

# Conversational Figures: A Conversational Question-Answering Dataset Grounded in Scientific Figures and Text

Anirudh Sundar, William Gay, Antonia Rabisheva, Larry Heck

AI Virtual Assistant Laboratory, Georgia Institute of Technology

{asundar34, wgay7, arabisheva3, larryheck}@gatech.edu

## Abstract

An emerging area of research in situated conversational AI is the creation of a Virtual Research Assistant (VRA). The VRA is a conversational agent that supports and amplifies human research. Among other challenges, the VRA must be capable of contextual dialogue grounded in scientific papers. An important element of conversational scientific papers is interpreting document-grounded figures to support an open dialogue question-answering interaction with the human researcher. This work introduces CONVERSATIONAL FIGURES (cFIGS), a dataset of multi-turn conversational question-answer pairs grounded in figures and their associated references from scientific documents available on arXiv.

## 1 Introduction

The development of conversational assistants capable of situated dialogue over structured knowledge sources remains an open problem (Sundar and Heck, 2022). An emerging area of research within this domain is conversational interactions over scientific documents with a Virtual Research Assistant (VRA). Given the rapid growth in the availability of scientific literature, developing a VRA and associated language models capable of understanding scientific documents is an important problem in Natural Language Processing (NLP). Scientific documents present an interesting challenge since the content is frequently multimodal. Besides textual paragraphs, researchers rely on various modalities to describe research methods. Images convey information about model architectures and pipelines, tables and graphical plots summarize experimental results for easy comparison, and equations describe mathematical formulations.

This paper focuses on the development of a dataset of conversations grounded in scientific figures and associated textual paragraphs. We introduce CONVERSATIONAL FIGURES (cFIGS), a

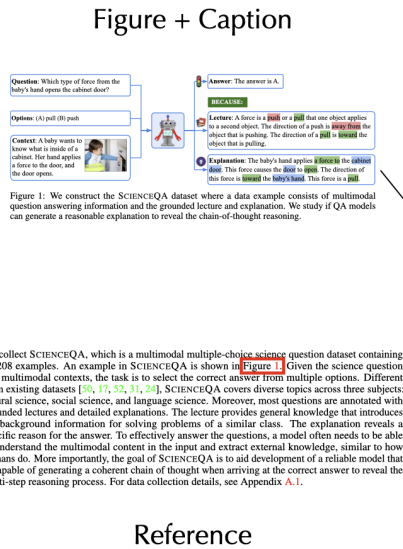
multimodal, multi-turn, conversational question-answering dataset grounded in content from both scientific figures and text. cFIGS is collected by building conversations from existing scientific datasets (Hsu et al., 2021).

## 2 Related Work

Early datasets on visual question-answering include VQA (Agrawal et al., 2016), Visual7W (Zhu et al., 2016), VisDial (Das et al., 2017), and MANY-MODALQA (Hannan et al., 2020). More recent datasets addressing the problem of open-domain conversations include IGC (Mostafazadeh et al., 2017), MOD (Fei et al., 2021), and Image-Chat (Shuster et al., 2020). However, the images in these datasets are collected from MS-COCO (Lin et al., 2014) or YFCC100M (Thomee et al., 2016), where visual content targets commonly seen everyday objects as opposed to scientific documents where images target specific information relevant to the explanation of a concept.

Recent work has addressed some of the challenges associated with multimodal image+text tasks situated in scientific documents. Tan et al. (2022) present a dataset of charts from scientific papers and associated natural language captions summarizing the information present in the chart. Gong et al. (2021) develop a method to link labels with relevant images in patents. Clark and Divvala (2015) introduce a dataset of 150 computer science papers, ground truth labels for the locations of figures, tables, and captions, and an approach to automatically extract this information from PDFs. SCICAP (Hsu et al., 2021) is a much larger dataset of 400,000 figures, their captions, and associated textual references from various scientific papers. The dataset is collected by scraping scientific preprints from arXiv<sup>1</sup>. For each figure, they provide the associated caption and all para-

<sup>1</sup><https://arxiv.org/>



## Generated Dialogue

```
{
  "dialogue":{
    "q1": "What does 'multimodal' mean in this context?",
    "a1": "Multimodal refers to the combination of different types of data. In SCIENCEQA, it means that to answer questions, one must consider information presented in both text and visual formats--like the written description of a scenario and an image illustrating it. This is important because it reflects how we often encounter and interpret multiple forms of information in real life.",
    "q2": "How does the explanation in the dataset help someone answer the question?",
    "a2": "The explanation helps by providing specific details on why the answer is correct. It connects the general concept from the lecture to the particular question. In the example, the lecture explains what a push or pull is, and the explanation tells us that the force applied by the baby is toward her, which means it's a pull. This detailed reasoning helps learners understand the application of concepts.",
    "q3": "Why is it useful for the dataset to cover topics from natural science, social science, and language science?",
    "a3": "Diversity in subjects ensures that the dataset can help develop QA models that are well-rounded and can handle a wide range of real-world problems. By including different sciences, SCIENCEQA encourages models to learn and apply various types of knowledge, which is more challenging but also more reflective of complex human reasoning.",
  }
}
```

Figure 1: Description of the dataset collection process used in CONVERSATIONAL FIGURES. We prompt GPT-4 to generate conversational turns using figures, their captions, and document references.

graphs that mention the figure.

