A Descriptive Review of Image Datasets for Accessible Alternative Descriptions in STEM Domains

Marco Cardia
University of Pisa
Pisa, Italy
marco.cardia@di.unipi.it

Marina Buzzi
CNR - IIT
Pisa, Italy
Marina.Buzzi@iit.cnr.it

Giulio Galesi
University of Pisa
Pisa, Italy
giulio.galesi@unipi.it

Barbara Leporini
University of Pisa
Pisa, Italy
CNR - ISTI
Pisa, Italy
barbara.leporini@unipi.it

Abstract

STEM disciplines rely heavily on visual content, such as charts, diagrams, formulas. These figures are often inaccessible to blind or visually impaired users due to the lack of meaningful alternative text. While automated image captioning has progressed, existing datasets are largely oriented toward general images and overlook the structural and semantic complexity of STEM visual contents. This paper presents a descriptive review of publicly available image datasets, evaluating their applicability for generating accessible descriptions of STEM images. Our analysis reveals major gaps: limited support for complex scientific content, shallow annotations, and little consideration for accessibility standards. We argue for the creation of a specialized dataset with rich, structured annotations aligned with accessibility goals. By identifying critical gaps, this work supports the development of AI tools and datasets that enhance inclusive access to STEM content.

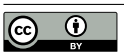
CCS Concepts

• **Human-centered computing** → **Accessibility technologies**.

Keywords

Alternative Descriptions, Alternative Text, Accessible Alternative Descriptions, Accessibility, Visually Impaired Students, Accessible Images, Image Captioning, STEM

ACM Reference Format:
Marco Cardia, Marina Buzzi, Giulio Galesi, and Barbara Leporini. 2025. A Descriptive Review of Image Datasets for Accessible Alternative Descriptions in STEM Domains. In *CHItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), October 06–10, 2025, Salerno, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3750069.3750122



CHItaly 2025, Salerno, Italy

ACM ISBN 979-8-4007-2102-1/25/10
https://doi.org/10.1145/3750069.3750122

1 Introduction

Images are essential in educational contexts as they enhance comprehension and engagement by visually representing complex concepts [27]. This issue is particularly significant in Scientific, Technical, Engineering, and Mathematics (STEM), where diagrams, graphs, formulas, and charts convey data and information. Individuals who are unable to see, such as blind persons, risk losing important content when accessing information or, particularly, when learning. So, this issue becomes critical in the educational field, where students often learn new concepts supported by images equipped with poor or limited textual explanations.

Alternative descriptions for non textual contents, like images, play a crucial role in making contents more inclusive and accessible to people with visual impairments [17, 30]. Adding alternative text (alt text) to images ensures that the information embedded in them is accessible to everyone, including those who rely on screen reading software. Alt text provides visual content into a more inclusive format that can be perceived via voice synthesizer or braille display. This helps in making it possible for individuals with visual impairments to access the same content as sighted users. Thus, according to universal design principles, adding alt text to images makes learning and communication inclusive for everyone. This benefits not only individuals with disabilities but also people temporarily unable to see a screen, such as when traveling or in challenging environmental conditions (e.g., very intense sunlight).

Research explores best practices for preparing effective alternative descriptions, considering factors such as the level of detail needed, the purpose of the image, and the context. Studies suggest that descriptions should vary depending on whether the image is decorative, informative, or complex, such as scientific diagrams or data visualizations [17, 30]. Ensuring that alt text is both concise and meaningful is a challenge in the accessibility field to improve the inclusivity of digital content. Therefore, preparing alternative and narrative descriptions is not an easy task and requires competence and accuracy, especially in the STEM field [21, 22]. The alternative text should be accurate and descriptive while being as concise as possible, to avoid overloading the user, interacting via screen reader, with unnecessary information [17]. Numerous guidelines have been proposed for the web and for preparing graphical content, including

