A Challenging Benchmark for Testing the Understanding and Reasoning Capabilities of Scientific Text-to-Image Generation Models

Anonymous ACL submission

Abstract

We present Scientific Text-to-Image Reasoning (STIR), a dataset of 200 scientific graph tasks in which each task is associated with a textual description, a reference latex code, and a reference image. Further, we propose a taxonomy that defines the . Based on this dataset, we define three challenging tasks: Text-to-Image generation; Text-to-Latex generation; and xxxxxxx. In this paper, we focus on the Text-to-Image task – providing a comprehensive evaluation of STOA ChatGPT-4 a strong baseline model with an extensive error analysis with proposed taxonomy. Our analysis suggest that LLMs' mistakes and under-performance reflect both spatial and non-spatial factors. These findings suggest that LLMs appear to capture certain aspects of scientific image implicitly, but room for improvement remains.

1 Introduction

SE: more are comments

SE: this is a quick summary from my side

- Multimodal models (e.g., text-to-image models) are super exciting, e.g., Dall-E 3
- Multimodal models for the scientific domain could accelerate scientific progress by enabling scientists to faster, better and more effectively generate scientific figures, thus sharing their ideas more effectively
- Generating high-quality scientific images from text would also foster inclusiveness and diversity, as not all authors are capable of doing so
- However, this field is in its infancy and while models have been proposed models recently (e.g. Automatikz), it is unclear what kind of understanding these models have of spatial visual relations or forms of visual reasoning
- we fill this grap in this work

Our contributions are

- we provide a taxonomy of tasks testing the understanding of scientific text-to-image generation models, including spatial understanding, understanding color attributes, numeric reasoning, etc.
- From the taxonomy, we create a dataset with textual captions and desired reference images
- We apply several models capable of performing (scientific) text-to-image generation on the dataset
- We provide human annotation
- We explore how metrics correlate with human annotation

2 Related work

In computer vision and multimodal studies, there are benchmarks and datasets serving for various purposes, including object detection, e.g. MS COCO (Lin et al., 2014), image classification, such as CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009), hand-written digits recognition: MNIST (Deng, 2012), and image captioning (Sharma et al., 2018 and Microsoft COCO Captions by Chen et al., 2015). In the scientific domain, Paper2Fig100k (Rodriguez et al., 2023) contains over 100k images of figures and texts extracted from research papers.

2.1 Benchmarks for Reasoning

In adddition to image perception, various datasets or benchmarks have emphasize more challenging cases in model assessment, such as visual reasoning (image as input in Section 2.1.1) or the reasoning

in text-to-image generation (Section 2.1.2) for instance reasoning about understanding spatial, temporal, and logical concepts and generating images as output.

2.1.1 Image as Input

In this subsection, we introduce several benchmarks where images serve as input and are utilized to evaluate models' performance in image comprehension.

The first type is visual question answering benchmarks, where images serve as input accompanied with a question regarding the image content. Multimodal Visual Patterns (MMVP) Benchmark (Tong et al., 2024) concentrates on the most challenging cases, comprising 150 CLIP-blind pairs (images that CLIP model perceives as similar although with clear visual distinctions) with questions aimed at probing image details such as relative position (e.g., whether the dog faces left or right), counting (e.g., how many eyes are in the image), or other attributes of objects. In CLEVR (Johnson et al., 2017), questions are designed to assess attribute identification, counting, comparison, spatial relationships, and logical operations.

In the scientific domain, there are also a few benchmarks that challenge multimodal image reasoning abilities by providing an image and a question as input to the model. MMMU (Massive Multi-discipline Multimodal Understanding, Yue et al., 2023) is a domain-specific benchmark designed with visual question-answering (an image with a question as the input and the answer to the question as the output) in six disciplines and data is mainly collected from college exams, quizzes and textbooks. This challenging benchmark is crafted to assess models' image comprehension and domain expertise.

MathVista (Lu et al., 2024) is a benchmark consisting of around 6000 mathematical questions along with an image reference. ScienceQA (Lu et al., 2022) is also a reasoning benchmark, but it employs an image as the context for input questions rather than posing questions directly about the image's content. Additionally, it incorporates the Chain of Thought (CoT) into its datasets, enhancing interpretability alongside the answers.

The second type for visual understanding is alignment of caption-image pairs. For instance, Winoground (Thrush et al., 2022) is a benchmark devised to correctly match images and captions. The dataset is organized in pairs, with a pair of caption that contain the same words but differ in word order. Some objects or predicates in captions are swapped, such as "there is a mug in some grass" and "there is some grass in a mug".

Apart from two types introduced above, there are also dataset contains mixed data types. Visual Genome (Krishna et al., 2017) serves as a dataset for grounding visual concepts to language, facilitating the connection of structured image concepts to natural language and knowledge bases. The dataset provides scene graphs, relationships between entities, and question-answer pairs regarding the image content.

