DOCCI: Descriptions of Connected and Contrasting Images

Yasumasa Onoe[1†], Sunayana Rane[2†*], Zachary Berger[1], Yonatan Bitton[1], Jaemin Cho[3*], Roopal Garg[1], Alexander Ku[1], Zarana Parekh[1], Jordi Pont-Tuset[1], Garrett Tanzer[1], Su Wang[1], and Jason Baldridge[1]

[1]Google, [2]Princeton University, [3]UNC Chapel Hill
https://google.github.io/docci

Abstract. Vision-language datasets are vital for both text-to-image (T2I) and image-to-text (I2T) research. However, current datasets lack descriptions with fine-grained detail that would allow for richer associations to be learned by models. To fill the gap, we introduce **Descriptions of Connected and Contrasting Images (DOCCI)**, a dataset with long, human-annotated English descriptions for 15k images that were taken, curated and donated by a single researcher intent on capturing key challenges such as spatial relations, counting, text rendering, world knowledge, and more. We instruct human annotators to create comprehensive descriptions for each image; these average 136 words in length and are crafted to clearly distinguish each image from those that are related or similar. Each description is highly compositional and typically encompasses multiple challenges. Through both quantitative and qualitative analyses, we demonstrate that DOCCI serves as an effective training resource for image-to-text generation – a PaLI 5B model finetuned on DOCCI shows equal or superior results compared to highly-performant larger models like LLaVA-1.5 7B and InstructBLIP 7B. Furthermore, we show that DOCCI is a useful testbed for text-to-image generation, highlighting the limitations of current text-to-image models in capturing long descriptions and fine details.

1 Introduction

The past several years has produced a continual, marked evolution of text-to-image (T2I) generation models (e.g. [6,48,49,51,64], and many more), leading to not only improved capabilities and progress on research benchmarks, but deployment in user-facing applications (e.g., [1,38,42,49], and many more). Nevertheless, even the best current models still exhibit weaknesses in key areas, including precise handling of spatial relationships between objects, correct object counting, and accurate text rendering [4,13,34,64]. As we look to improve our research understanding of T2I models and the impact of their limitations on real-world applications, it is essential to identify their weaknesses precisely and efficiently.

Many test prompt sets have been developed [11,12,51,61,64] to assess model behaviors in a controlled manner (e.g., image-text alignment). The common

[†]Equal contribution. [*]Work done as a Student Researcher at Google.

Fig. 1: An example detailed description of a DOCCI image. The color of the text corresponds to each aspect of the details listed below the description. A more comprehensive list is presented in Appendix B. NOTE: this figure illustrates rich visual information in our descriptions, but we do not annotate spans with these information types.

practice involves generating images for test prompts and then obtaining automatic evaluation scores, either through embedding-based approaches [21] or VQA-based approaches [11, 23, 62]. But, these test prompts are often simplistic and fail to specify critical details, such as the orientation, direction, and fine-grained attributes of the key visual subjects (e.g., *"a cat standing on a horse"* can be as specifically described as *"a left-facing grey British short-hair perched on a white and brown-spotted Mustang horse."*). Crucially, these prompt sets lack ground-truth images, making it impossible to directly compare generated images with corresponding reference images. One way to address this issue is to use existing human-annotated image-caption datasets like COCO [36]. Unfortunately, the captions in these datasets are typically brief (e.g., COCO captions average around 10 words) and lack details of the visual features in the images. The recently introduced Densely Captioned Images (DCI) dataset provides descriptions with over 1,000 words per image [59]. But, those descriptions concatenate short captions of image segments, which lack rich linguistic structures and coherence. Additionally, their images are sampled from SA-1B [29], which were not taken specifically with the intent of evaluating T2I models.

To fill this gap, we introduce a new vision-language dataset, **Descriptions of Connected and Contrasting Images** (DOCCI, pronounced *doh-chee*). Fig. 1 demonstrates the level of detail included in our descriptions, including its coverage of multiple aspects of the image. DOCCI contains 15k images – all taken, selected, framed and curated by the lead author, Jason Baldridge – along with manually annotated detailed text descriptions. The images intentionally focus on many of the failure modes noted in section 6.3 of the Parti paper [65] – in fact, the initial set of DOCCI pictures were taken during the course of the development of the Parti model. Example challenge aspects that were targeted include

Fig. 2: Examples of related images. DOCCI images were intentionally collected to have substantially similar, contextually related groups of *distractor* images. The text descriptions must be detailed enough to differentiate each image from related ones.

precise visual properties such as complex attribute-object binding, spatial relationships, multimedia blending, counting, and different types of optical effects. The complexity of images varies from very simple ones (text on a blackboard) to highly complex ones (detailed street wall murals and their surrounding context). Additionally, there are multiple images of the same or similar objects, e.g., each with slight differences in their spatial orientations and counts, in line with the concept of contrast sets [19]. This approach enables a precise and localized investigation of model behaviors, thereby making the evaluation more rigorous and challenging. DOCCI images are free of personally identifiable information (PII) and have been donated to the public domain under the CC-BY license. Equipped with the newly-curated images and detailed descriptions, DOCCI covers a wide range of outstanding issues for T2I models.