STEM images. However, these guidelines are primarily designed for experts and professionals responsible for formulating descriptions of graphical content and images. Williams et al. (2023) investigate how the accessibility of images is implemented by designers and developers in productive contexts, noting that the perspectives of accessibility practitioners might differ from those of researchers [31]. Leotta et al. (2022) investigate if AI-based services such as Azure Computer Vision Engine, Amazon Rekognition, Cloudsight, and Auto Alt-Text for Google Chrome are suitable for processing images and return textual descriptions (alt-text) in web content, highlighting that none of the analyzed systems is mature enough to replace the human-based preparation of alternative texts, although some tools can generate good descriptions for specific categories of images [17]. Williams et al. (2022) investigated the state of alt text in HCI publications by analyzing 300 figures (including data representations and diagrams), observing that the quality of alt text is highly variable, and nearly half of the figure descriptions contain little helpful information [31]. More guidelines must be developed to address different content types of complex images composed of multiple elements. More recently, generative AI has opened new possibilities for obtaining image descriptions without requiring expert input. However, the fields where generative AI has been most applied appear to be related to photos and images in social or humanities related contexts. The STEM domain, on the other hand, still seems to be underexplored, possibly because it is perceived as more niche or catering to a specialized audience. Studies indicate that while AI-driven image recognition and description models have been widely adopted in mainstream applications, their adaptation to scientific, mathematical, and engineering visuals remains limited [2, 10, 29]. This gap may be due to the complexity of scientific images, which often contain symbolic notations, intricate graphs, and specialized diagrams that require domain-specific understanding to describe meaningfully. Addressing this challenge could significantly enhance accessibility in STEM education and research, making complex scientific content more inclusive for visually impaired users and broader audiences [29, 30]. There are several datasets containing alternative descriptions but few of them are available for STEM images such as charts, plots, graphs, automata, diagrams, trees, and set theory representations. This work provides a descriptive review of existing image datasets with the aim of assessing their suitability for generating accessible alternative descriptions of STEM images. While many datasets have been developed for general image captioning, their applicability to STEM-specific visuals—such as charts, diagrams, and scientific figures—remains underexplored, particularly from an accessibility perspective. By comparing key properties such as annotation type, image content, and descriptive depth, this study highlights the current limitations and identifies opportunities for leveraging or improving available resources in support of inclusive image description. In this context, the study aims to address the following research questions:

(RQ1) Which existing datasets include STEM images accompanied by detailed and accessible textual descriptions that can support the automatic generation of alternative descriptions/texts?

(RQ2) What are the limitations of existing datasets in addressing accessibility needs in STEM education, and what characteristics should an ideal dataset include to overcome these limitations?

2 Related Work

Image description generation can be seen as the intersection of computer vision and natural language processing. It has advanced significantly with deep learning, evolving from handcrafted techniques to sophisticated neural network models (Convolutional Neural Networks, Recurrent Neural Networks) [10]. Sharma et al. (2023) provide a comprehensive survey of this evolution, categorizing methods and highlighting open research issues [29]. The concept of alternative descriptions for images, in the context of accessibility for users with disabilities, has also been explored [29]. Guidelines such as WCAG 2.0 explicitly recommend the use of text alternatives for informative images. These alternatives aim to convey the essential information of the image to users who cannot perceive it visually. Research in accessibility has also considered the use of image-related texts, such as captions and surrounding text, as potential mechanisms for conveying long text alternatives for complex images [30].

A variety of datasets have been developed to support research in automatic image description. These datasets pair images with textual descriptions and vary in size, the format of descriptions, and the collection methods. A variety of datasets have been developed to support research in automatic image description, though most focus on everyday scenes. For example, the Flickr8K/30K datasets [7, 34] pair thousands of web images with Amazon Mechanical Turk (AMT) crowdsourced captions. Other notable resources include the Abstract Scenes dataset with clip-art images [36], the early IAPR-TC12 benchmark [3], the large-scale SBU1M dataset [26], Microsoft Common Objects in Context (MS COCO) with its complex scenes [20], the VizWiz dataset containing images from blind users [4], and the Yelp dataset for business-related images [33]. These foundational datasets, however, do not address the specific complexities of STEM content, providing a foundation for understanding the relationship between general visual data and natural language.