While these datasets address tasks related to multimodal content in scientific documents, they are not conversational in nature. In contrast, cFIGS addresses the shortcomings of prior datasets as a multimodal, multi-turn, conversational dataset grounded in scientific documents. The questions and answers address specific contextual visual information grounded in scientific documents.

## 3 Dataset Creation

### 3.1 Figure Extraction

We obtain scientific figures by leveraging the expanded SCICAP dataset (Hsu et al., 2021)<sup>2</sup>. The dataset consists of approximately 400,000 scientific figures from preprints on arXiv with their accompanying captions, and all paragraphs referring to the image in the document.

### 3.2 Dataset Creation

For each figure, we prompt GPT-4 (OpenAI et al., 2023) with an instruction as well as the caption and any paragraphs referencing the figure. The prompt instructs the model to generate three-turn conversational question-answer pairs while utilizing the provided contextual information included for each figure. Our instruction is:

<sup>2</sup><https://huggingface.co/datasets/CrowdAILab/scicap>

*I will provide you with the figure, the figure’s caption, and the figure’s reference paragraph. Using the provided elements, generate 3 conversational turns in the form of alternating questions and answers. Return the questions in JSON format like so: “dialogue”: {“q1”: “question 1”, “a1”: “answer 1”, “q2”: “question 2”, “a2”: “answer 2”, “q3”: “question 3”, “a3”: “answer 3”}*

The process of utilizing the prompt as well as the figure and textual context to generate the conversational turns in cFIGS is described in Figure 1 for a specific data sample from SCICAP. It is evident from the figure that the questions target both the visual content in the image (“How does the explanation help?”) as well as the textual content (“What does ‘multimodal’ mean in this context?”).

## 4 Next Steps

This paper outlines current work on collecting the CONVERSATIONAL FIGURES(cFIGS) dataset. The first step is to expand the existing pipeline to complete the collection of the entire dataset. Following the collection of the entire dataset, we will present a baseline approach using large vision+language models to address the problem of generating answers to the collected questions. Both the dataset and baseline will be released publicly to encourage further research on the task.

## Acknowledgements

This work was supported by NSF IIS-2112633 and by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

## References

- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. [VQA: Visual Question Answering](#). *arXiv:1505.00468 [cs]*. ArXiv: 1505.00468.
- Christopher Clark and Santosh Divvala. 2015. Extracting Figures, Tables, and Captions from Computer Science Papers.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. [Visual Dialog](#). *arXiv:1611.08669 [cs]*. ArXiv: 1611.08669.
- Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. [Towards Expressive Communication with Internet Memes: A New Multimodal Conversation Dataset and Benchmark](#). *arXiv:2109.01839 [cs]*. ArXiv: 2109.01839.
- Ming Gong, Xin Wei, Diane Oyen, Jian Wu, Martin Gryder, and Liping Yang. 2021. Recognizing Figure Labels in Patents. In *SDU@ AAAI*.
- Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. [ManyModalQA: Modality Disambiguation and QA over Diverse Inputs](#). *arXiv:2001.08034 [cs]*. ArXiv: 2001.08034.
- Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. [SciCap: Generating captions for scientific figures](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. [Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation](#). *arXiv:1701.08251 [cs]*. ArXiv: 1701.08251.
- OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perialman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki

Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. [Gpt-4 technical report](#).

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. [Image Chat: Engaging Grounded Conversations](#). *arXiv:1811.00945 [cs]*. ArXiv: 1811.00945.

Anirudh Sundar and Larry Heck. 2022. [Multimodal conversational AI: A survey of datasets and approaches](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Hao Tan, Chen-Tse Tsai, Yujie He, and Mohit Bansal. 2022. [Scientific Chart Summarization: Datasets and Improved Text Modeling](#). In *SDU@AAAI*.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. [YFCC100M: The New Data in Multimedia Research](#). *Communications of the ACM*, 59(2):64–73. ArXiv: 1503.01817.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. [Visual7W: Grounded Question Answering in Images](#). *arXiv:1511.03416 [cs]*. ArXiv: 1511.03416.