FigurA11y: AI Assistance for Writing Scientific Alt Text

Nikhil Singh*
nsingh1@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Lucy Lu Wang
lucylw@uw.edu
University of Washington & Allen Institute for AI
Seattle, WA, USA

Jonathan Bragg
jbragg@allenai.org
Allen Institute for AI
Seattle, WA, USA

ABSTRACT

High-quality alt text is crucial for making scientific figures accessible to blind and low-vision readers. Crafting complete, accurate alt text is challenging even for domain experts, as published figures often depict complex visual information and readers have varied informational needs. These challenges, along with high diversity in figure types and domain-specific details, also limit the usefulness of fully automated approaches. Consequently, the prevalence of high-quality alt text is very low in scientific papers today. We investigate whether and how human-AI collaborative editing systems can help address the difficulty of writing high-quality alt text for complex scientific figures. We present FigurA11y, an interactive system that generates draft alt text and provides suggestions for author revisions using a pipeline driven by extracted figure and paper metadata. We test two versions, motivated by prior work on visual accessibility and writing support. The base **Draft+Revise** version provides authors with an automatically generated draft description to revise, along with extracted figure metadata and figure-specific alt text guidelines to support the revision process. The full **Interactive Assistance** version further adds contextualized suggestions: text snippets to iteratively produce descriptions, and hypothetical user questions with possible answers to reveal potential ambiguities and resolutions. In a study of authors (N=14), we found the system assisted them in efficiently producing descriptive alt text. Generated drafts and interface elements enabled authors to quickly initiate and edit detailed descriptions. Additionally, interactive suggestions from the full system prompted more iteration and highlighted aspects for authors to consider, resulting in greater deviation from the drafts without increased average cognitive load or manual effort.

CCS CONCEPTS

• **Human-centered computing → Accessibility systems and tools**; **Human computer interaction (HCI)**; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Accessibility, Alt text, Image descriptions, Scientific figures, Human-AI interaction, Natural language generation, Large language models, Writing assistance systems

1 INTRODUCTION

Digital dissemination has allowed scientific authors to reach broad audiences with their work. However, one audience that continues to face barriers is blind and low-vision (BLV) readers. BLV individuals typically rely on alternative text (alt text) descriptions to access key data and concepts communicated visually in figures. Distinct from and complementary to captions, which typically provide figure context or commentary, alt text descriptions convey figure *content* including that which may be visually apparent.

Despite its vital role, alt text is often absent or of inadequate quality in scientific papers [4, 34]. One key reason for this that has been elucidated in prior work is that authors face challenges in producing high-quality descriptions [46]. Scientific figures can depict intricate concepts and relationships through numerous visual encodings, making translation to textual descriptions cumbersome. Authors must determine which aspects to describe and how to adequately convey them. Additionally, accurate and complete alt text requires deep domain knowledge and contextual insight into the figure's interpretation. This can make it challenging for non-authors and/or automated systems lacking such expertise to replace or supplement author effort. As such, insufficiently detailed or even entirely lacking descriptions remain prevalent.

Guidelines designed to assist authors in writing effective alt text[1,2] are often narrow in scope as they focus on specific figure types like line or bar plots, or tree diagrams. This makes it difficult for authors to extend their principles more broadly, such as to compound figures [46]. Similarly, though fully automated approaches are rapidly improving in quality, they are often also constrained to specific types of figures such as plots or natural images, limiting applicability to scientific communication more broadly which often involve complex diagrams and multi-part scientific figures [12, 44, 48]. Model errors also risk creating misinformative alt text if authors or publishers over-rely on automated methods. Despite advances in computer vision and natural language processing for processing and describing many types of images [15, 26, 55], scientific figures often contain nuanced details and contextual factors that might hamper the applicability of these models to producing high-quality accessible descriptions. As such, it is important to effectively engage authors, equipping them to critically review and refine auto-generated descriptions.

More specifically, there is a need for methods that generalize across authors' open-domain figures while providing tailored guidance within interactive alt text drafting workflows. To inform the design of such an interactive alt text authoring system, we first conducted a formative study with authors (N=6). This study used an initial prototype which provided authors suggestions during drafting, using large language models conditioned on figure metadata. The study revealed needs for more guidance during drafting and increased control over text generation.

Based on these findings, we developed an interactive system for alt text authoring[3] that combines human and AI capabilities, with

[*] Work completed during an internship at Semantic Scholar, Allen Institute for AI.

[1] http://diagramcenter.org/table-of-contents-2.html
[2] http://diagramcenter.org/poet.html
[3] Code available at https://github.com/allenai/figura11y

specific features detailed below. Authors upload their paper into the system, which then automatically extracts figures along with corresponding captions, mentioning paragraphs, figure text, and an estimated data table (for plots). This establishes a knowledge base to use for AI suggestions. Subsequently, a suite of drafting and editing features ranging from pre-generated drafts to iterative description snippet generation and queries to elicit author input are provided to assist them in efficiently producing detailed alt text. We conducted a within-subjects user study to evaluate our system (N=14), which, to our knowledge, is the most realistic and general study of AI-assisted alt text authoring to date, with authors writing descriptions for their own figures across a diverse set of figures and fields of study.

Overall, our work contributes:

(1) A formative study (N=6) with authors, using a technology probe which offered alt text writing suggestions. This revealed needs for guidance, control, and varied suggestions.
(2) An automated pipeline to generate descriptive draft alt text for open-domain figures without requiring ground truth data. It uses an ensemble of methods to extract metadata, assembles knowledge-based prompts, and uses large language models for generation. In contrast to prior work, this is training- and data-free, fast, generalizes to arbitrary figure types, doesn't require ground truth data or scene graphs, and allows us to incorporate existing accessibility guidelines, all necessary features for real-world alt text applicability.
(3) An interactive alt text authoring system which (A) scaffolds alt text production by providing extracted paper and figure context with figure type-specific accessibility guidelines to support reviewing and revising generated drafts, and (B) two additional features: *Generate at Cursor*, which interactively expands descriptions at user-directed points based on writing support approaches, and *Potential User Questions* (and suggested answers), which prompt authors to address ambiguous elements following from prior work using queries and templates.
(4) A within-subjects user study (N=14) where authors described their own figures, mimicking real-world use. Findings show the system assisted rapid drafting and editing of descriptive alt text through different strategies based on author needs. Interactive features enhanced experience without increased cognitive load or effort on average, and enabled greater deviation from generated drafts by supporting iterative refinement.

2 RELATED WORK

2.1 Figure Accessibility in Scientific Publishing

A systematic analysis of alt text practices across scientific disciplines has yet to be conducted. However, smaller-scale studies highlight significant issues with low alt text prevalence. For example, only 4.6% of figures had valid alt text in a sample of Accessibility and HCI papers [4], despite explicit author guidelines from venues in these fields. Even lower prevalence has been observed in other fields like biomedicine: an examination of recent papers from 16 leading biomedical and ophthalmology journals found no meaningful alt text beyond basic information [34]. While these results highlight the need for improvements, more work is required to characterize issues in figure accessibility across domains. Still, in response to the lack of quality alt text, we aim to address a broad set of scientific figures without an explicit domain or type constraint.

Beyond prevalence, the quality of alt text is also important to consider. Web accessibility guidelines suggest that alt text should convey the same information or function as visual content[4]. Scientific figures are information-dense, making full coverage of relevant information difficult to judge. Rubrics have been proposed to assess the descriptiveness and structure of alt text content. Williams et al. [46] developed a rubric to assess the overall descriptiveness of figures in HCI papers, building on prior work for other types of images [11]. Lundgard and Satyaranayanan proposed an influential four-level semantic model of descriptions for data-driven figures like plots [29]. This model decomposes the descriptions of such figures into *elemental and encoded*, *statistical and relational*, *perceptual and cognitive*, and *contextual and domain-specific* content. We factor this semantic model into our system, in order to steer language models towards generating structured, meaningful descriptions and suggestions.

2.2 Author Challenges in Alt Text Writing

One well-documented reason for inadequate alt text is that authors face challenges in effectively describing figures. Interviews by Williams et al. [46] reveal that their author participants were confused about what information to include in the alt text (or, as one participant put it, "what's missing" beyond the figure caption). Their results point out that interviewed authors wanted advice on the structure and content of their descriptions, given the density of visual elements and relationships depicted in their figures.

Guidelines are a traditional mechanism by which authors have previously been supported in writing alt text. For example, SIGACCESS provides guidelines for computing publications[5], the American Chemical Society (ACS) provides guidelines for ACS authors[6], and the multidisciplinary publisher *Taylor & Francis* provides guidelines for authors submitting to their journals[7], among others. However, guidelines are often based on example figures. Authors must interpret such guidelines and adapt them to their own figures and even figure types. Additionally, the content of guidelines can be difficult to interpret. For instance, such guidelines often emphasize brevity, but this can come at the expense of accessibility, especially for complex figures as Williams et al. also note. In our system for supporting authors, we leverage guidelines to generate figure-specific drafts and suggestions.

2.3 Automated Image Description Generation

Early work in automated image description, often associated with computer vision, typically relied on detecting objects and relations or constructing patterned templates [9, 22]. While these initial

[4] https://www.w3.org/WAI/tutorials/images/
[5] https://www.sigaccess.org/welcome-to-sigaccess/resources/describing-figures/
[6] https://pubs.acs.org/doi/full/10.1021/acsguide.60108
[7] https://www.tandfonline.com/pb-assets/tandf/authors/tf-alt-text-guide-1636994956097.pdf

approaches produced reasonable descriptions for constrained domains of images, they were limited in flexibility and language quality. Additionally, figures frequently exhibit higher complexity than natural images. Recent work has explored automated figure captioning [2, 16], including doing so with knowledge-augmentation in the form of mentioning paragraphs and OCR text [50]. We use a similar knowledge-augmented approach in conjunction with alt text guidelines and zero-shot large language model inference to achieve broad applicability for figures without training, especially since datasets of high-quality alt text for diverse scientific figures are not readily available.

More recently, large language and multimodal models have been used to generate improved image descriptions [45, 55]. Language models can produce varied, high-quality text conditioned on input information, making them useful for this task. In multimodal models, this input can also often be visual [1, 25, 27, 40], allowing direct input of figure information. However, scaling and providing interactive access to cutting-edge multimodal models remains challenging due to computational demands and rapid changes in their capabilities. Additionally, these models' visual capabilities are still error-prone and often not evaluated on tasks as complex as describing figures in scientific research.