Annotating detailed yet concise descriptions for images from scratch is challenging. For efficiency, we divide the text annotation process into three stages (see Figure 3). In the first stage, annotators write short descriptions of objects based on the predefined rubric, ensuring they capture all the salient details. The second stage consolidates those short descriptions into one detailed, coherent natural language description. The final stage enriches the description by adding important details such as colors, textures, and the relationships between vari-

ous elements. We rigorously implement quality control steps to ensure that each description meets our high standards of annotation.

We evaluate current highly-performant T2I and I2T models with DOCCI to conduct both quantitative and qualitative analyses. We first demonstrate that, combined with a sample efficient model such as PaLI 5B, DOCCI can greatly improve I2T generation. To assess this, we introduce a framework for evaluating long image descriptions, including the side-by-side human evaluation setup with the precision (i.e., hallucinations) and recall (i.e., details) ratings. Our experimental results also show that the T2I models still exhibit numerous error modes including those related to spatial relationships, counting, and text rendering. We show that the limited input length of most T2I models is problematic as it causes significant parts of the description (i.e., prompt) to be omitted, making it impossible to include those details in the generated image. We also show the unreliability of automatic metrics such as FID [22] and CLIPScore [21], which do not align with the results of our human evaluation.

2 Dataset Construction

DOCCI is unique in its curation and annotation, as described overall in this section and in Appendices A and C in further detail.

2.1 Images

We summarize here the collection and curation of DOCCI images. See Appendix A for more details. All 15k annotated DOCCI images were taken by one of the authors, Jason Baldridge, and his family. The majority of these images were taken in the United States, spanning over fifteen states (especially California, Florida, Nevada, New York, Arkansas and Texas). A few were taken in other countries such as India, Iceland, and Italy. Most images are natural scenes captured in both indoor and outdoor settings and feature different types of lighting conditions. The choice of subjects was driven largely by opportunity – interesting scenes and things encountered over the course of August 2021 to September 2022, as well as a selection of relevant images taken before that period. Additionally, many images were specifically arranged or framed to test known limitations of text-to-image models, such as counting and spatial relationships and mixed media images (e.g. an image of a cat shown on a TV with a live cat in front of the TV). The images range from very complex ones containing intricate murals fronted by plants and signs, to quite simple ones like short handwritten words in chalk on pavement. Since the images capture everyday scenes, common objects include domestic/wild animals, plants, artwork, vehicles, toys, and elements of natural and urban landscapes (e.g., rivers, rocks, and buildings).

Most images are captured using an iPhone camera in landscape or portrait orientation. Typically, their size is 2048×1536 pixels, but some are smaller due to cropping that ensured the focus was on a specific element in the original shot. In addition, we release 8,932 *unannotated* **DOCCI-AAR** images curated in similar fashion from October 2022 to November 2023. These images also span multiple regions of the USA (especially New York, Texas, California, Michigan,

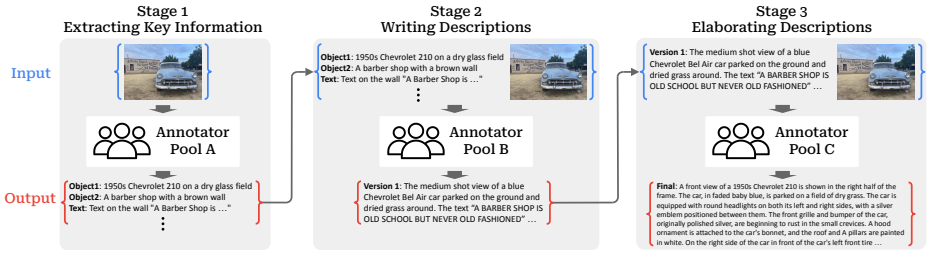


Fig. 3: Data Annotation Process. *Stage 1*: extract the key aspects, such as objects, from the image and write short descriptions. *Stage 2*: extend and combine these short descriptions into one overall description. *Stage 3*: elaborate and refine the description.

Arkansas, and Arizona) but also include a large number of images from Canada, Germany, Switzerland, and France. These images are not constrained to portrait or landscape mode; instead, they are cropped to select the most salient components and thus cover arbitrary aspect ratios (AAR).

Given the nature of their collection, DOCCI's images necessarily are a biased sample in terms of content and geographical extent. We hope others will donate images in similar fashion to expand the visual diversity available for research.

Contrastive Images Figure 2 shows examples of related images in DOCCI. The images were intentionally collected to include groups of related, substantially similar *distractor* images. For instance, a group of images depicts the same cats but in different orientations, poses, and actions. There could be several images of green apples placed on a table in various numbers and arrangements. Words, characters, and numbers can appear on various surfaces or materials, such as paper, brick walls, and stone, in diverse formats, including print, stickers, and handwriting. Those similar images are intentionally taken to challenge both T2I and I2T models, to test if they can correctly reflect the details in either direction.

Reoccurring Entities There are 15 distinct entities that occur in multiple images, including specific cats, dogs, vehicles and graffiti tags. All instances of these entities are tagged with their corresponding images in the dataset, and we will release these for future work on consistent character generation with DOCCI using methods like DreamBooth [50] (and its descendants).

