Was any preprocessing/cleaning/labeling of the data done? We manually reviewed all images for personally identifiable information (PII), removing some images and blurring detected faces, phone numbers, and URLs to protect privacy. For text descriptions, we instructed annotators to exclude any PII, such as people's names, phone numbers, and URLs. After the annotation phase, we employed automatic tools to scan for PII, ensuring the descriptions remained free of such information.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? No

Is the software used to preprocess/clean/label the instances available? No

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes

H.5 Dataset Distribution

How will the dataset be distributed? The dataset is available at `https://google.github.io/docci`.

When will the dataset be released/first distributed? What license (if any) is it distributed under? We release the dataset in March 2024. DOCCI will be released under the CC-BY 4.0 license.

Are there any copyrights on the data? No

Are there any fees or access/export restrictions? No

H.6 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? This dataset will be maintained by the authors of this paper.

Will the dataset be updated? As DOCCI is designed for evaluation purposes, we do not anticipate any future updates. However, should significant errors be discovered within the dataset, we may consider making modifications.

How will updates be communicated? Updates will be posted on the dataset website.

If the dataset becomes obsolete how will this be communicated? Updates will be posted on the dataset website.

H.7 Legal and Ethical Considerations

Were any ethical review processes conducted (e.g., by an institutional review board)? Yes

Does the dataset contain data that might be considered confidential? No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No

Does the dataset relate to people? Very few images, as taken, contained PII including people.

Does the dataset identify any subpopulations? We manually reviewed all images for PII. We removed some images and otherwise scrubbed any detected faces, phone numbers, and URLs by blurring them.

Is it possible to identify individuals, either directly or indirectly from the dataset? No (see above)

Does the dataset contain data that might be considered sensitive in any way? No

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources? We collected text descriptions from human annotators whom we hired.

Were the individuals in question notified about the data collection? Yes