An emerging solution for improving vision-language reasoning is to decompose the task into vision and reasoning components through a number of different strategies [14, 43, 47, 51, 52, 55]. This can allow using separate specialized models for each part. For example, dedicated vision models can efficiently handle image information extraction as a frontend, while language models can focus on reasoning over these visual features. This has been done for general images, but also leveraged for tasks like question-answering based on plots and charts [26]. Prior work has also shown promise in generating descriptions for data visualizations from metadata alone. VisText [44] produces descriptions for plots based on data tables and scene graphs available during visualization design. Interestingly, this work's experiments found visual inputs did not improve over metadata-only methods. For open-domain figures, visual information could still be advantageous. However, these results demonstrate language models can perform well (in the plot domains covered by their models) given sufficient contextual information.

Our methods stem from a similar motivation, but we tailor them to open-domain scientific figures without original data or metadata available. Since describing figures is knowledge-intensive, we look beyond just visual features to extract contextual information from paper text and writing guidelines. Integrating this knowledge aims to assist language models in producing high-quality, tailored alt text suggestions by providing critical context beyond what is visually evident. Overall, our approach selectively combines strengths of language models, computer vision models, knowledge extraction methods, and human input to provide robust assistance for authoring accessible figure descriptions.

It is important to note, however, the rapid advances occurring in multimodal models. Future vision-language models might well provide strong automated generation capabilities. However, we believe supplementary information and human interaction will remain valuable. Extracted paper content can provide essential contextual knowledge beyond visual inputs, both for generation and revision. Human workflows also enable assessing accuracy, eliciting additional details, evaluating coverage sufficiency, and customization.

2.4 Alt Text Writing Support

Crowdsourcing has been identified as one viable avenue for communicating visual information to blind and low vision (BLV) users. VizWiz is an influential platform that leverages crowdsourcing to describe visual content in real-time [3]. However, extending this paradigm to scientific figures poses challenges: describing figures often requires additional domain knowledge, as well as added effort to ensure accuracy and detail. Other recent work has explored how crowdsourcing can be combined with other strategies, including automation and retrieval, to generate alt text for images on Twitter [12]. Like this work, we rely on a human-in-the-loop, specifically an author, and propose a suite of features to allow figure-specific description workflows. We introduce a collaborative AI-based system in order to distribute the workload of producing detailed alt text for complex figures. Rather than asking crowd-workers to acquire sufficient knowledge of the figure, we represent extracted knowledge as a structured prompt for a language model which can rapidly create content for the author to evaluate and incorporate.

Work targeted to author support has explored templates and queries. Morash et al. [33] explored the use of templates to elicit information from non-specialists in order to produce effective alt text. They found that this *queried image description* (QID) approach resulted in improved results compared to *free response image description*. Mack et al. [30] observed that templates helped authors write better alt text compared with automatically generated options, which were brief and regarded by authors as a gold standard, leading to reduced final quality. Text generation has made significant strides in recent years, however, which result in generated descriptions no longer being limited to brief and general content. Further, templates require per-image crafting. We generalize queries into our *Potential User Questions* feature which leverages text generation to elicit author input on possible ambiguities. These questions are also motivated by VizWiz's approach, which treats questions and answers as a mechanism for making images non-visually accessible.

2.5 Language Models for Writing and Editing Support

Large language models (LLMs) have recently shown promise in providing contextual suggestions for diverse writing and revision tasks [8, 10, 23]. A common application is to open-ended tasks such as creative writing, which allow for wide-ranging suggestions useful for inspiration [32, 42, 53]. In contrast, alt text requires faithfulness to the source visual information. It has aspects in common with *expository* writing tasks [41], requiring steps such as reasoning over and synthesizing information, and facilitating composition. Our approach aims at these components in the specific case of alt text writing. Extracted information provides a knowledge base for faithful generation. Refinement interactions support accuracy verification and content enhancement. Together, these aim to leverage the capabilities of advanced LLMs to assist authors in efficiently producing high-quality, accessible figure descriptions.

3 FORMATIVE STUDY AND TOOL DESIGN

3.1 Initial Prototype

We created a high-fidelity interactive prototype, serving as a technology probe, containing early versions of two key features designed based on reviewing prior work and proposing methods to generalize across open-domain figures: text continuations and question-answer pairs. The continuations appended generated text conditioned on figure metadata, a common strategy in writing support which allows suggestions that build on user-authored text [35, 42, 53]. The question-answer pairs were motivated by *queried image description* [33] wherein authors were provided a predetermined series of guideline-derived questions depending on figure type. With this feature, we aimed to highlight elements the author might consider describing for the figure.

3.2 Formative Study

We conducted a formative study using this initial prototype with paper authors (N=6) to inform the design of our main system. Participants had varying levels of experience with authoring alt text, ranging from none to over 5 years of experience. We used a think-aloud protocol and semi-structured interviews during 45-minute remote sessions conducted via video-conferencing. Participants were provided with access to our prototype. We asked participating authors to verbalize their thought processes while trying out these features on their own figures. Session audio and screencast were recorded and transcribed. One member of the research team performed inductive thematic analysis on the data via open coding, guided by discussion with the full research team.

3.3 Feedback and System Redesign

Through analyzing and interpreting participant feedback, we identified key design goals for improvement:

DG1 More guidance during the drafting process, such as feedback to ensure authors provide sufficient coverage of key information in their descriptions. For example, one participant (P4) suggested the system could provide a hypothetical question like *“you didn’t actually mention anything about the axes, do you want to do that?”* to prompt the author to describe missing elements.

DG2 Increased control over where and how much automatic text generation occurs within the description, e.g. targeted expansions of specific parts based on author needs. For instance, P6 suggested *“what I would really like is something… where I can put my cursor somewhere and say get continuation from here.”* P3 proposed a similar interaction: *“Something that could be cool is if I could highlight something and then say generate more about this.”*

Some participants also noted that the two suggestion types (continuations and Q&A pairs) could emphasize similar information, though in other cases participants found both independently useful. This highlighted an opportunity for differentiating the two to provide complementary guidance. For example, P5 noted: *“a continuation I could see including some information that I might not have thought would be relevant… the [Q&A pairs], I could also see that working in a similar way,”* while P1 noted that *“They felt useful for different things”* like the continuation helping to create an *“outline or skeleton.”* Based on this feedback, we concluded that differentiating these suggestions would help offer both benefits, i.e., interactively creating outlines and highlighting missed or ambiguous content.

In response, we implemented modified versions of the original features to provide potential user questions and on-demand text generation at user-selected points which we call *Potential User Questions* and *Generate at Cursor* respectively, described in detail in the following section. To compare with, we also implemented a simplified interface without these two suggestion features. In summary, author feedback highlighted needs for improved guidance and disambiguating suggestion types. Our redesign addressed these by providing targeted author feedback and control over text generation.

Additionally, our pilot interface positioned metadata in a menu, and used this information to prompt for suggestions. However, we observed participants referencing this metadata to get context for beginning, editing, and evaluating their descriptions and the system’s suggestions. To account for this usage, we moved the metadata into the main interface so the user can easily cross-reference as needed.

3.4 Improving the AI Assistance

In addition to improving the features, we also sought to improve the quality of the AI assistance provided. We iterated on model choice and prompts by optimizing on a development set of scientific figures and captions.

Figure Sampling for Development Set. We constructed a development set of figures and metadata to iterate on the AI assistance. We started with the SciCap [16] challenge[8] validation set, which contains figures, captions, and paragraphs for a large number of figures. Initially, we observed imbalance in figure types and research fields. We quantified this using pretrained classifiers for figure type (DocFigure [19]) and field of study from the mention paragraphs (S2-FOS[9]), finding highly skewed distributions. Figures from Physics were highly overrepresented, as well as line plots. We resampled with replacement to the third most populous categories for type and field, then dropped duplicates to obtain a broadly representative set without overly distorting the original distribution. To create a modest-sized development set, we embedded and vectorized the figures using CLIP [37], caption and mentions using SPECTER [7], and figure type as one-hot encodings. We used a facility location submodular optimization algorithm from the apricot [39] package to efficiently select a diverse subset of 30 figures. We confirmed through manual review that the figures had low content overlap, visual distinctiveness, and representation across scientific fields. We used this set to iterate on prompts by generating descriptions for these figures and spot-checking the results.

Guidelines for Suggestion Generation. In feedback from the formative study, authors noted that there were errors in some of the AI suggestions. In response, we updated the OCR model used for

[8] http://scicap.ai/
[9] https://github.com/allenai/s2_fos

figure text extraction from Tesseract[10] to EasyOCR,[11] which produced more accurate textual figure representations. We also greatly expanded the set of guidelines used; initially we only implemented two sets of guidelines for plots and non-plot figures respectively. In the re-designed system, we collected an extensive set of guidelines from sources including the DIAGRAM Center[12] and SIGACCESS[13] guidelines. We adapted these guidelines to (1) remove references to specific example figures, (2) remove presentational guidelines, such as conciseness, and focus on those relating to content, and (3) organize them as a nested list indexed by figure type. From the previous version, we maintained general guidelines applicable to all figures, and additional guidelines for all plots including the first three levels of Lundgard and Satyanarayan's four-level model [29]. We included the fourth in the formative system, but removed it based on the observation that the first three levels are more often found useful by end-users, and that the fourth typically requires significantly more exogenous context to integrate, which may not be available from the extracted metadata.

Base Model Selection. We compared baseline generations from GPT-4 [36] and LLaVA [27] for 5 figures sampled randomly from our development set of 30. Among the outputs, GPT-4 tended to produce more descriptive alt text with fewer hallucinations. This, coupled with the higher likelihood of LLaVA failing to generate any alt text at all, led us to choose GPT-4 as our base model. This choice can be reconsidered in the future with the emergence of more powerful language and vision-language models. Note that GPT-4 with Vision was not available for comparison at the time that this work was conducted.