License and Privacy As noted, the DOCCI images were donated by a single person and shared under the CC-BY 4.0 license. Very few images, as taken, contained personally identifiable information (PII). We manually reviewed all images for PII. We removed some images and otherwise scrubbed any detected faces, phone numbers, and URLs by blurring them by hand.

2.2 Text Descriptions

We hypothesize that good descriptions include sufficient details of the key objects and their attributes as well as salient information of secondary objects and background. In addition, a good description should be well-organized and read like a newspaper article: important information is covered in early sentences, while secondary information is mentioned later, thereby effectively triaging key

Table 1: Statistics for DOCCI and other datasets. #Words and #Sent. give the average number of words and sentences per description, respectively. For DCI, we only use the extra captions and exclude descriptions of submasks (see Appendix G).

	Images		Descriptions		
Dataset	Sources	Size	Size	#Words	#Sent.
DOCCI (ours)	Author donation	14,847	14,847	135.9	7.1
DCI (extra caption)	SA-1B	8,012	8,012	144.7	10.1
Stanford Vis. Par.	COCO, Visual Genome	19,561	19,561	68.5	6.3
Localized Narratives	COCO, Open Images	848,749	873,107	41.0	2.6
COCO	Flickr	123,287	616,767	11.3	1.0

details. To clarify the goal of our annotation task, we focus on the key visual features such as **objects, attributes, spatial relationships, text rendering, counting, world knowledge, scenes, views, and optical effects**. See Appendix B for further detail on annotation interfaces and guidelines.

Annotation Protocol Writing detailed and high-quality descriptions for images demands a broad skill set, including extensive knowledge about various objects and proficient writing skills. During pilot studies, it was clear that composing a detailed description of an image from scratch is time-consuming and tiring, even for expert annotators. To enhance efficiency, we divide our annotation process into three stages (Figure 3), distributing the required skills and workload more effectively. In the first stage, we extract the key aspects (e.g., the main objects and their attributes) and create concise descriptions of each. In the second stage, we combine these brief descriptions into a preliminary draft. Finally, in the third stage, we add further detail and refine the description. In some images, collecting detailed information is difficult when relying solely on visual cues. To ensure that the final descriptions are deeply grounded in the context of the images, we provide background information (e.g., specific car makes, whether it was sunrise or sunset) to the annotators when available.

3 Dataset Analysis

We analyze the features, functionalities and quality of DOCCI, and compare it with existing datasets including DCI [59], Stanford Visual Paragraphs [31], Localized Narratives [46], and COCO Captions [36].

3.1 Dataset Statistics

Table 1 lists key statistics for DOCCI and prior datasets. On average, DOCCI's descriptions are substantially longer than those in the Stanford Visual Paragraphs dataset and have similar length to DCI's. However, the average sentence count in DOCCI descriptions is lower than in DCI: DOCCI's sentences are denser. This discrepancy becomes even larger when compared to larger datasets such as Localized Narratives and COCO, which are less detailed.

We further investigate the length of the descriptions, as this serves as a reliable proxy for identifying recall errors (i.e., missing information). Figure 4 displays the distribution of description lengths across each dataset. DOCCI has

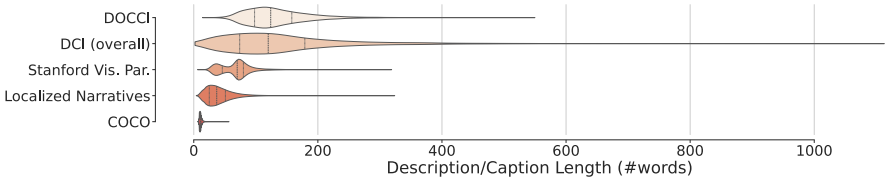
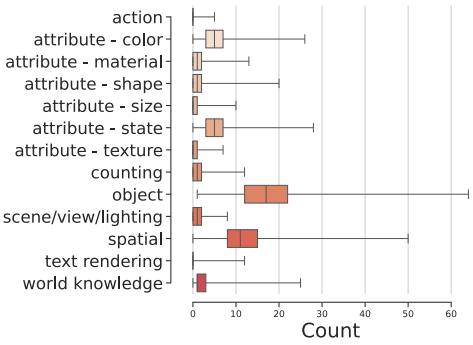


Fig. 4: The distribution of description lengths. The x-axis represents the number of words, and the vertical dotted lines in the violin plot indicate quartiles.

Table 2: The percentage of descriptions that contain each challenge type and the average count of that particular challenge type per image. Additionally, we show boxplots depicting the distributions of each challenge type over all images.

Challenge Type	Descriptions (%)	Avg.
action	20.9	0.3
attribute - color	97.3	5.3
attribute - material	60.9	1.3
attribute - shape	62.1	1.4
attribute - size	47.4	0.9
attribute - state	97.8	5.3
attribute - texture	25.7	0.3
counting	54.6	1.0
object	100.0	17.7
scene/view/lighting	63.6	1.2
spatial	99.9	11.5
text rendering	23.3	0.4
world knowledge	76.2	2.0



the highest median description length compared to other datasets, including DCI (which has the highest mean). The plot reveals the presence of outlier descriptions exceeding 1,000 words in DCI – which elevate its mean length.