Individuals with visual impairments and blindness (VIB) are underrepresented in STEM careers, partly because these fields rely on visual content that poses accessibility challenges [1, 5, 6]. Providing domain-specific alternative descriptions is crucial, as the absence of effective text for figures in scientific articles creates significant barriers to understanding core concepts. This need is particularly acute in fields like biomedicine and computer science, where visual information is often semantically dense. While detailed descriptions also benefit sighted readers, their primary role is ensuring accessibility. Furthermore, the development of high-quality, STEM-specific alternative descriptions is crucial for advancing automated interpretation and retrieval of information from scientific documents. Generic captions generated by general-purpose models often lack the necessary domain-specific vocabulary and analytical depth required to accurately represent the information contained in STEM figures. Reframing scientific figure captioning as a knowledge-augmented task highlights the importance of incorporating domain-specific knowledge from the surrounding text and within the figures themselves to

generate more informative and contextually relevant descriptions [16, 32].

3 Method

We conducted a comparative analysis of publicly available image datasets to evaluate their suitability for generating accessible descriptions in STEM. Our analysis focuses on datasets pairing images with textual annotations and considers their applicability for blind users.

To identify relevant datasets, we conducted a comprehensive search across major academic and developer platforms. Our search strategy involved querying Google Scholar, arXiv, PapersWithCode, GitHub, and Hugging Face using keywords such as 'STEM image dataset', 'chart captioning', 'scientific figure dataset', 'plot question answering', and 'accessible diagrams'. The initial search yielded over 50 potential resources, which were then screened against our inclusion criteria. We have applied the following inclusion criteria:

- IC1 The dataset contains visual material typical of STEM disciplines (such as charts, diagrams, graphs, formulas, equations, or scientific figures)
- IC2 Each figure includes textual annotations (e.g., captions, labels, summaries, or question–answer pairs), excluding datasets limited to classification or object detection
- IC3 The dataset should be publicly available and documented through sources like arXiv, PapersWithCode, GitHub, or Hugging Face.

The exclusion criteria are:

- EC1 Datasets focused exclusively on natural scenes or general photographs.
- EC2 Datasets with restricted access or with no public documentation.
- EC3 Datasets whose annotations were not in English.

Each selected dataset was analyzed along a set of key dimensions relevant to the accessibility and usability of STEM image content:

- Caption type: identifies the primary form of annotation: captions, question-answer (QA) pairs, classification labels, multiple choice questions (MCQs), mention-paragraphs.
- Annotation method: specifies whether the textual content was generated manually by humans, through amazon mechanical turk (AMT), extracted from the paper caption.
- Image type: classifies the dataset to the predominant image types (e.g., charts, diagrams, graphs, plots, equations, scientific illustrations).
- Dataset source: indicates the origin of the dataset, such as publicly available repositories, institutional archives, web scrapes, or curated collections from academic projects.
- Dataset scale: provides the size of the dataset in terms of number of images.
- License: specifies the legal terms under which the dataset is released. This includes open licenses such as MIT, Apache 2.0, Creative Commons (e.g., CC BY 4.0), or more restrictive licenses like GPL-3.0. The license determines how the dataset can be reused, modified, or redistributed, and impacts its integration into new projects, particularly when combining multiple datasets with different license types.
- Accessibility relevance: assesses the dataset's alignment with accessibility goals, particularly the presence of detailed, structured, or semantically rich descriptions that could support screen reader use.

These dimensions provide a holistic view of each dataset's suitability. 'Caption type' and 'Annotation method' define the textual data's origin; 'Image type' and 'Dataset source' establish the visual domain; 'Dataset scale' addresses practical training needs. Critically, 'Accessibility relevance' assesses alignment with the needs of visually impaired users, moving beyond generic captioning.

In our study, we focus on publicly available datasets that provide STEM images together with their textual annotations. Our goal is to assess the availability of datasets with STEM images with accessible alternative descriptions to support the generation of these. While we aimed for comprehensive coverage, the analysis prioritizes well-documented datasets referenced in academic and developer communities. Resources with restricted access or limited metadata, including descriptions not in english, are not included. We emphasize that our focus is on structured dataset characteristics rather than on direct model performance. Therefore, we do not evaluate image description models per se, but rather the descriptive quality and accessibility readiness of the training data they might consume.

4 Results and discussions

Our analysis of existing datasets relevant to image understanding reveals a diverse range of efforts across domains and tasks, yet also underscores critical gaps in addressing the specific requirements for generating accessible alternative descriptions of complex STEM figures. The datasets identified, presented in Table 1, span various domains and tasks.