Prompt Engineering. We identified unhelpful patterns in model-generated suggestions through testing and observations during the formative study. To address these, we made several prompt adjustments:

(1) Added instructions like "Respond with only x" to avoid chat-like responses and keep suggestions focused on the requested task (e.g. text continuation).
(2) Added an instruction and logit biases to avoid explicit references to metadata. Metadata should inform responses without being directly referenced (e.g. "the OCR-extracted text contains..."). We experimented with reiterating the instruction at the end of the instruction set, finding this further reduced such occurrences.
(3) Motivated by prior work [55], we added an "uncertainty prompt" to mitigate sensitivity to metadata extraction errors. In our version, we acknowledge they may exist and encourage inferring details despite this to provide helpful suggestions.
(4) Added instruction to focus on the figure visual metadata and key information, reducing suggestions derived from the text that do not describe visual aspects of the figure.

Although it is difficult to systematically evaluate the effect of such changes or their reproducibility, we include them here to describe our design process for improving AI assistance and mitigating observed issues.

We also created prompt variants for different contexts like generating initial summaries versus later continuations, adding placeholder text and instructions to improve infilling around user text. The appropriate context is inferred from the system's state:

(1) Initial High-Level Summary Prompt: Generates a high-level summary when no description exists yet.
(2) Continuation + Infilling Prompt: Extends existing text by referencing the description context. For infilling, we tested multiple strategies:
    - Naive infilling (without providing post-cursor context): often resulted in duplicate content
    - In-prompt context (adding post-cursor text as a prompt metadata element): still resulted in duplicate generations
    - Placeholder text at the cursor position: reduced duplication, so we selected this approach
(3) Draft Prompt: Variation on the initial high-level summary prompt to generate a full description, used for pre-generating drafts.
(4) A separate prompt for generating *Potential User Questions* and corresponding suggested answers.

4 SYSTEM DESIGN

FigurA11y consists of a backend architecture for processing and extracting figures from scientific PDFs (Section 4.1), figure metadata extraction (Section 4.2), and figure description prompting (Section 4.3), as well as a user interface for AI-supported figure alt text writing (Section 4.4).

4.1 Overall Pipeline Architecture

The overall system architecture consists of several steps, as depicted in Fig. 1. The first stage involves uploading an academic paper in PDF format. The system then extracts figures, captions, and paragraphs mentioning each figure. The figures and captions are extracted using PDFFigures2 [6], and the mentioning paragraphs are extracted from the paper using GROBID [28] to extract the text and a regular expression to match mentions with the figure number in each caption. These methods are similar to those used in recent work on knowledge-augmented figure captioning [50], but we incorporate the caption as well since our goal is to generate alt text, in addition to information extracted from figures and hierarchical guidelines (reviewed next).

4.2 Metadata Extraction

Metadata is then extracted from each figure, including classifying the figure type using the pre-trained DocFigure [19] classifier (e.g. bar plot, tree diagram, etc.). We focus our methods on plots and diagrams, and so we construct an "Other" figure category to account for figure types outside of plot and diagram sub-types. For plots, the plot data table is extracted using the pre-trained UniChart [31] model by default, or the DePlot [26] model if desired. The latter is slower, but we find that it sometimes yields better results, depending

[10] https://github.com/tesseract-ocr/tesseract
[11] https://github.com/JaidedAI/EasyOCR
[12] http://diagramcenter.org/table-of-contents-2.html
[13] https://www.sigaccess.org/welcome-to-sigaccess/resources/describing-figures/

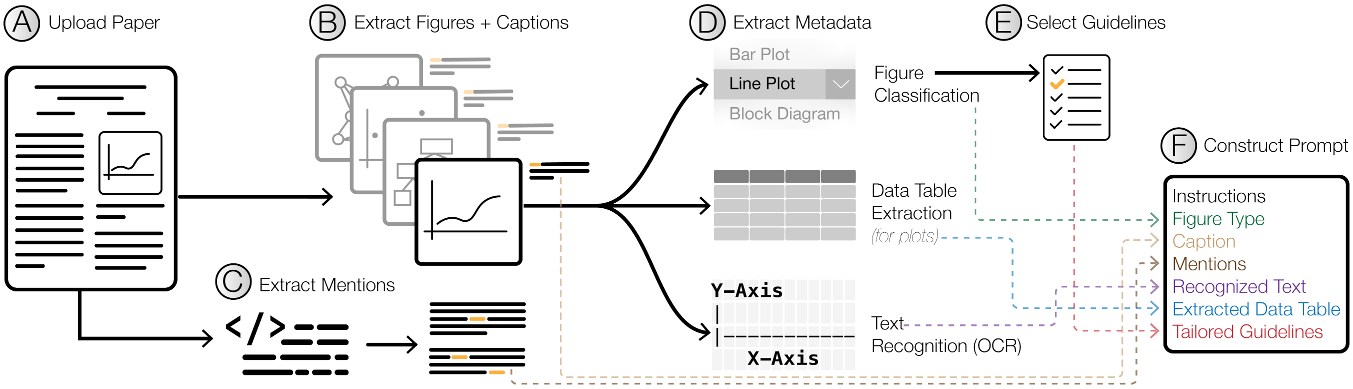


Figure 1: Pipeline for extracting information from figures, and using this information in a prompt to generate draft alt text and suggestions for enhancement. The author first (A) uploads a paper, from which (B) figures and their captions, and (C) mentions of each figure in the paper are extracted. Then, (D) the figure is classified, a data table is extracted if it is a plot, and the figure text is recognized. Finally, (E) based on the figure type, a set of guidelines are selected. (F) all of this information is put together with instructions into a prompt for the LLM to use in generating drafts and suggestions.

on the figure. Text in the figure is extracted using EasyOCR,[14] with the layout preserved; UniChart's results [31] suggest that this can help with LLM reasoning over charts, and we adapt this to our context of open-domain scientific figures.

Finally, this extracted information is assembled into a prompt for the LLM to use in generating text, along with tailored guidelines based on figure type. The full pipeline synthesizes disparate recommendations from prior work, as discussed above, to construct a prompt with tailored instructions and hierarchically-selected guidelines depending on figure type. We next discuss the structure of this prompt.

4.3 Prompt Structure

We use structured prompts to effectively harness large language models (LLMs), specifically GPT-4 in our current system, for generating useful alt text suggestions. Prompts contain two main elements, described in more detail below.

4.3.1 Instructions. We designed several prompts with detailed instructions supporting the core interactions: content generation and potential user questions, as briefly described in the previous section.

- **Initial Summary**: instruct the LLM to introduce the figure in 1-2 sentences focusing only on the most important elements and relationships shown, without additional commentary.
- **Continuation**: prompt the LLM to expand on the existing description by adding 1-4 sentences conveying missing details and relationships relevant to understanding the figure, avoiding repeating content already provided.
- **Infilling**: use the *Continuation* prompt but with added context. Includes placeholder text at the cursor location and instructs the LLM to provide distinct suggestions that fill in gaps within the existing description section.
- **Drafts**: adapt the *Initial Summary* instructions to generate a full figure description.
- **Potential User Questions**: instruct the LLM to generate pointed questions querying unclear visual elements that need explanation in the description. These come with 1-4 suggested answers each, also generated by the LLM based on figure metadata. To maintain this structured format in the generations, we use OpenAI's function calling API[15]. We define a function which accepts a question, along with 1 required answer argument and 3 optional answer arguments, to construct the question-answer sets.
- **Summarization**: provide a brief summary (~1 paragraph) of longer alt texts, to align with guidelines around conciseness and short/long alt text versions[16,17].

The complete prompts are provided in Appendix B.

4.3.2 Metadata. As noted in Section 4.2, prompts incorporate metadata extracted from the figure and paper to ground the LLM's generations. These act as visual information for the LLM to reason over, allowing us to leverage its advanced information processing capabilities without relying on newer, less robust multimodal approaches that result in less descriptive and sometimes empty outputs as we found in our prototyping (Section 3.4). Additionally, VisText [44] found that metadata-driven description outperformed visually-improved description in their case, for plots. Though our circumstance differs (our two model options are not directly comparable), we also find that the combination of metadata we use can produce detailed and grounded descriptions.

- **Figure type** provides high-level context.
- The **caption** often summarizes main ideas depicted and can contain useful details about visual elements.
- **Mentioning paragraphs** give further context from the paper, e.g., describing key concepts or results shown.

[14] https://github.com/JaidedAI/EasyOCR

[15] https://openai.com/blog/function-calling-and-other-api-updates

[16] https://www.w3.org/WAI/tutorials/images/complex/,

[17] https://authors.acm.org/proceedings/production-information/describing-figures

- **Extracted text** conveys lower-level visual details like axis labels and diagram text.
- For plots, the **extracted data table** approximates the values depicted.

4.4 Interface Design and Implementation

The FigurA11y interface was designed to provide authors with AI-assisted support throughout the alt text drafting process, while scaffolding the review and revision process by concisely presenting figure metadata. The left side of the interface displays the figure along with extracted metadata like the figure type, caption, paragraphs which mention the figure, and extracted data values for plots (see Fig. 2 for the **Interactive Assistance** version, for instance). These metadata components serve as prompts to inform the initial AI-generated draft and subsequent suggestions.

The right side contains the main alt text authoring field where authors can write and iteratively refine descriptions. The **Interactive Assistance**'s two augmentative features are engaged by clicking buttons in the authoring field's toolbar, or using the corresponding key commands: `TAB` for *Generate at Cursor* and `(CMD|CTRL)+/` for *Potential User Questions*. The results of the former are shown in the description field; the generated text is highlighted in red and with a differently tagged underlying HTML element. Then, the user can click on a generated snippet, and decide whether to accept or reject it. If accepted, it becomes part of the description and the special formatting is removed. If rejected, it is discarded.

In **Draft+Revise** (shown in full in Fig. 3), the interactive features are replaced with a simple text box with which to prompt GPT-4 as the user desires, to simulate access to an LLM as authors may have in their normal writing workflows. After drafting in either of the system versions, authors can run the summarization workflow. Additional interface features are described in Appendix D.

The full system was implemented in around 6100 lines of TypeScript and 2000 lines of Python using ReactJS, Next.js, Zustand, and Mantine for the frontend interactions, Tiptap and Prosemirror for the interactive editor specifically, Flask for the backend server, and PostgreSQL for the database.

5 STUDY DESIGN

We designed a study to evaluate the usefulness of our system for assisting authors in producing alt text. In particular, we sought to examine (1) whether authors perceive benefit from our pipeline's scaffolding and pre-generated drafts, (2) if the added interactive features in **Interactive Assistance** support authors in further enhancing descriptions beyond editing pre-generated drafts, (3) whether added features incur additional cognitive load, and (4) what strategies participants used when integrating our tool's features into their alt text authoring workflows.