We split **DOCCI** into four sets: **9,647 train, 5,000 test, 100 qualification-dev, and 100 qualification-test**. The test set is intended for computing automatic metrics. The qualification sets comprise manually selected images that specifically test prominent challenges in T2I models, intended for manual inspection or human evaluation. QUAL-DEV can be used by experimenters for their own qualitative comparisons. QUAL-TEST is intended to be held out for rating by human judges. We also split the **DOCCI-AAR** images into **3,932 train** and **5,000 test** sets, with the expectation that this will facilitate future experiments with automatic high-quality captioning (or, we hope, further human annotation).

3.2 Challenge Types

DOCCI's descriptions cover various types of challenges for T2I models, and one description can contain multiple challenges. We analyze the challenge types using DSG [11], which extracts challenge types from descriptions (e.g., Attribute-color). This automatically generated by an LLM and thus may contain errors, but it serves as an effective proxy for estimating the distribution of challenge types. Table 2 summarizes the percentage of descriptions per challenge type and the average number of challenge types per description. The descriptions include

Table 3: Language complexity and readability scores.

Dataset	Syntactic (↑)	Semantic (↑)	SMOG (↑)	FRE (↓)	#Errors (↓)
DOCCI (ours)	8.6	50.5	8.7	77.7	0.3
DCI (overall)	8.1	52.0	7.9	82.5	1.2
Stanford Vis. Par.	6.0	23.9	6.7	88.6	0.8
Localized Narratives	5.8	13.5	6.0	87.7	1.2
COCO	4.5	4.9	7.3	82.2	0.1

an average of 17.7 objects,[1] and their spatial relationships are mentioned in 99.9% of the descriptions. Object attributes are well covered: *color* and *state* are described in 97% of descriptions. Additionally, *count* is present in 54.6% of the descriptions, and *text rendering* in 23.3%. Each single description encompasses multiple challenge types, making DOCCI a challenging benchmark.

3.3 Description Quality

Detail Are DOCCI descriptions detailed enough to differentiate similar/related images? To answer this, 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 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. Given the description and the five images (pivot and four distractors), three annotators correctly identified the true (pivot) image 97.1% of the time, achieving Fleiss' kappa of 0.98. We confirmed that all negative cases were due to human errors. The high accuracy and strong agreement among annotators demonstrate that the descriptions capture essential and unique details of the pivots.

Language Complexity Table 3 compares language complexity and readability. For assessing language complexity, we evaluate two dimensions: the syntactic complexity, measured by the maximum depth of the dependency tree [39], and semantic complexity, indicated by the number of nodes in a scene graph. DOCCI and DCI – the datasets with longer descriptions – generally achieve higher complexity scores. DOCCI exhibits the highest syntactic complexity, while DCI achieves the highest semantic complexity score. For readability scores, we report the Simple Measure of Gobbledygook (SMOG) score [33] and the Flesch Reading Ease (FRE) score [28]. The scores indicate that DOCCI's descriptions are generally written in plain English, yet are not overly simplistic. Additionally, we count the average number of suggestions by an off-the-shelf spelling/grammar

[1] This number includes both primary and secondary objects. DSG often detects nested objects (e.g., tires of a car), leading to a higher count of objects detected.

Fig. 5: A side-by-side comparison of descriptions from DOCCI and one generated by GPT-4v. Blue-highlighted spans indicate details present in one description but absent in the other. Red-highlighted spans denote incorrect information (i.e., hallucination).

checker. On average, DOCCI generates 0.3 error suggestions per description, whereas DCI generates 1.2, indicating better quality control in DOCCI.[2]

4 Evaluating I2T Generation Models with DOCCI

We demonstrate the utility of DOCCI for image-to-text (I2T) generation by evaluating SOTA I2T models with both automatic metrics and side-by-side (SxS) human evaluation. Additionally, we conduct a SxS human evaluation of DOCCI descriptions compared to GPT-4v, to better understand key differences between human descriptions and high-quality machine-generated descriptions.

Setup We generate detailed descriptions for images from the test set using InstructBLIP (Vicuna-7B) [14], LLaVA-1.5 7B [37], and PaLI 5B [8,9]. Following their original setup, we use a different prompt for each model as described in their paper. PaLI has not been trained on captioning tasks during its pretraining phase; thus, we finetune it using the DOCCI training set (9,647 examples) and the COCO training set [36]. We report reference-based metrics for captioning such as BLEU@4 [44], ROUGE-L [35], METEOR [5], CIDEr [60], and the average number of words as proxies of the detail and density of descriptions.

SxS Human Evalution For SxS human evaluation, we focus on PaLI 5B *finetuned on DOCCI* compared to InstructBLIP, LLaVA, and GPT-4v, and generate descriptions with each model for the 100 DOCCI-QUAL-TEST images. Since GPT-4v generates lengthy descriptions, we prompted it to create shorter descriptions.