To address the first research question (RQ1), we assessed 15 publicly available datasets that pair images with textual annotations and are relevant to STEM domains. These datasets span a variety of tasks, ranging from image captioning and chart summarization to question/answering, and vary widely in scope, content, and accessibility potential. The selected datasets originate from diverse sources including scientific publications (e.g., M-Paper, Multimodal arXiv, SciCap, MMSci), educational materials (e.g., AI2D), statistical platforms (e.g., Chart-to-Text, ChartSumm), and synthetic or crowd-annotated benchmarks (e.g., DVQA, FigureQA). Collectively, they encompass a broad spectrum of visual content types common in STEM communication, such as: bar, line, and pie charts, mathematical equations and tables, schematic illustrations, diagrams, and scientific figures extracted from peer-reviewed articles. A summary of the key characteristics of these reviewed datasets is presented in Table 1. It includes information such as caption type, annotation method, source of the images, image type, dataset scale, and license.

Among the datasets reviewed SciCap+ and Chart-to-Text exhibit more structured and semantically rich annotations that could serve as partial models for accessibility-oriented applications [13, 32]. In terms of scale, the datasets range from smaller collections such as AI2D (5,000 annotated diagrams) to massive repositories like SciCap (>2 million figure-caption pairs), reflecting the trade-off between breadth and annotation depth. Notably, large-scale datasets often prioritize quantity over descriptive granularity, while smaller or

Table 1: Properties of datasets relevant for the description of images related to the STEM fields.

Name, Reference	Caption type	Annotation method	Image type	Dataset source	Dataset scale	License
FigureQA, Kahou et al. 2017 [12]	Question Answering: Yes/No	Template based	Charts	Synthetic	140,000	MIT
PlotQA, Methani et al. 2020 [24]	Question Answering	AMT	Charts	Real world source	224,377	MIT
DVQA, Kafle et al. 2018 [11]	Question Answering	Template based	Bar charts	Synthetic	300,000	CC by 4.0
ChartQA, Masry et al. 2022 [23]	Question Answering	AMT, generated questions	Charts	Statista, Pew Research, OWID, OECD	20,882	GPL-3.0
Chart-to-Text, Kantharaj et al. 2022 [13]	Captions	Human-written captions	Charts	Statista and Pew Research	44,096	GPL-3.0
Chart2Text, Obeid et Hoque 2020 [25]	Captions	Human-written captions	Charts	Statista	8,305	GPL-3.0
ChartSumm, Rahman et al. 2023 [28]	Captions	Human-written captions, generated summaries	Charts	Knoema and Statista	84,363	CC by 4.0
AutoChart, Zhu et al. 2021 [35]	Captions from template	Template generated summary	Charts	World Bank Open Data, Nutritional Analysis Data	10,232	CC by 4.0
SciCap, Hsu et al. 2021 [8]	Captions	Captions from original article	Graph plots, tables, equations, flowcharts	Computer science arXiv papers	2,170,719	CC by 4.0
SciCap+, Yang et al. 2024 [32]	Captions	Captions from original article, mention-paragraphs	Plots, tables, equations, flowcharts	Computer science arXiv papers	414,809	CC by 4.0
AI2D, Kembhavi et al. 2016 [15]	MCQs	AMT	Diagrams	Google Image Search	5,000	CC by 4.0
MMSci, Li et al. 2024b [19]	Captions, MCQs	Captions from original article / (MCQs are generated)	Charts, diagrams, microscopic images	Nature Communications	742,273	CC by 4.0
M-Paper, Hu et al. 2024 [9]	Captions	Captions from original paper	Chart, diagrams and tables	LaTeX source files of arXiv papers	343,546	Apache 2.0
Multimodal ArXiv, Li et al. 2024a [18]	Captions and QAs	Extracted from original article	Geometry shapes, and plots	arXiv academic papers	6,400,000	CC by 4.0
ACL Fig, Karishma et al. 2023 [14]	Labels (for classification)	Captions and mention-paragraph	Charts, Graphs, Trees, Venn diagram	ACL Anthology	112,052	CC by 4.0

manually curated datasets sometimes offer more domain-specific or linguistically enriched annotations.