Rather than using a standardized task with predetermined figures, we chose to conduct the study with authors describing figures from their own recent papers. Since our prototype aims to support alt text writing across diverse open-domain figures, it was essential that our lab study be grounded in a realistic context using authors' knowledge of their own content. Our formative results and prior work have also emphasized authors' contextual knowledge as essential for informing alt text drafting.

Beyond assessing overall usefulness, our goal was to understand how different features supported the process of creating complete and accessible descriptions. To compare feature sets, we used a within-subjects design with the two system versions discussed earlier: **Draft+Revise** and **Interactive Assistance**. The **Draft+Revise** condition allowed us to evaluate the draft-generation pipeline and overall revision-support interface. The **Interactive Assistance** condition focused on specific writing assistance interactions. Using both versions allowed us to gather comparative insights. We did not include a baseline without access to any generated text because we do not believe it is realistic to restrict author access to LLMs, given their wide use; however, we note that **Draft+Revise** is a strong baseline not previously available to alt text writers, as it uses our refined alt text draft generation pipeline.

5.1 Materials: Figure Selection

We invited authors recruited for the study to share two to three recent papers containing figures for which they had not yet written alt text. We extracted figures from these papers, and then selected two figures per participant (one for each system version condition).

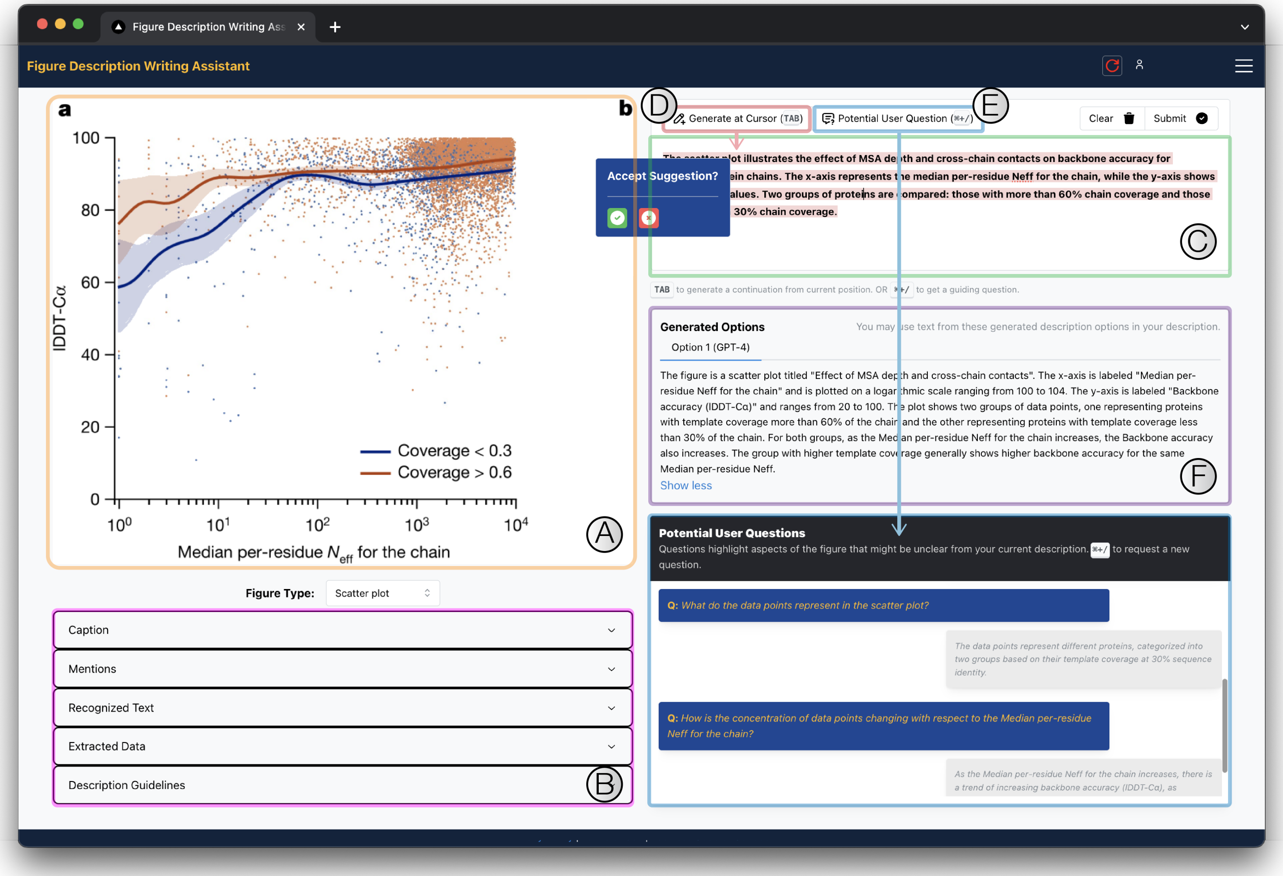
One challenge with this design is that participants could apply the guidelines and suggestions from the first condition to the subsequent condition, if the figures are sufficiently similar. To avoid this, we aimed to select different figure types within participants when possible, typically one chart and one diagram. In cases where participants did not have both types available (e.g., results presented in tables instead of charts, as is common in some domains), we aimed to select substantially different instances (e.g., different plot types, or diagrams that were visually very distinct and did not represent overlapping information).

A second concern was figure complexity. Since figures have a different prior complexity for description tasks (e.g., by being compound, or having many variables or components), varied complexity could produce biased results. Since there is no validated metric for the complexity of scientific figures, we aimed to minimize the impact of this in two ways. First, we randomized the assignment of figures to conditions within participants. This ensured that figure complexity does not systematically factor into the difference between conditions. Second, given our small participant pool, we sought to further reduce this bias. We heuristically selected figures with comparable numbers of visual elements (prior to random assignment) and, if this was difficult to determine, overall subjective complexity. This was to minimize large mismatches in complexity between a participant's two figures, subject to the availability of figures from participants' submitted papers.

We pre-loaded figures into our system to save participants time and effort during the study compared with the full workflow of paper upload and figure selection. We wanted to focus the tasks on writing the alt text itself. Participants were given URLs with figure IDs, which pre-populated the interface with the figure information.

5.2 Study Procedure

We conducted this study remotely via video-conferencing. Participants were assigned to one of two counterbalanced groups determining the order of writing with the two system versions. Group 1 used the **Draft+Revise** version first, followed by **Interactive**



The scatter plot illustrates the effect of MSA depth and cross-chain contacts on backbone accuracy for ein chains. The x-axis represents the median per-residue Neff for the chain, while the y-axis shows alues. Two groups of proteins are compared: those with more than 60% chain coverage and those 30% chain coverage.

TAB to generate a continuation from current position. OR ⌘+/ to get a guiding question.

Generated Options

You may use text from these generated description options in your description.

Option 1 (GPT-4)

The figure is a scatter plot titled "Effect of MSA depth and cross-chain contacts". The x-axis is labeled "Median per-residue Neff for the chain" and is plotted on a logarithmic scale ranging from 100 to 104. The y-axis is labeled "Backbone accuracy (IDDT-Cα)" and ranges from 20 to 100. The plot shows two groups of data points, one representing proteins with template coverage more than 60% of the chain and the other representing proteins with template coverage less than 30% of the chain. For both groups, as the Median per-residue Neff for the chain increases, the Backbone accuracy also increases. The group with higher template coverage generally shows higher backbone accuracy for the same Median per-residue Neff.

Potential User Questions

Questions highlight aspects of the figure that might be unclear from your current description. ⌘+/ to request a new question.

Q: *What do the data points represent in the scatter plot?*

The data points represent different proteins, categorized into two groups based on their template coverage at 30% sequence identity.

Q: *How is the concentration of data points changing with respect to the Median per-residue Neff for the chain?*

As the Median per-residue Neff for the chain increases, there is a trend of increasing backbone accuracy (IDDT-Cα), as

Figure 2: Screenshot of our Interactive Assistance alt text authoring assistant interface. On the left, it shows (A) the figure and (B) extracted metadata. On the right, it shows (C) the description authoring field, (D) the *Generate at Cursor* feature with generated initial text below, (E) the *Potential User Questions* request button and results, and (F) a pre-generated draft description. Example figure is taken from [20].

Assistance. Group 2 used the reverse order. For all tasks, participants were instructed to write descriptions that were as descriptive as possible, rather than aim for conciseness. This allowed for participants to take a more consistent approach towards maximizing information content to make accessible, rather than employing intuitive strategies for conciseness, and also to avoid challenges authors face deciding whether to include a piece of information [46]; we believe that given diverse alt text reader preferences [29] that reader customization should happen at a later stage. At the end of the workflow we provided a semi-automated step to allow authors to create a more concise version.

The study procedure consisted of four main components. First, participants were given a brief 5 minute introduction to alt text and shown examples of effective alt text for a tree diagram and scatter plot from the DIAGRAM Center guidelines. They also received an overview of the study tasks and timeline. Second, there were two 10 minute alt text writing sessions, one for each system version. We determined this time through piloting and observation during our formative study. Participants were allowed to conclude each writing session early, if desired (e.g., if they felt their description had saturated available information to describe). Prior to each one, the experimenter provided a brief, structured walk-through of the features available in the interface. Each session was followed by a 5 minute survey gathering feedback. After the second survey, participants completed an additional 5-10 minute comparison survey. For the first few sessions and those ending with sufficient time remaining, we also conducted a semi-structured follow up interview probing participants' overall impressions, the usefulness of different features, and the strategies they employed beyond what we observed. In these interviews, we asked participants to walk us through their process writing alt text with each system, to offer feedback, and additional questions based on their interactions and comments. This multi-stage procedure allowed us to observe system use, gather both immediate and retrospective feedback, and have an open-ended discussion.