Even so, GPT-4v's average response length was the longest, at 147 words. Annotators indicate their preference in terms of **precision** and **recall** errors [27] (see Fig. 5). Here, precision primarily governs incorrect information (i.e., hallucinations), and recall penalizes generic or uninformative descriptions. We do not consider aspects of writing quality (e.g., fluency and word choice).

Quantitative Metrics Table 4 compares three I2T models on the qual-test set, using common reference-based metrics. Pali 5B (finetuned on DOCCI) generates longer descriptions (121.8 words on average), substantially improving all

[2] To ensure that DOCCI remains purely annotated by humans, we do not alter or modify descriptions based on suggested errors.

Table 4: I2T performance on the DOCCI test set. PaLI 5B finetuned on DOCCI outperforms other models by a substantial margin, indicating that DOCCI is an effective training data for I2T generation.

Model	Eval Mode	BLEU@4	ROUGE-L	METEOR	CIDEr	#Words
PaLI 5B, FT on COCO	finetune	0.0	11.3	3.6	0.0	15.1
PaLI 5B, FT on DOCCI	finetune	10.1	29.1	17.9	16.0	121.8
InstructBLIP (Vicuna-7B)	zero-shot	3.5	20.5	10.6	5.9	84.4
LLaVA-1.5 7B	zero-shot	3.5	22.0	11.3	6.4	89.5

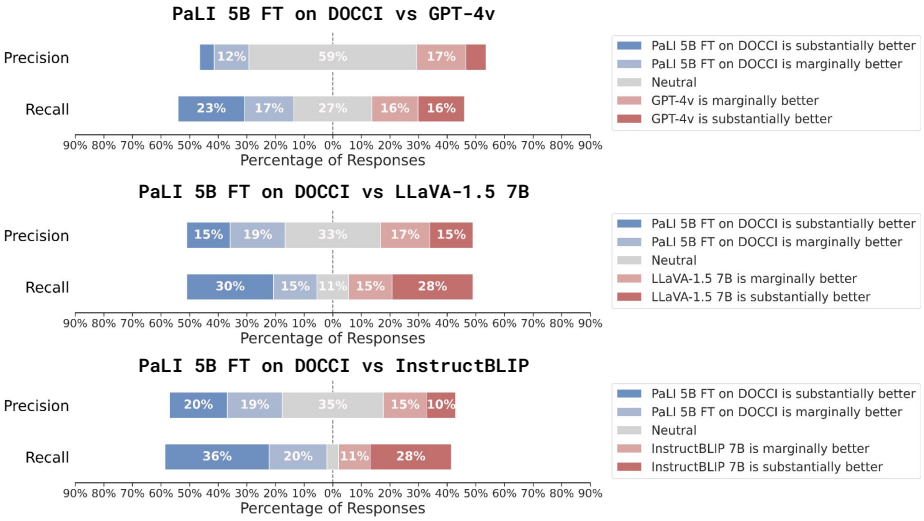


Fig. 6: Side-by-side human evaluation of descriptions generated by PaLI, GPT-4v, LLaVA, and InstructBLIP, with a specific focus on the visual features listed in Section 2.2. Note that we do not assess writing quality (fluency and word choice). In summary, descriptions by finetuned PaLI 5B contain more details compared to those three models (better recall scores), but it falls behind GPT-4v in terms of precision.

metrics and outperforming larger instruction-tuned models. This indicates that **the DOCCI training set is effective for fine-tuning and can drastically change the output length despite its relatively small size.** Note that we use only one reference description per image, and the choice of reference description impacts those scores [17]. Additionally, we still lack reliable automatic metrics for evaluating detailed and long image descriptions. Given this, we do not assess the content of the generated descriptions, leaving it for future research.

Human Evaluation Results Figure 6 plots the Likert scale for each model pair. The top bar plot shows that GPT-4v is more accurate than PaLI 5B finetuned on DOCCI, but PaLI includes more details. GPT-4v typically produces fluent and accurate descriptions, though they are not always concise, sometimes including speculative statements. Conversely, PaLI provides more details (e.g., spatial relationships, named entities), but this comes at the risk of generating inaccurate information. The middle plot indicates that human annotators slightly prefer PaLI over LLaVA on both precision and recall, while the bottom plot

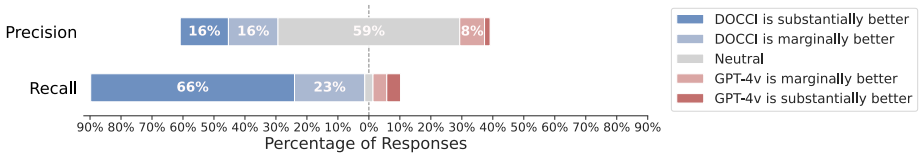


Fig. 7: Side-by-side human evaluation of the DOCCI descriptions and those generated by GPT-4v, with a specific focus on the visual features listed in Section 2.2. Note that we do not assess the quality of the writing such as fluency and word choice.

suggests that PaLI is preferred over InstructBLIP. Note that LLaVA has been trained on the instruction tuning data generated (158k) by GPT-4 [40], and InstructBLIP has been trained on a range of vision-language datasets adapted for instruction tuning (13 publicly available datasets). **Despite the fact that the DOCCI training set is relatively small (9.6k), the finetuned PaLI achieves remarkable performance, demonstrating the strong supervision in the DOCCI training set and the sample efficiency of PaLI 5B.**