To assess the suitability of the analyzed datasets for generating alternative descriptions of STEM images, we evaluated each dataset across seven core dimensions introduced in our methodology: caption type, annotation method, image type, dataset source, dataset scale, license, and accessibility relevance. A majority of datasets provide short, flat captions typically one sentence descriptions. While appropriate for general-purpose captioning, these annotations often lack the structure needed to describe relationships in STEM figures. Exceptions include datasets like Chart-to-Text, which provide longer human-written summaries, and SciCap+, which offers figure captions extracted from scientific articles alongside contextual text. However, none of these dataset fully deals with the description of images specific for the STEM contexts, such as mathematical notations, formulas, circuit diagrams, automata, flowcharts, graphs, and trees. Annotation methodologies varied significantly. Crowdsourced approaches, specifically exploiting Amazon Mechanical Turk, are used in datasets such as PlotQA, ChartQA, AI2D. Captions extracted from the original paper are used in ACLFig, Multimodal Arxiv, MMSci, M-Paper, SciCap. Instead, SciCap+ included captions extracted from the original paper together with the description available in the text. Notably, none of the reviewed datasets document the involvement of users with disabilities or accessibility experts in their annotation workflows, a critical shortcoming for ensuring practical usability in assistive contexts. Furthermore, in

most cases, the captions accompanying images are extracted directly from original scientific papers and were not authored with accessibility for visually impaired users in mind. These captions often serve communicative or referential purposes for sighted readers, lacking the structured detail and descriptive clarity typically required for effective screen reader interpretation. Dataset sources varied from academic repositories (e.g., arXiv, ACL Anthology) to commercial platforms (e.g., Statista, Knoema) and synthetic generators. This variation impacts annotation quality, as synthetic or extracted sources often lack the context required for accessibility. Finally, the heterogeneity of data licenses visible in Table 1, ranging from permissive (e.g., MIT, CC BY 4.0) to restrictive copyleft (GPL-3.0), creates a significant but often overlooked barrier to combining these valuable resources. Table 2 provides a comparative overview of the accessibility focus stated in the design of each dataset. As shown, only a few datasets, most notably SciCap, SciCap+, Chart-to-Text, and ChartSumm, include accessibility in their stated motivations. Chart2Text also references accessibility as a future direction. For the majority of datasets (11/17), accessibility is either entirely absent from the design rationale or only indirectly addressed through downstream tasks such as captioning or summarization. This highlights a substantial gap in the current landscape and underscores the need for datasets that are built to support inclusive image description in STEM domains.

Despite broad coverage of graph types (bar, line, pie), few datasets contain symbolic, mathematical, or logic-based images, such as automata, graphs, trees, or annotated code diagrams, key areas in computer science and STEM education. The number of figures for each dataset ranged from fewer than 5,000 images (AI2D) to well over two million (SciCap, Multimodal Arxiv). Larger datasets generally offered broader visual diversity but often at the expense of annotation detail and quality. Smaller, curated datasets tended to feature more context-aware or semantically meaningful annotations, though their size may limit model generalization.

4.1 Limitation in STEM context

In response to RQ2, we observe that despite the valuable contributions of the identified datasets, significant limitations remain when it comes to generating high-quality, accessible alternative descriptions for STEM images. One major challenge is the lack of domain alignment in many foundational datasets. While resources built on general web images have driven progress in basic captioning, their content and associated descriptions lack the specific vocabulary, symbolic notation (e.g., mathematical equations), and conceptual depth required to accurately represent complex scientific diagrams, plots, or schematics. Even within datasets explicitly focusing on STEM content, such as SciCap, M-Paper, or resources focused on chats like ChartQA and PlotQA, a critical task mismatch often exists. These datasets primarily target caption generation, typically concise summaries highlighting key findings, or QA, which retrieves specific factual details. Neither task directly aligns with the goal of creating comprehensive alternative text for accessibility. Effective alt text often requires a more detailed, structured description of the visual elements, their relationships, and the data presented (e.g., describing axes, data series, trends in a plot, or components and connections in a diagram or in a graph), going beyond a simple summary or a single factual answer. Consequently, the annotation depth and structure within these datasets frequently prove insufficient. The provided captions or summaries are often single, relatively short sentences and may not capture the relational structure presented in many STEM visual contents. Collectively, these limitations concerning domain relevance, task definition, annotation granularity, and modality mean that existing datasets, while advancing visual understanding in general, provide an inadequate foundation for training robust models capable of generating truly effective and accessible alternative descriptions for the nuances of STEM visual contents. The diversity of data licenses, as shown in Table 1, presents another barrier to progress. While the availability of datasets under permissive licenses like MIT and Creative Commons is encouraging for reusability, the mix with more restrictive copyleft licenses such as GPL-3.0 complicates the landscape. This licensing fragmentation makes it legally and practically difficult to combine different datasets, hindering the creation of larger, more comprehensive resources required to train effective and widely applicable models for STEM accessibility.