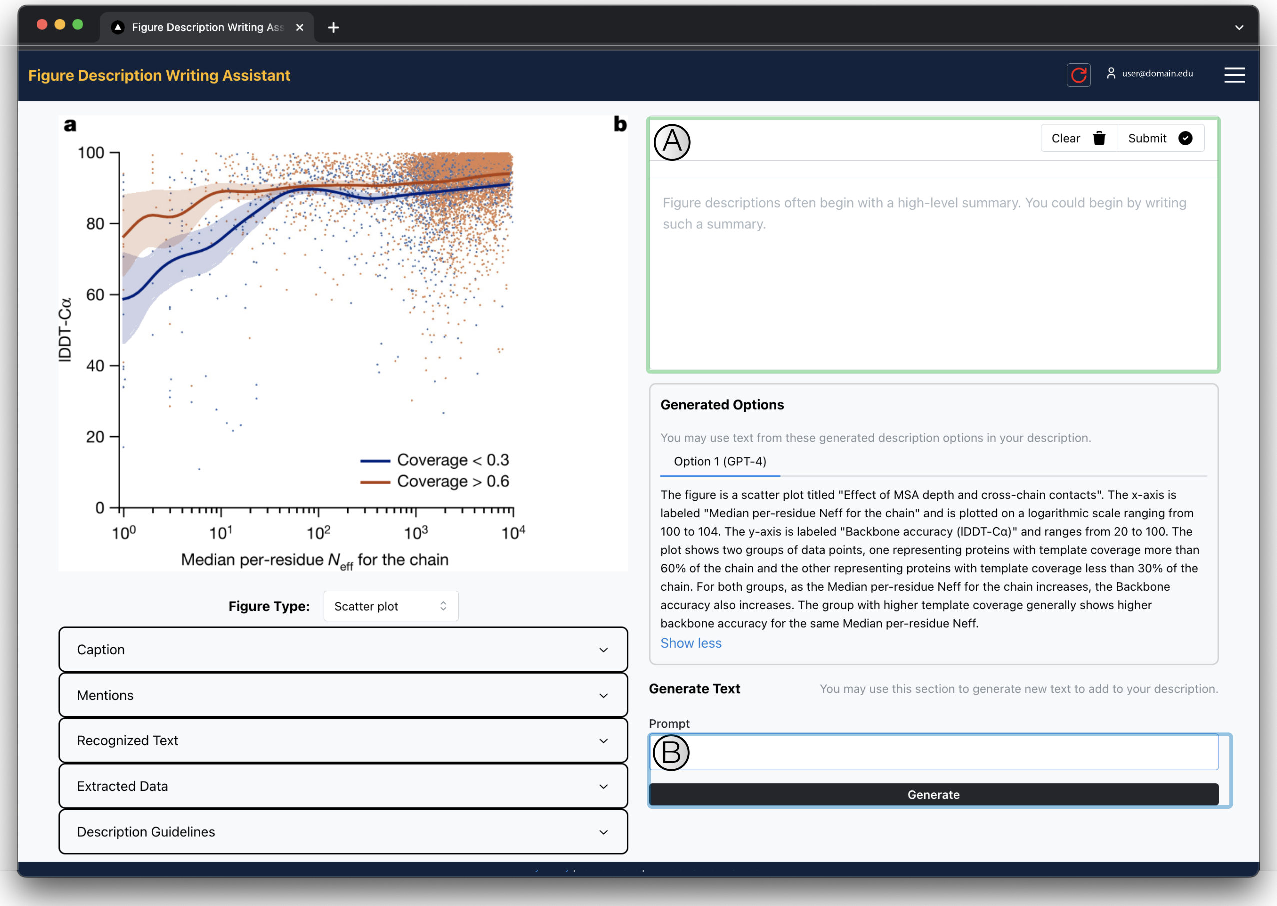


Figure descriptions often begin with a high-level summary. You could begin by writing such a summary.

You may use text from these generated description options in your description.

The figure is a scatter plot titled "Effect of MSA depth and cross-chain contacts". The x-axis is labeled "Median per-residue Neff for the chain" and is plotted on a logarithmic scale ranging from 100 to 104. The y-axis is labeled "Backbone accuracy (lDDT-Cα)" and ranges from 20 to 100. The plot shows two groups of data points, one representing proteins with template coverage more than 60% of the chain and the other representing proteins with template coverage less than 30% of the chain. For both groups, as the Median per-residue Neff for the chain increases, the Backbone accuracy also increases. The group with higher template coverage generally shows higher backbone accuracy for the same Median per-residue Neff.

You may use this section to generate new text to add to your description.

Figure 3: Screenshot of our Draft+Revise alt text authoring assistant interface, showing some of the same features as the Interactive Assistance version: figure and metadata on the left side; and the description authoring field and a pre-generated draft description on the right side. However, there are two differences: (A) the description authoring field does not contain the *Generate at Cursor* and *Potential User Questions* features, and (B) we provide a box to freely prompt the LLM to generate text that the author can integrate into their description. Example figure is taken from [20].

5.3 Recruitment and Participants

We recruited participants using the authors' academic social networks, snowball sampling, and institutional mailing lists. Our study included a total of 14 participants: 9 women, 4 men, and 1 non-binary participant. Their ages ranged from 18 to 44 years old, with most (10 participants) aged 25–34. In terms of roles, there were 7 graduate students/research assistants, 3 postdoctoral researchers, 2 assistant professors, 1 lawyer and researcher, and 1 scientific assistant. The participants' fields of study were diverse, including 5 in formal sciences like computer science and math, 3 in applied sciences like engineering and medicine, 3 in human-computer interaction or design, 2 in social sciences, and 1 in information sciences. The participants also varied in their amount of prior research experience, with 4 having published 1-5 works, 4 having published 5-10 works, 2 having published 10-20 works, and 4 having published over 20 works. Most participants indicated that the majority or all of their prior published works contained figures. However, many had limited experience writing alt text for these figures, with 5 having written no alt text previously and 7 having written alt text for 50% or less of figures. In terms of familiarity with alt text guidelines, 6 were somewhat familiar and 2 were very familiar, while 6 were not very or not at all familiar. When asked about AI writing assistants, 8 had tried them before, 4 used them regularly, and 2 were aware of them but had not used them.

5.4 Data Collection, Evaluation Methodology, and Measures

5.4.1 Questionnaires. Participants completed the following:

(1) Cognitive Load and Usability (completed after each system variant):

- NASA TLX dimensions: mental demand, temporal demand, effort, frustration. We also included own performance, but factor it out in our analysis to differentiate self-assessed performance from experienced cognitive load.
- A usability or system acceptance scale based on recent work on AI assistance [21].

(2) Comparative Preference: A single preference rating on a divergent scale ranging from 1 (**Draft+Revise**, referred to as *Without Suggestions*) to 7 (**Interactive Assistance**, referred to as *With Suggestions*).

(3) Open-Ended Questions: A set of questions covering topics such as in which situations the system variants were helpful or unhelpful, and suggestions for improvement.

5.4.2 Description Measures. We computed metrics to compare the final descriptions against the generated draft. In particular, we sought to capture the degree to which participants' descriptions diverged from these drafts. We assessed this using a range of metrics like the Levenshtein edit distance [24] and Zlib-based normalized compression distance (NCD) [5], using implementations from the `textdistance` package[18]. We also used cosine similarity of embeddings produced by the `all-distilroberta-v1` [38] pretrained language model from the `sentence-transformers` package[19] for a less length-sensitive and more semantic view.

5.4.3 Logs. In addition to logging participants' descriptions, we logged key presses (split into "Input" (additions) and "Backspace or Delete" (deletions), as well as whenever text was pasted from the clipboard (e.g. copied from the draft or suggested answer for a Potential User Question). Examining keylogs allows us to assess task effort and compare against reported cognitive load, to assess whether the added features in the full **Interactive Assistance** system induced or saved additional effort. We also examine *traces* of the interaction through these logs over time, to illustrate different strategies used by participants to produce alt text descriptions with the features available in both systems.

5.4.4 Screen Recordings and Transcripts. The study sessions were screen-recorded to capture participants' on-screen interactions. We also recorded audio and transcripts of the participants during interviews. These were later examined to compare against usage logs, and to keep track of observations made during the sessions.

5.4.5 Challenges for Evaluating Quality. We considered using a descriptiveness metric from prior work [46] to evaluate the level of detail of alt text descriptions. However, the descriptiveness measure was defined based on the range of human-written figure descriptions, with a substantial part of the scale dedicated to low descriptiveness or not descriptive alt texts. The pre-generated alt text drafted by large language models used to seed our system variants introduced a distributional shift from human-written alt text. These generated descriptions tend to be sufficiently long and detailed, such that the descriptiveness metric is no longer effective for distinguishing between pre-generated and human-edited versions of these alt texts.

During our system redesign, we piloted an annotation task with two base models generating draft descriptions: GPT-4 [36] and LLaVA [27]. We asked three individuals with undergraduate training in physical and life sciences to annotate descriptions generated for 5 figures from our development set: 2 each with GPT-4 and LLaVA, and one for each model with and without description guidelines. We adapted annotation guidelines based on the previously defined levels for descriptiveness [46], while introducing half-step levels (9 total levels) to capture finer-grained differences. We found that there was low correlation between pairs of annotators (Spearman's rho 0.246-0.462), challenging the use of this metric in our high-descriptiveness regime. Instead, we evaluate description detail through metrics like divergence from drafts and length, and leave establishing robust descriptiveness metrics to future work.

6 RESULTS

Overall, the results indicate that participating authors preferred the **Interactive Assistance** system version over the **Draft+Revise** version. The **Interactive Assistance** version helped users produce longer, more detailed alt text that diverged more from the initial AI-generated drafts on average. Participants appreciated the pre-generated drafts in both systems, but found features like *Potential User Questions* and *Generate at Cursor* useful for highlighting additional details and supporting incremental drafting in **Interactive Assistance**.

6.1 User Preferences and Responses

Participants generally preferred the **Interactive Assistance** interface as shown in Fig. 4, with 13 participants indicating preference for the **Interactive Assistance** tool to varying degrees. The one participant who preferred **Draft+Revise** noted that they found the workflow of editing the pre-generated description to be less effortful. We tested that these ratings deviated from the neutral level (4) with a one sample t-test, which showed a statistically significant result with a large effect ($t(13) = 4.16$, $p < 0.01$, Cohen's $d = 1.1$).

Participants who preferred the **Interactive Assistance** offered a number of reasons for this, including:

- *Potential User Questions* highlighting elements that might have been missed.
- *Generate at Cursor* allowing incremental drafting.
- *Generate at Cursor* anticipating user needs or replacing user effort in context.