DOCCI Descriptions vs GPT-4v GPT-4v [41] demonstrates impressive abilities in generating fluent, well-written descriptions. However, can GPT-4v create more detailed descriptions than those of human annotators? We generate descriptions of similar length to those in the DOCCI using images from DOCCI-QUAL-TEST. Figure 7 plots the 5-point Likert scale for precision and recall selected by annotators. For precision, as both DOCCI and GPT-4v descriptions rarely include incorrect information, the annotators selected "Neutral" 59% of the time. Other than "Neutral", annotators judge DOCCI descriptions as more accurate than those of GPT-4v. The annotators prefer DOCCI descriptions 89% of the time in terms of recall. Figure 5 showcases the details present in one description but absent in the other (blue-highlighted spans) and inaccurate information (red-highlighted spans). Although DOCCI descriptions are shorter, they include more detailed information compared to GPT-4v descriptions. **These findings show that large models like GPT-4v demonstrate remarkable capabilities in producing detailed descriptions, but that there are still important gaps with descriptions created by human annotators.**

5 Evaluating T2I Generation Models with DOCCI

In this section, we investigate how T2I models behave with long and detailed descriptions. We report the performance of current high-performing T2I models on DOCCI. For this, we compute automatic metrics for image quality and text-image alignment, along with side-by-side (SxS) human evaluation.

Setup We generate images based on DOCCI descriptions using three T2I models: a variant of Imagen [51], DALL-E 3 [42], and Stable Diffusion XL (SDXL) [45]. We report on three image quality metrics: FID [22], CMMD [25], and $\text{FD}_{\text{DINOv2}}$ [56], and two text-image-alignment metrics, CLIPScore [21] and DSG [11] on the test set.[3] For FID, CMMD, and $\text{FD}_{\text{DINOv2}}$, we use DOCCI

[3] We used 4,966 examples as DALL-E 3's content filter rejected 34 rewritten prompts (Our descriptions do not contain any sensitive content.).

Table 5: T2I performance by Imagen, SDXL, and DALL-E 3 on the DOCCI test set. For image quality metrics, we report random training images (RANDOM) and retrieved training images based on descriptions (TEXT RET.) as baselines. For image-text alignment metrics, we report scores using with the original images as an oracle (TEST).

	Image Quality			Image-Text Alignment		
Model	FID (↓)	CMMD (↓)	FD_{DINOv2} (↓)	CLIPScore (↑)	DSG_{VQA} (↑)	DSG_{Human} (↑)
Imagen	28.13	1.016	300.8	81.2	69.2	77.3
SDXL	23.69	0.823	267.2	85.9	65.2	69.8
DALL-E 3	32.37	1.691	300.8	80.1	76.3	85.6
RANDOM	13.71	0.002	142.8	–	–	–
TEXT RET.	13.43	0.003	133.1	–	–	–
TEST	–	–	–	80.8	78.7	91.7

images to compute the statistics of the reference distribution. We also report random training images (RANDOM) and retrieved training images based on descriptions (TEXT RET.) as baselines. For DSG, we compute the final score using a VQA model (DSG_{VQA}), with PaLM 2 340B [3] for question generation and PaLI 17B [9] for VQA. In addition, we ask human annotators to assess 100 samples from the test set to observe the correlation between the scores given by the VQA model and human judgment (DSG_{Human}). As oracle performance, we report the scores computed with the original test images (TEST). Additionally, we conduct side-by-side human evaluation using the 100 DOCCI-QUAL-TEST set, focusing on *user preference*. In this human evaluation, we ask annotators to rank three generated images based on the same description, considering both image quality and fit to the prompt, and report the mean rank of each model.

Automatic Metrics and User Preference Table 5 shows the zero-shot T2I generation performance of the models with automatic metrics. All three models substantially underperform the RANDOM and TEXT RET. baselines, and SDXL consistently achieves better scores than Imagen and DALL-E 3 (for FID, CMMD, and FD_{DINOv2}). These results run counter to our human evaluation, which rate DALL-E 3 and Imagen higher: **In our user preference evaluation, DALL-E 3 was rated the highest with a mean rank of 1.42, followed by Imagen at 1.84 and SDXL at 2.38.** This discrepancy between FID and human judgment is also reported in previous studies [43, 56].