4.2 Need for a STEM specific dataset

The limitations discussed so far highlight the need for the development of a dataset specifically designed and annotated for generating accessible alternative descriptions of STEM images. Existing resources, while advancing general computer vision and multimodal understanding, are insufficient because they were not designed with the primary goal of enabling equitable access to complex scientific visuals for users with visual impairments. Overcoming the identified gaps in domain focus, task definition (captioning/QA vs. structured description), annotation depth, and modality is essential for progress in this area. This need is further amplified by increasing normative and regulatory requirements globally. Accessibility standards, such as the Web Content Accessibility Guidelines (WCAG), and legislation like Section 508 in the US or the European Accessibility Act, mandate that digital content, including scientific publications and educational materials, be accessible. Currently, generating compliant and truly useful alternative text for intricate STEM visual contents often requires significant manual effort from domain experts, a process that is costly and does not scale. Automated tools often fail on such content precisely because they lack training data that reflects the complexity of the visuals and the requirements of a good description. A dedicated dataset, annotated according to accessibility best practices, is therefore crucial for developing AI models capable of generating descriptions that meet both user needs and legal obligations. Furthermore, the benefits of such a dataset extend beyond generating alt text for static images in documents or web pages. Enhancing AI's ability to understand and describe complex STEM visual contents has implications for accessibility across broader STEM environments. Consider, for example, virtual laboratories, interactive simulations, data analysis software, or assistive technologies for physical lab work. These environments rely heavily on visual interaction. Models trained on a dataset rich in descriptions of STEM components, relationships, processes, and data representations could form the foundation for tools that provide real-time descriptions, or alternative representations in these more dynamic contexts. Improving the descriptive

Table 2: Stated Accessibility Relevance of Reviewed Datasets. While several datasets can indirectly support accessibility tasks such as caption generation, few are explicitly designed with accessibility for visually impaired users as a primary objective.

Dataset	Accessibility Relevance
FigureQA [12]	Designed for visual reasoning grounded in synthetic figures. Authors do not explicitly state accessibility as a goal.
PlotQA [24]	Developed for reasoning over scientific plots via Q&A using real-world data. Authors do not explicitly state accessibility as a goal.
DVQA [11]	A synthetic dataset for understanding bar charts via Q&A, focusing on structure and data retrieval. Authors do not explicitly state accessibility as a goal.
ChartQA [23]	A benchmark for Chart Q&A focusing on visual/logical reasoning using real-world charts. Authors do not explicitly state accessibility as a goal.
Chart-to-Text [13]	The Chart-to-Text dataset serves as a large-scale benchmark for chart summarization. Authors explicitly state that automatic chart summarization offers an important benefit of making charts more accessible to visually impaired people.
Chart2Text [25]	Dataset for chart summarization. Future goals include enhancing accessibility of charts for visually impaired people via an interactive chart summarization system.
ChartSumm [28]	Large-scale benchmark for automatic chart summarization. Automatic chart-to-text summarization is explicitly stated as effective for visually impaired people.
AutoChart [35]	Large dataset for analytical description of charts generated automatically. Authors do not explicitly state accessibility as a goal.
SciCap [8]	Dataset for scientific figure captioning. Motivation includes making scientific figures more accessible to blind or visually impaired readers.
SciCap+ [32]	Augmented version of SciCap. Extends the dataset supporting the original motivation, which includes making scientific figures more accessible to visually impaired readers.
AI2D [15]	Dataset of diagrams for grade school science. Used for creating graph diagrams in another dataset. Authors do not explicitly state accessibility as a goal.
MMSci [19]	Dataset from scientific articles for multimodal scientific understanding. Contains figure-caption pairs which can support accessibility tasks like captioning, but authors do not explicitly state accessibility as a primary goal.
M-Paper [9]	Dataset collecting figure-caption pairs. Contains figure-caption and table-caption pairs which can support accessibility tasks like captioning and analysis, but authors do not explicitly state accessibility as a primary goal.
Multimodal ArXiv [18]	Derived from arXiv papers for scientific comprehension. ArXivCap contains figure-caption pairs, which can support accessibility tasks, but authors do not explicitly state accessibility as a primary goal.
ACL Fig [14]	Dataset for scientific figure classification. Supports tasks like classification, QA, and auto-captioning. Auto-captioning can support accessibility, but authors do not explicitly state accessibility as a primary goal.