Several participants found the initial pre-generated draft (available in both conditions) useful, with some even indicating that the usefulness of this option diminished the value of the *Generate at Cursor* feature.

Finally, participants identified some usability issues and potential changes to improve experience when working with their own figures. These ranged from behaviors in edge cases (e.g., rapidly double triggering suggestions produced unexpected behavior) to interface features that would assist in smoother review (e.g., visibility of multiple types of metadata at the same time).

6.2 Workload, Usability, and Utility

Both system versions show comparable cognitive load (Fig. 5), despite the added interactive features in **Interactive Assistance**. We

[18] https://github.com/life4/textdistance
[19] https://www.sbert.net/

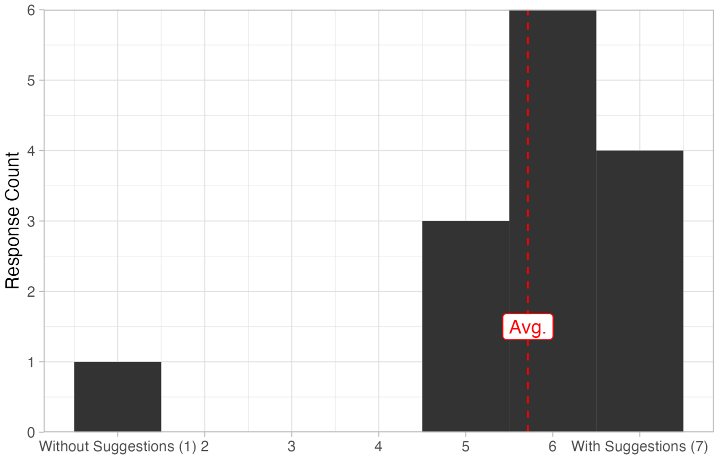


Figure 4: Overall participant preference between the system versions. Results favor the Interactive Assistance version.

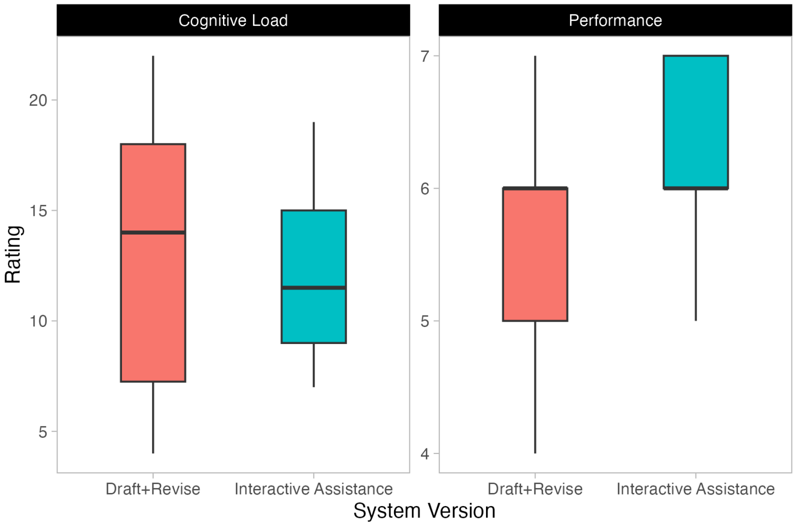


Figure 5: Partial raw NASA TLX results, summing the demand scores (left; with the *Physical Demand* item removed), and factoring out the *Performance* item (right). The score distributions are comparable between the two system versions, overall.

report the four-item raw NASA TLX score representing cognitive load without the *own performance* and *physical effort* items, and the factored-out item representing participants' assessment of their own task performance.

Usability and utility questions indicated an overall preference for the **Interactive Assistance** version, but generally favorable results for the base **Draft+Revise** version as well (shown in Fig. 6). The first four items relating to general tool usability and acceptance showed comparable results for both system versions. However, more participants strongly agreed that **Interactive Assistance** improved efficiency compared to **Draft+Revise**. This was also true for quality ("better alt text"); however, one fewer participant agreed overall for **Interactive Assistance** despite 11 strongly agreeing. **Interactive Assistance** was also reported to more effectively prompt participants to describe elements they may have otherwise missed, a core goal of the added *Potential User Questions* feature. The *Potential User Questions* in **Interactive Assistance** received positive feedback, and the *Generate at Cursor* feedback was mixed but biased positive as well.

While the pre-generated draft feature was identical in both systems, it was rated as slightly more helpful in **Interactive Assistance**. This could suggest it was used differently in conjunction with the interactive features. Overall, the user feedback indicates broad acceptance for the core draft generation, with added value from the interactive assistance features in **Interactive Assistance**.

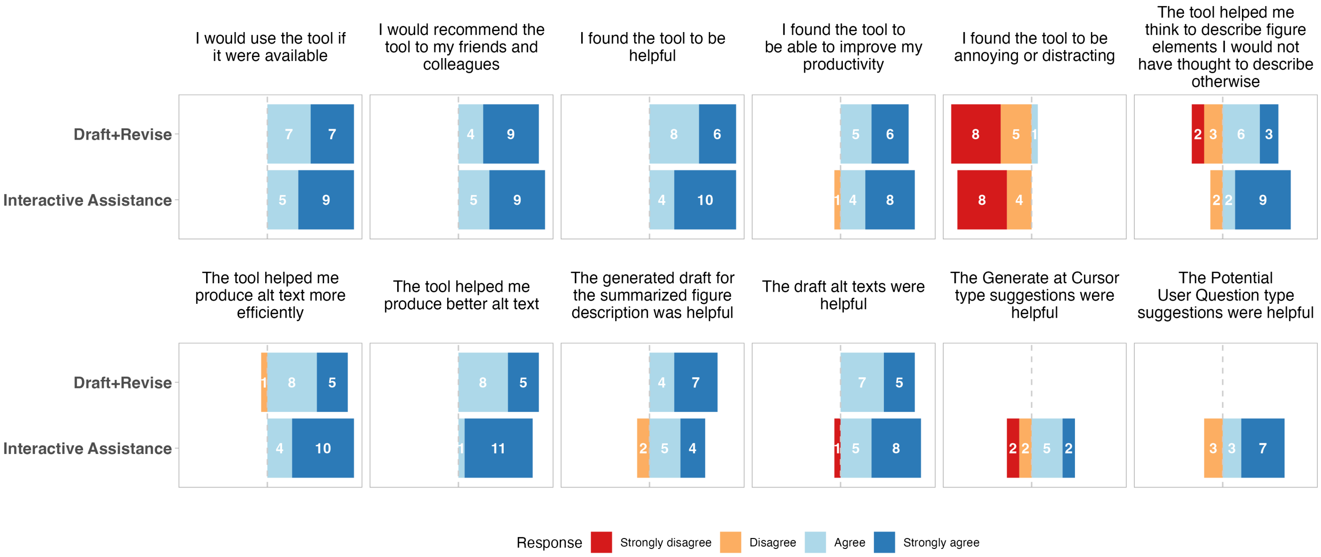


Figure 6: Usability and utility ratings of both versions of the system.

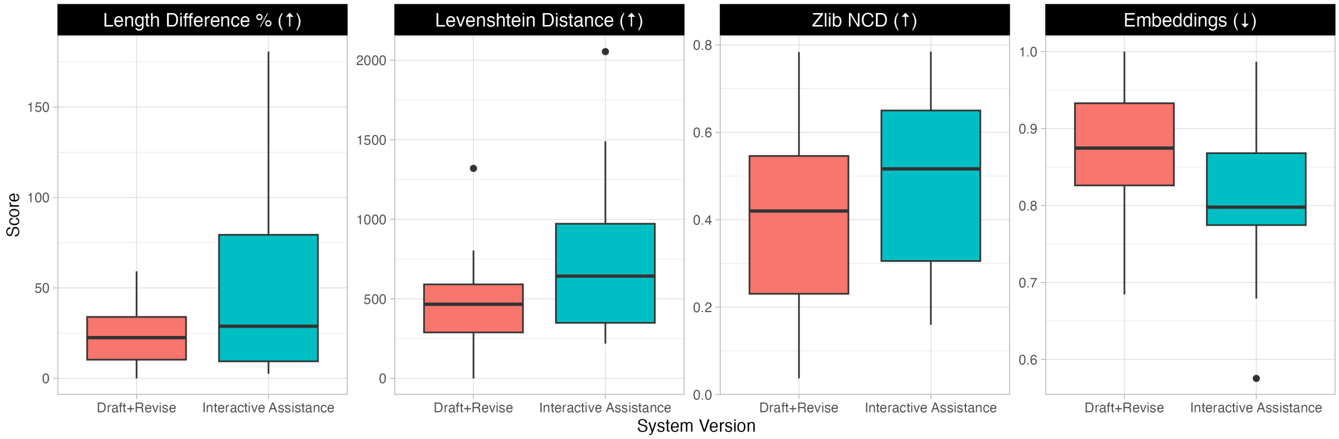


Figure 7: Measures of divergence between the pre-generated alt text drafts and authors' final alt text. Overall, descriptions in the Interactive Assistance condition deviated substantially more from pre-generated drafts across methods (note that the *Embeddings* scores are cosine similarity, and as such are inverted compared with the other metrics; higher similarity indicates *lower* divergence).

6.3 Change in Final Descriptions

One of our core design goals was to encourage greater detail and reflection from authors when writing alt text. We compute several metrics to quantify the textual divergence between the pre-generated draft alt text and authors' final alt text (i.e., how much authors edited the generated alt text) across conditions as a proxy for detail and reflection (Fig. 7).

Between the draft and final alt texts, we computed the absolute percentage length difference, the Levenshtein edit distance [24], the normalized compression distance (NCD) [5], and the cosine similarity of language model embeddings as computed by Sentence-BERT, `all-distilroberta-v1` [38]. Across all metrics on average, authors deviated more from the initial AI-generated draft when provided interactive assistance. This suggests that the added interactive features (*Generate at Cursor* and *Potential User Questions*) provided more opportunity for authors to revise and customize the alt text.