Image-Text Alignment The right half of Table 5 lists three metrics for image-text alignment. SDXL achieves the highest CLIPScore, while DALL-E 3 performs the worst. However, DSG_{VQA} results in a conflicting pattern which aligns better with our human evaluation. Basically, CLIPScore is not suitable for long descriptions as the CLIP text encoder truncates just 77 tokens. While one can summarize a long description to fit this input limit, there will still be information loss. In contrast, DSG extracts atomic validation questions from the full description, distilling its full specification in a detailed and interpretable manner. It thus serves as a better proxy for image-text alignment. We additionally report the DSG results by human annotators instead of a VQA model to verify its reliability (DSG_{Human}). The absolute scores are higher than DSG_{VQA}

Fig. 8: Text-to-Image Reconstruction Quality. Top row (a) shows high-fidelity reconstructions by Imagen, DALL-E 3, and SDXL with CLIP similarities over 88%, due to detailed descriptions. Middle row (b) DALL-E 3 generates an image of a box truck instead of an open truck, viewed from an aerial perspective, and includes additional, unintended road signs. Bottom row (c) depicts all models' overemphasis of "green" from a vague description, highlighting the impact of inadequate detail in the input.

as human annotators can make better judgments in areas where VQA models fall short (e.g., spatial relations). The overall trend of $\text{DSG}_{\text{Human}}$ matches with DSG_{VQA} as well as our user preference evaluation. **DALL-E 3 tops the DSG scores likely due to the low truncation-caused information loss with its context length of 4k characters, in contrast to Imagen's 128 tokens and SDXL's 77 tokens.** We provide detailed error analysis in Appendix E.

Text-to-Image Reconstruction The detailed descriptions in the DOCCI dataset enable a benchmarking of text-to-image models' ability to recreate original images (meaning: compare a generated image to a reference image). This is an analysis not possible with prompt-only evaluation sets such as Parti Prompts [65] or Drawbench [52]. We utilize 5,000 test descriptions from the dataset to generate images using Imagen, DALL-E 3, and SDXL. The fidelity of these reconstructions to the original images is quantified using two metrics: CLIP (ViT-L/14@336px) [47] (image-to-image) and DreamSim [18], a newer metric designed to assess the resemblance of generated images to a reference. The resulting CLIP similarity scores—85.1 for SDXL, 82.8 for DALL-E 3, and 85.8 for Imagen and DreamSim scores—53.7, 54.1, and 51.6, respectively—while suggesting models perform comparably at a high level, conceal nuanced deficits in their understanding and recreation of complex imagery. Clearly, more work is needed on automatic metrics with respect to the level of detail given in DOCCI.

In-depth analysis reveals further insights, exemplified in Figure 8: (a) High similarity instances, depicted in the top row, where all models achieve close resemblance to the original images, typically occur with comprehensive descriptions. (b) The middle row showcases mixed similarity scenarios, highlighting certain models' superiority over others and exposing their relative strengths and weaknesses. (c) The bottom row presents cases of low similarity, where the generative models struggle due to underspecified visual features [24]. **These performance variations pinpoint the current models' limitations and establish DOCCI as useful means to identify the strengths and weaknesses in visual reconstruction by these models.**

6 Related Work

Over the past decade, the vision-language research community has developed various image-text datasets. In the early years, datasets such as Flickr30k [63], COCO [10, 36], and Visual Genome [32] provided annotations in the form of human-written captions for images depicting common objects from everyday scenes. Since then, captioning datasets have been evolving, for example, nocaps [2] annotated captions to more diverse objects [30], Localized Narratives [46] used more modalities (e.g., mouse tracking) for annotation, Stanford Visual Paragraphs [31] annotated dense and descriptive captions, and WIT [55] and Crossmodal 3600 [57] considered multilinguality. Another line of research focuses on scale, building much larger image-text pair datasets. YFCC100M [58] includes 100M images/videos that have been collected from the web. Conceptual Captions [7, 54] collected up to 12M images together with alt-text. RedCaps [15] provides 12M image text pairs collected from Reddit. WIT [55] is large scale as well as multilingual, providing 11.5M images with text in 108 languages. CLIP [47] and ALIGN [26] have been trained on large-scale web datasets containing 400M and 1.8B image alt-text pairs respectively. This trend continues further: LAION-5B [53] extended its size to 5B and WebLI [9] consists of 10B image-text pairs from 109 languages.

DOCCI primarily focuses on the density and quality of descriptions and is directly comparable with prior work such as Stanford Visual Paragraphs [31] and DCI [59], which have a similar balance of size and density. DAC [16] improves the quality of descriptions using an LLM and achieves higher performance on downstream tasks. However, our human evaluation results (Section 4) indicate that human annotations still have an advantage over (proprietary) machine generated/elaborated dense descriptions in terms of detail and lack of hallucinations.

7 Conclusion

In this work, we introduced **Descriptions of Connected and Contrasting Images (DOCCI)**, a new vision-language dataset that consists of 15k newly curated images with detailed descriptions annotated by humans. Using DOCCI, we showcased outstanding problems in T2I models and evaluation such as their limited input length and the unreliability of automatic metrics. We encourage the research community to develop improved model architectures and evaluation metrics that are better suited for detailed visual descriptions in future work.

Acknowledgement

First of all, we would like to express our gratitude to all members of the annotator team for their diligent and hard work on a very challenging and long-running task. We also give a huge thanks to Soravit Changpinyo and Radu Soricut for their thorough review and the constructive feedback provided on our paper. Many thanks also to Cristina Vasconcelos and Brian Gordon for their support with our experiments, and to Andrea Burns for insightful suggestions for the paper. Finally, Jason Baldridge is incredibly grateful to his family members who contributed by helping arrange scenes, taking pictures, and being patient while he took so many pictures – Cheryl, Olivia, Nash, Gray, and Esme Baldridge and Mary and Justin Reusch – and to pets Ivy, Tiger, DD and Yoshi for their roles as frequent subjects.