capabilities for static images builds a fundamental visual-semantic understanding applicable to these wider accessibility challenges within STEM education and practice.

4.2.1 Characteristics of an ideal dataset: Key requirements for alternative descriptions in STEM images. Based on our critical analysis, we outline a set of essential characteristics that an ideal dataset should possess to effectively support the development of accessible alternative descriptions for STEM images. These requirements are informed by both accessibility best practices and the specific challenges of describing complex scientific visuals. An ideal dataset must include a wide range of image types commonly found in STEM disciplines, such as: charts (bar, line, scatter, pie), mathematical notations and formulas, circuit diagrams and automata, flowcharts, graphs, trees, and structural representations. This diversity ensures that models trained on such data can generalize across disciplines and support varied accessibility needs. Effective alternative descriptions should go beyond short captions. An ideal alternative description should provide:

- Structured descriptions that decompose visual content, a physical description of the image.
- Semantic description, offering context about what the image represents and why it matters.
- Concise and informative phrasing suited for screen reader output, avoiding unnecessary verbosity.

Additionally, all annotations should conform to accessibility standards to ensure compatibility with assistive technologies. To promote reproducibility and enable broad community use, the dataset should be publicly available under an open license and accompanied by comprehensive usage guidelines and documentation. These requirements synthesize our findings and provide a direct response to RQ2 regarding the essential features of a dataset designed to support accessible image descriptions in STEM domains.

5 Conclusion and future work

Our descriptive review of public datasets, evaluated across seven key dimensions, provides a comprehensive view of their suitability for generating accessible descriptions of STEM visuals. We found that current datasets largely fall short of supporting accessibility needs. In response to RQ1, we identified a subset of datasets, such as SciCap+ and Chart-to-Text, that show promise but ultimately fall short of fully supporting accessibility in STEM contexts. To address RQ2, we outlined key limitations in current datasets and proposed a set of essential characteristics for a purpose-built resource that meets both accessibility and domain-specific requirements. Our analysis reveals a clear gap in the current landscape: while many datasets support general image captioning and scientific visualization, few are designed with accessibility as a primary objective. Most lack the semantic detail, structured annotations, and alignment with accessibility standards necessary to serve users with

visual impairments in the STEM domain. These limitations highlight the need for a dedicated dataset that includes diverse STEM visual contents and provides multi-level, structured, and semantically rich descriptions created with accessibility in mind. Such a dataset would serve not only as a benchmark for evaluating AI models but also as a foundational resource for developing inclusive educational and research tools.

Future work will focus on the design and development of this specialized dataset. This includes identifying representative STEM image types, defining annotation schemas aligned with accessibility standards (e.g., WCAG), and involving domain experts and users with disabilities in the annotation process. In parallel, we plan to evaluate current AI models on their ability to generate meaningful alternative descriptions using this dataset, and to explore hybrid human-AI pipelines to improve annotation quality and scalability. Ultimately, this work aims to bridge the gap between advances in AI and the practical needs of inclusive STEM education and research.

Acknowledgments

This work was carried out within the STEMMA PRIN project thanks to the funding by the European Union - Next Generation EU, Mission 4 Component 2 CUP B53D23019500006.

References

[1] Karla Antonelli, Anne Steverson, and Jamie O'Mally. 2018. College graduates with visual impairments: A report on seeking and finding employment. *Journal of Visual Impairment & Blindness* 112, 1 (2018), 33–45.