First, we observed descriptions in the **Interactive Assistance** condition to be longer on average than those in **Draft+Revise** (1348 vs. 1075 characters on average). On average, the **Interactive Assistance** condition saw significantly greater changes in length (mean of ∼52% change) compared to **Draft+Revise** (mean of ∼23% change). Individual differences from generated descriptions ranged up to 150% in the **Interactive Assistance** condition. Though informative, length alone does not fully capture textual changes. The Levenshtein edit distance [24] count how many insertions, deletions, and substitutions are needed to transform one string into another. Edit distance also revealed significantly more alterations in the **Interactive Assistance** condition. However, as Levenshtein distance can be influenced by the previously reported length differences, we report two additional metrics. Normalized compression distance (NCD) [5] measures how much compressing two strings together differs from compressing them separately. Unlike Levenshtein distance, NCD is less sensitive to length differences. Additionally, cosine similarity of language model embeddings captures semantic similarity beyond length. With both these metrics, we again see greater divergence from the pre-generated drafts in the **Interactive Assistance** condition (higher on average for the former, and lower on average for cosine similarity in the latter).

6.4 Semi-Structured Interviews

We interviewed 7 participants (half of the total 14), selecting the first 7 whose sessions left sufficient time remaining (typically 5-10 minutes) after the interactions and surveys within the total 1-hour time slots. The variation in time available mainly had to do with time spent on surveys' free-response items, if participants arrived late to the session start, and any connectivity issues, rather than time spent drafting (though some participants did finish particular descriptions before 10 minutes). Specifically, this included P1, P2, P3, P4, P5, P10, and P12. One researcher reviewed these transcripts using a hybrid approach: deductively, to add nuance to the primary findings in the surveys and observed behaviors, and also inductively, to discover factors not covered by other feedback.

The interviews highlighted the value of different system features for prompting potentially missed details and incremental drafting. Participants appreciated the pre-generated drafts, and most also found the *Potential User Questions* and *Generate at Cursor* features useful.

6.4.1 Potential User Questions helped authors reflect on missing elements. Authors reported ways in which the *Potential User Questions* highlighted elements they might have missed otherwise. P1 noted *"they were asking questions of images I wouldn't necessarily think of because I've seen most of those images hundreds of times,"* and similarly P5 commented: *"The potential user question function was super helpful because it was asking some questions that I never thought of and pointing out certain points that I might miss."* P12 more broadly noted that *"it made the captions so much richer and so much fuller and better."*

6.4.2 Generate at Cursor encouraged thinking and supported iteration. P4, who made use of the *Generate at Cursor* suggestions to produce a very long and detailed description of a complex figure, highlighted how they found these suggestions useful: *"it could be a bit overwhelming when you're looking at the full generated text and then figure out how you're gonna tweak this... You could generate more chunks based on what you're writing as well. So it's very collaborative,"* and specifically pointed out *"the chunks are nice in the sense that they encourage you to use your own brain as you're writing, and then use that as an aid."* P2 expressed a similar idea, that *"the cursor feature... makes you think more,"* however disliked this aspect and preferred the **Draft+Revise** workflow. P3 similarly commented that *"it takes a little more time but it gives you much deeper breadth to the text."* P5 highlighted the value for rewriting: *"it's helpful for especially just editing a certain portion of the text rather than rewriting the whole entire text."*

6.4.3 Pre-generated drafts were a useful starting point. P1 noted how the generated drafts brought their attention to the difference between captions and alt text: *"I think comparing [the generated description] to [those] that were already written in my papers is adjusting to what alt text would look like versus just a regular image caption."* P10 remarked that the generated draft was sufficiently helpful that the *Generate at Cursor* didn't add much beyond it, *"I tried a little bit the add text at cursor... I just felt like for mine, at least, I either liked parts of the generated text better or it just wasn't really adding anything more to what I already had there."* They also commented more specifically that *"it pulled in pieces of data that I just wouldn't have."* P12 noted *"I like being able to copy the whole paragraph and edit as I needed,"* and that *"it was pretty great and then it didn't really need that much editing."* P2 reported, when asked to compare this process to their prior experience writing alt text: *"To be honest, the alt text descriptions generated... are much more exhaustive and it's covering all the major parts."* P2 also noted having to correct an error in the generated description, but explained that this was easy to do: *"it generated text for bar chart... it somehow detected the other category which was not there, but it was very easy to do it."*

6.4.4 Interface features helped authors review descriptions. P10 commented that they *"hadn't looked before [for] guidelines for alt text, but it was nice to have that as a reference."* P2 noted that though they did not directly use the guidelines in their interaction with the specific figures, they *"usually have a lot of different types of graphs and I used to struggle with finding the guidelines and then again I had to open the tab and Google and search about it,"* and so could see the automatic guideline selection and presentation being useful. P2 also noted that *"the mentions were good"*, but that they did not find value from reviewing the OCR-extracted figure text and extracted data table. We also observed from participants' screens that several participants referred to the caption, mentions, and guidelines to check against their alt text draft.

6.4.5 Authors perceived value for authoring tasks beyond alt text. Some participants identified how the tool might be useful for broader contexts in their academic writing. P12 remarked that *"[potential user questions] made me think more about the paper, and things that I might want to include in the discussion section or limitations."* P1 commented on the *Generate at Cursor* feature's initial high-level summary that *"I think those short summaries could be really helpful in writing my presentation script, to have something to describe the images on the screen especially [during an] oral presentation. I don't want to go into too much detail or depth."* P10 noted *"I also may use it [for] generating my captions as well, because I noticed my captions are really lackluster."*

6.5 Participant-identified limitations

Participants flagged instances of incorrect generations, including figure classification errors, mis-recognized characters in the OCR and these leaking into the description (e.g. "1" for "I"), or mistaking values or value ranges. Though some errors were identified as *"nothing I couldn't easily correct," "very minor," "not a big issue,"* or needing *"very little effort,"* P1 noted that *"It might be unhelpful if the tool generates false captions for something that isn't in the image and the author doesn't read it over,"* emphasizing the importance of author revision. For example, P8 noted that *"the tool had missed the most important category [in the figure] (in my opinion, since that category was central to the paper's argument)."* Additionally, participants made several usability recommendations that we plan to adopt, such as supporting parallel review of multiple figure metadata attributes to compare to the description.

6.6 Log Analysis

We also investigated event traces for insights into how participants engaged with our features. We found that authors expended comparable manual effort between conditions, measured in additions and deletions. Examining individual interaction traces showed how participants employed different strategies in using the features; ranging from reviewing and submitting a pre-generated draft to incrementally building a draft using snippets interleaved with manual writing. This demonstrates that participants who found the tool's features useful may have used them in diverse ways, adapted to their needs and figure context.

6.6.1 Logs Sample Analyzed. Due to database sync issues, event logs from the first five participants were incomplete (specifically, from the base **Draft+Revise** system). As such, results in this section proceed with the remaining 9 participants. Since participants were scheduled by availability, and the group order was alternated between subsequently scheduled participants, we do not expect this to systematically bias our results in any way. As a robustness check, we alternately dropped the first and last participants in this list to create balanced sets of 8 (4 in each group) and the subsequent results did not substantially change.

6.6.2 Aggregate Counts of Events. As a first analysis of participants' logged interactions, we examine aggregates by type (deletions, additions, and text-pasting). Median counts of deletions ("Backspace or Delete"; 60 vs. 56) and additions ("Input"; 450 vs. 399) are comparable across the two system variants (we observe very slightly higher medians and moderately higher dispersion for **Draft+Revise**). The narrower range of key-presses in the **Interactive Assistance** version could indicate that the added features allow a more efficient writing process in some cases. However, the lowest addition and deletion count for **Draft+Revise** is 0, lower than for **Interactive Assistance**, as P12 did not make any edits in this condition, but was satisfied with the pre-generated draft after reviewing for some time. The median count of paste events is higher in **Interactive Assistance**(5 vs. 2); this might account for the added pasting from suggested answers to *Potential User Questions*, in addition to pasting from the pre-generated draft and from elsewhere within the figure metadata or participants' working descriptions.

6.6.3 Event Traces. To obtain a more fine-grained view into participants' interaction and writing strategies, we examined event logs by participant as histograms over time (starting from the first in-session event). Three examples of this are shown in Fig. 8, to highlight the differences in how the tools' affordances supported the participants' alt text authoring.

P10, shown in the first row, incrementally built up their description by pasting text from the draft at various points, interleaving this with their own writing. They used the *Potential User Questions* to test *"how it was coming across,"* and found this useful despite not directly pasting in suggested answers.

P12 only performed one action directly towards producing the description in the **Draft+Revise** condition; they pasted in the pre-generated description, and then reviewed and accepted it. This is apparent in the second row, where the **Interactive Assistance** condition shows substantial presence of deletions, additions, and paste events over the session, but the **Draft+Revise** condition shows only one paste event at the beginning and no other events logged. The screencast recording of P12 shows that they used the open-prompt box, and even tried to obtain a similar effect to the *Potential User Questions* by asking which aspects were unclear from the description (this participant was in group 2 and had already interacted with **Interactive Assistance**). They ultimately decided that, from the resulting questions, *"none of this is helpful."* This process took almost 5 minutes, including time spent reviewing figure metadata.

P14, on the other hand, pasted parts of the generated draft at the beginning in both conditions, but then in **Interactive Assistance** proceeds to paste additional text. Some of this came from the draft (particularly before 300 seconds), and then subsequently from suggested answers to the *Potential User Questions*. Towards the end of this description, P14 pasted two separate answers to the same question into their description in sequence, as they contained complementary details relating to the function of the same visual cues shown in the figure.

In summary, participants used the available features in individual ways reflecting their needs and preferences to craft detailed figure descriptions. Strategies we observed varied even more widely, including patterns like pasting generated drafts and then extensively editing them. The examples illustrate the diversity of strategies employed to balance writing, integrating suggestions, and revising, with support from the system. We include all participants' individual event traces in Appendix C.