References

1. Adobe: Adobe Firefly (2023)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: ICCV (2019)
3. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv (2023)
4. Bakr, E.M., Sun, P., Shen, X., Khan, F.F., Li, L.E., Elhoseiny, M.: Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In: ICCV (2023)
5. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (2005)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K.P., Freeman, W.T., Rubinstein, M., Li, Y., Krishnan, D.: Muse: Text-To-Image Generation via Masked Generative Transformers. In: ICML (2023)
7. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)
8. Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., Salz, D., Xiong, X., Vlasic, D., Pavetic, F., Rong, K., Yu, T., Keysers, D., Zhai, X., Soricut, R.: PaLI-3 Vision Language Models: Smaller, Faster, Stronger (2023)
9. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Ruiz, C.R., Steiner, A.P., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: PaLI: A Jointly-Scaled Multilingual Language-Image Model. In: ICLR (2023)
10. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO Captions: Data Collection and Evaluation Server. ArXiv (2015)
11. Cho, J., Hu, Y., Baldridge, J.M., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-Image Generation. In: ICLR (2024)

12. Cho, J., Zala, A., Bansal, M.: DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In: ICCV (2023)
13. Conwell, C., Ullman, T.D.: Testing Relational Understanding in Text-Guided Image Generation. ArXiv (2022)
14. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In: NeurIPS (2023)
15. Desai, K., Kaul, G., Aysola, Z.T., Johnson, J.: RedCaps: Web-curated image-text data created by the people, for the people. In: NeurIPS: Datasets and Benchmarks Track (2021)
16. Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., Ullman, S., Karlinsky, L.: Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models. In: NeurIPS (2023)
17. Freitag, M., Grangier, D., Caswell, I.: BLEU might be Guilty but References are not Innocent. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) EMNLP (2020)
18. Fu*, S., Tamir*, N., Sundaram*, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. arXiv (2023)
19. Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N.F., Mulcaire, P., Ning, Q., Singh, S., Smith, N.A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., Zhou, B.: Evaluating Models' Local Decision Boundaries via Contrast Sets. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of EMNLP (2020)
20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., au2, H.D.I., Crawford, K.: Datasheets for Datasets (2021)
21. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning (2021)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (2018)
23. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In: CVPR (2023)
24. Hutchinson, B., Baldridge, J., Prabhakaran, V.: Underspecification in Scene Description-to-Depiction Tasks (2022)
25. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a Better Evaluation Metric for Image Generation (2024)
26. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
27. Kasai, J., Sakaguchi, K., Dunagan, L., Morrison, J., Le Bras, R., Choi, Y., Smith, N.A.: Transparent Human Evaluation for Image Captioning. In: NAACL (2022)
28. Kincaid, P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel (1975)
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. arXiv (2023)

30. Krasin, I., Duerig, T., Alldrin, N., Veit, A., Abu-El-Haija, S., Belongie, S., Cai, D., Feng, Z., Ferrari, V., Gomes, V., Gupta, A., Narayanan, D., Sun, C., Chechik, G., Murphy, K.: OpenImages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2016)
31. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A Hierarchical Approach for Generating Descriptive Image Paragraphs. In: CVPR (2017)
32. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
33. Laughlin, G.H.M.: SMOG Grading-a New Readability Formula. Journal of Reading (1969)
34. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H.B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.Y., Fei-Fei, L., Wu, J., Ermon, S., Liang, P.: Holistic Evaluation of Text-To-Image Models (2023)
35. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out (2004)
36. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)
37. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: NeurIPS (2023)
38. Midjourney: Midjourney (2022)
39. Ohta S, Fukui N, S.K.: Computational principles of syntax in the regions specialized for language: integrating theoretical linguistics and functional neuroimaging (2013)
40. OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J.,

Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: GPT-4 Technical Report (2024)
41. OpenAI: GPT-4V(ision) system card (2022)
42. OpenAI: DALL·E 3 system card (2023)
43. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In: CVPR (2023)
44. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: ACL (2002)
45. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: ICLR (2024)
46. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting Vision and Language with Localized Narratives. In: ECCV (2020)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
48. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents (2022)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022)
50. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
51. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: NeurIPS (2022)
52. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
53. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B:

An open large-scale dataset for training next generation image-text models. In: NeurIPS: Datasets and Benchmarks Track
54. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
55. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning. In: SIGIR (2021)
56. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, E., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In: NeurIPS (2023)
57. Thapliyal, A., Pont-Tuset, J., Chen, X., Soricut, R.: Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In: EMNLP (2022)
58. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: the new data in multimedia research. Commun. ACM (2016)
59. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions (2023)
60. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
61. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., Baldridge, J., Norouzi, M., Anderson, P., Chan, W.: Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In: CVPR (2023)
62. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What You See is What You Read? Improving Text-Image Alignment Evaluation. In: NeurIPS (2023)
63. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL (2014)
64. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling Autoregressive Models for Content-Rich Text-to-Image Generation (2022)
65. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)