[2] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55 (2016), 409–442.

[3] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, Vol. 2.

[4] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 417–434.

[5] Jesse R Hairston, Derrick W Smith, Tania Williams, William T Sabados, and Steven Forney. 2020. Teaching cybersecurity to students with visual impairments and blindness. *Journal of Science Education for Students with Disabilities* 23, 1 (2020), n1.

[6] Cicely Hayes and Michael J Proulx. 2024. Turning a blind eye? Removing barriers to science and mathematics education for students with visual impairments. *British Journal of Visual Impairment* 42, 2 (2024), 544–556.

[7] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[8] T.Y. Hsu, C.L. Giles, and T.H. Huang. 2021. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624* (2021).

[9] Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[10] Shubh Jain, Siddhant Zawar, Yash Rupchandani, and MA Chimanna. 2024. Image Description Generation using Deep Learning: A Comprehensive Overview. In *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*. IEEE, 1–9.

[11] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5648–5656.

[12] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300* (2017).

[13] Siddhartha Kantharaj, R.T. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4005–4023.

[14] Z. Karishma, S. Rohatgi, K.S. Puranik, J. Wu, and C.L. Giles. 2023. Acl-fig: A dataset for scientific figure classification. https://arxiv.org/abs/2301.12293.

[15] Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*, Vol. 14. Springer, 235–251.

[16] Jaeyoung Kim, Jongho Lee, Hong-Jun Choi, Ting-Yao Hsu, Chieh-Yang Huang, Sungchul Kim, Ryan Rossi, Tong Yu, Clyde Lee Giles, Ting-Hao'Kenneth' Huang, et al. 2025. Multi-LLM Collaborative Caption Generation in Scientific Documents. *arXiv preprint arXiv:2501.02552* (2025).

[17] M. Leotta, F. Mori, and M. Ribaudo. 2023. Evaluating the effectiveness of automatic image captioning for web accessibility. *Universal Access in the Information Society* 22, 4 (2023), 1293–1313.

[18] L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong, and Q. Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 14369–14387.

[19] Z. Li, X. Yang, K. Choi, W. Zhu, R. Hsieh, H. Kim, J.H. Lim, S. Ji, B. Lee, X. Yan, et al. 2024. MMSCI: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903* (2024).

[20] 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*. Springer, 740–755.

[21] Alex Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1073–1083.

[22] Katie Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing tools for high-quality alt text authoring. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[23] Ahmed Masry, Dongxu Long, Jiaqi Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244* (2022).

[24] N. Methani, P. Ganguly, M.M. Khapra, and P. Kumar. 2020. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1527–1536.

[25] Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142* (2020).

[26] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* 24 (2011).

[27] M. Pateşan, A. Balagiu, and C. Alibec. 2018. Visual aids in language education. In *International Conference Knowledge-Based Organization*, Vol. 24. 356–361.

[28] R. Rahman, R. Hasan, A. Al Farhad, M.T.R. Laskar, M.H. Ashmafee, and A.R.M. Kamal. 2023. ChartSumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. In *Canadian Conference on Artificial Intelligence*.

[29] Himanshu Sharma and Devanand Padha. 2023. A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artificial Intelligence Review* 56, 11 (2023), 13619–13661.

[30] Bruno Splendiani. 2015. A proposal for the inclusion of accessibility criteria in the authoring workflow of images for scientific articles.

[31] C. Williams, L. de Greef, E. Harris III, L. Findlater, A. Pavel, and C. Bennett. 2022. Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference*. 1–12.

[32] Z. Yang, R. Dabre, H. Tanaka, and N. Okazaki. 2024. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing* 31, 3 (2024), 1140–1165.

[33] Yelp Inc. 2014. Yelp Dataset Challenge. http://www.yelp.com/dataset_challenge. Accessed: 21 April 2025.

[34] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics* 2 (2014), 67–78.

[35] J. Zhu, J. Ran, R.K.W. Lee, K. Choo, and Z. Li. 2021. AutoChart: A dataset for chart-to-text generation task. *arXiv preprint arXiv:2108.06897* (2021).

[36] C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3009–3016.