2.1.2 Image Generation

Compared to visual reasoning, the text-to-image generation's reasoning ability remains relatively underexplored. One benchmark serves for this purpose is T2I-CompBench (Huang et al., 2023). It comprises 6,000 compositional text prompts and real-life images categorized into three groups: attribute binding (e.g., colour and shape), object relationships (e.g., spatial relationships), and complex compositions (involving both attributes and relationships, such as "The brown shoes were on top of the blue rug"). Multi-Modal CelebA-HQ (Xia et al., 2021) is another dataset for text-guided image generation, but it focuses specifically on real face images.

2.2 Evaluation of Scientific Image Generation

General benchmarks or metrics used to evaluate image generation may not necessarily be applicable to the scientific graph domain. We highlight the distinctions between scientific graphs and general graphs and assume that scientific graphs should prioritize accuracy in representing scientific concepts and ideas. This includes ensuring the accuracy of numerical values and adhering to certain conventions when simplifying real-life objects into graphical representations (e.g., representing a "battery" with a rectangle in a circuit diagram).

Although image-text alignment for general objects can be evaluated using metrics like the CLIP score (Hessel et al., 2021), such evaluations have not yet been extended to the scientific domain. In the scientific context, accuracy in both image generation and text description is paramount. It remains uncertain whether nuances such as the precision of numerical representations or the proportions of bars in a bar chart can be adequately captured by existing evaluation metrics.

Moreover, evaluating scientific graphs poses challenges even for human evaluators, as it requires specialized scientific or professional knowledge. Crowd evaluation may be difficult due to the complexity of the scientific concepts involved. Even when captions and images are provided, human evaluators may struggle to assess the quality of scientific images if the information is presented in an uncontrolled manner. To address these challenges, we propose a novel approach.

Our benchmark is designed for XXX with format of xxx

Furthermore, we provide the TikZ code alongside the images, which has the potential to assess both TikZ code and image generation.

3 Method

- 1) We construct a taxonomy suitable for text-image understanding (based on what?)
- 2) We build a dataset of pairs of captions and reference images. The dataset includes 0,1,2,... reference images per caption. How?
- 3) We apply several multimodal models on our dataset to test their scientific text-to-image conversion abilities/understanding
- 4) We use human eval and automatic metrics
- 5) Optionally: we fine-tune automatikz on our dataset

Instantiation of 1) below:

In a scientific graph containing a single object, such as a circle or a square, we assume that the image meets the requirements of the caption if the following factors are well-controlled:

- **Object**: Ensuring the shape is accurately represented according to the scientific standards required.
- **Properties of Objects**: includes colour, line or contour, size, and other relevant physical attributes. Additionally, the object may be subject to operations such as rotation or flipping to enhance clarity or emphasis.
- **Position** of the object in the image

In a word, a correct scientific graph with a single object should meet the shape × properties × position requirement.

The types of objects include geometric shapes, including those of 2 dimensions e.g. square or perpendicular, and also those in 3 dimensional space like a sphere. They can also be specific to a particular field or discipline. [UNFINISHED PARAGRAPH]

Composition: For graphs that combine two or more objects, the number of objects, their arrangement (the way of composition), and their comparative properties are crucial for accurate scientific representation. The way of composition may include the following examples.

- Two triangles **share the same side**.
- Two lines **perpendicular or parallel** to each other.
- Five circles with the **same centre** (five **concentric** circles).
- Three circles **without intersecting**.

Annotation: Additionally, annotations are often necessary to provide clarity or emphasize specific features. These may include textual annotations, for instance: text '`Linear Function:` $y = 3x$' for its function graph, or the annotation of 'A', 'B' and 'C' in the following example: `C is the midpoint of line AB`. Numerical annotations are also useful, for instance, in labeling the values on a bar chart. Additionally, scientific graphs may incorporate symbols or arrows to highlight or direct attention to particular segments or details of the graph."

References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Li Deng. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023. Ocr-vqgan: Taming text-within-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3689–3698.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Dataset	size	domain	input	output	type of challenges
MMVP : Multimodal Visual Patterns benchmark	300	general real-life images	image and question	answer to the input question	images with similar CLIP embeddings despite visual distinctions.
MMMU: Multi-discipline Multimodal Understanding and Reasoning Benchmark	11.5k	college exam collection	image and question	answer to the input question	domain knowledge from college exams, with question domain and question difficulty level annotated
Genome	108,077	general	image	a collection of reasoning types, including scene identification.	
science QA	21k	scientific	natural image (as context) and question	answer and explanation (CoT)	scientific problem and image reasoning
MathVista	6,141	MATH	math-related image and question	answer to the question	math reasoning
Paper2Fig	100k	scientific	images from papers	caption	complex cases of scientific figure collection
T2I-CompBench	6000	general	text	image	attribute binding, object relationships, complex composition
Wino-ground	1600 pairs	general	caption-image pairs, input-output may be used in a reverse way		linguistic/visual reasoning with words swapped

Table 1: Multi-model (text and image) reasoning in previous studies