7 DISCUSSION

The present work demonstrates how a human-AI collaborative workflow can support authors in making their figures accessible through producing descriptive alt text. Our results show that automatically generated drafts and an interface supporting revision accelerated the authoring process. Additional interactive writing support features, including on-demand text generation (*Generate at Cursor*) and information-seeking queries (*Potential User Questions*), further helped most authors by progressively building comprehensive descriptions and highlighting points they may have otherwise missed. An analysis of system usage shows authors leveraged these features extensively and in diverse ways depending on their figures and preferences. In the interactive condition, authors produced longer alt text diverging more from the initial drafts, despite similar cognitive load and key-press counts on average. Overall, the system mitigated key challenges authors face in crafting complete figure descriptions. This human-AI collaborative approach highlights the

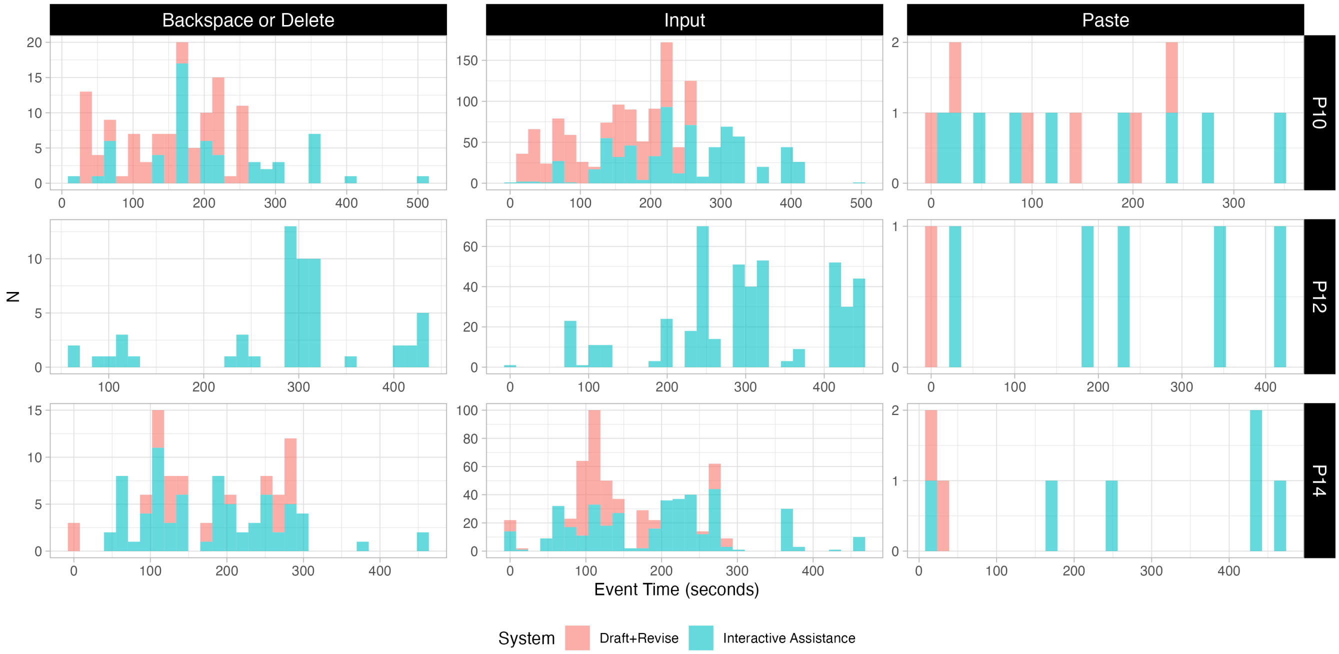


Figure 8: Traces of three participants' interaction by event type, highlighting how participants used different strategies to produce final descriptions.

opportunities for combining human contextual knowledge and AI capabilities in making scientific communication more inclusive. However, realizing the full potential of such collaborative authoring systems for accessibility requires addressing issues like generalization and robustness across real-world figures and alt text authoring contexts, integration into diverse author workflows, and responsible deployment, which we explore in this section.

7.1 Rise of Multimodal Models

Advances in multimodal language models, which incorporate vision and language, point toward expansive future capabilities for automated alt text generation. Our approach relied on metadata extracted from figures and papers to provide contextual grounding for language models, since today's large language models substantially outpace widely available multimodal models in terms of their generation capabilities and can better incorporate large amounts of metadata representing knowledge about figures. However, the ability to process complex figures directly could reduce dependence on potentially error-prone metadata extraction pipelines while incorporating the right kinds of contextual knowledge for support.

While this could enable purely automated description systems, risks accompany such approaches. Recent work has shown how current state-of-the-art multimodal models can make errors when processing complex figures such as scientific figures [17, 18]. Without human validation, model errors or biases could more easily propagate. Maintaining author discretion may prove wise, even as automated methods become more capable. Furthermore, descriptive tasks require not just visual recognition, but reasoning, inference, and judgment. The wisdom accumulated in authors and fields, who can respond to changing contexts, might allow tailoring descriptions for clarity and relevance. Thus, while future multimodal models may better parse figures, the role of human guidance and customization is unlikely to dissolve. Specialized metadata extraction models could also enhance such models' zero-shot capabilities. Visual control could, however, be useful; automatically decomposing complex compound figures into components for iterative description is a promising approach we did not explore.

7.2 Realizing Gains for Alt Text Consumers

A critical question for future work is whether the increased alt text length and apparent descriptiveness from the system translates to improved comprehension for blind and low vision readers. Evaluating alt text quality remains an open challenge, as illustrated by prior work showing divergent reader preferences. Even assessing descriptiveness in the presence of generated drafts may prove difficult, evidenced by low agreement in our annotation pilot. While we aimed to make it easier and faster for authors to produce detailed alt text, realizing accessibility gains requires considering the perspectives of and impact on readers. Follow-up work on evaluation methodology and studies which evaluate the impact of human-AI collaboratively written descriptions on figure comprehension could help to quantify this impact.

Future work should investigate how to incentivize adoption. Though our study aimed to mimic natural workflows, factors like time constraints, competing demands, and incentive structures also inevitably shape real-world use. Even if the system can help improve alt text completeness, lagging integration risks limiting its impact. Overall, while initial evidence is promising, confirming and extending the benefits requires both rigorous accessibility-focused evaluation and understanding practical barriers to mainstream integration.

7.3 Transforming Descriptions to Match Individual Needs

While comprehensive alt text can benefit accessibility, readers have diverse preferences [29] and may desire descriptions of varied lengths tailored to individual needs. Our approach focused on highly descriptive alt texts by design, so that this text can serve as a base to produce personalized derivative texts. As abstractive summarization techniques continue advancing [13, 49, 54], in addition to dialog and other interactive language processing approaches, future systems could apply these methods to accommodate diverse preferences and needs. For example, a concise 1–2 sentence overview could assist quickly grasping key ideas, while retaining the option to query for more information, or expand to more detailed versions for nuanced understanding. Appropriately customizing alt text poses challenges beyond generic document summarization, requiring preservation of visually salient information such as trends in depicted data. However, customization also holds promise to reconcile the objectives of maximizing completeness for authors while matching diversity in user preferences.

7.4 Ethical Considerations

A key ethical consideration is the risk of imposing additional burdens on marginalized communities. Blind and low vision readers already often face exclusion from scientific communication due to the low prevalence of alt text, in addition to other challenges. Providing them erroneous and verbose descriptions without thoughtful human involvement could create further challenges. Though relying on language models is core to the approach in this work, it also risks introducing hallucinations, errors, and biases. Our approach emphasizes author involvement to mitigate these issues, but incentives and workflows must ensure careful review if deployed at scale. The goal should be lightening authors' workload without absolving responsibility. Overall, we must weigh accessibility gains against potential harm, and ensure technical progress on aiding authors in describing figures aligns with the goals of assistive technology.

8 LIMITATIONS

While we evaluated our system on a diverse and realistic set of figures, the study still involved a limited number of author participants (N=14) describing a small set of their own figures (2 per participant). Evaluating the approach on a larger scale with more figures per author would provide stronger evidence. Relatedly, our participants covered a range of fields, but some areas like life sciences were still underrepresented despite our best recruitment efforts. Testing robustness across even more diverse figures and author backgrounds is an important next step towards deployment.

Additionally, our study instructions asked authors to maximize descriptiveness. A different motivation such as information density (maximizing amount of information conveyed in the shortest amount of text) could change how the system is used and the resulting alt texts. The interface features we designed for the initial goal may not generalize to other aspects of alt text that authors or readers may prefer to optimize for in certain settings.

The automated pipeline also occasionally produced errors (like incorrect figure classification or OCR errors) which propagated to the alt text drafts. Though authors could correct these errors (and pointed out such instances), robustness is critical for real-world utility. We did not systematically characterize authors' ability to resolve errors in final versions of their descriptions, but such an evaluation could also help gauge real-world effects of errors in drafts. Enhancing these components, or integrating uncertainty estimates to guide authors, could improve draft quality and adoption.

Finally, though we demonstrate that our system has the potential to improve alt text writing for scientific figures, the disconnect between assistive writing interfaces such as ours and the scientific publication process limits the true utility of our tool. While authors may be able to produce better alt text using FigurA11y, the processes around integrating this alt text into their publications and making the alt text easily accessible to those who need it are still cumbersome. We acknowledge this limitation and push for better and more intuitive processes around scientific paper accessibility that will make it easier and motivate more authors to include alt text in their publications.

9 CONCLUSION AND FUTURE WORK

We present FigurA11y, a human-AI collaborative approach to improve the accessibility of scientific figures through descriptive alt text. By combining a pipeline for automatically generated drafts with an interactive authoring interface that makes contextualized suggestions, our system helped authors efficiently craft detailed descriptions of their own figures. Interactive suggestions further assisted authors by highlighting aspects they may have missed describing, enabling iterative refinement of descriptions, and supporting longer descriptions which diverged more from pre-generated drafts without increasing cognitive load or taking more effort on average. Future work can extend this approach by pursuing strategies like incorporating visual information directly, improving robustness of parts of the pipeline, and integrating with real-world author workflows and incentives, to maximize the positive impact on the accessibility of scholarly communication.

ACKNOWLEDGEMENTS

We thank our user study participants for offering helpful feedback towards the final design and implementation of our FigurA11y system. We thank Bailey Kuehl, Erin Bransom, and Jenna Sparks for conducting a preliminary review of alt texts generated by our pipeline for factuality and information provenance. We thank members of the Semantic Scholar Research team for helpful feedback on project direction and earlier versions of the manuscript.

REFERENCES

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[2] Dana Aubakirova, Kim Gerdes, and Lufei Liu. 2023. PatFig: Generating Short and Long Captions for Patent Figures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2843–2849.

[3] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[4] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. A dataset of alt texts from HCI publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

[5] Rudi Cilibrasi and Paul MB Vitányi. 2005. Clustering by compression. *IEEE Transactions on Information theory* 51, 4 (2005), 1523–1545.

[6] Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. 143–152.

[7] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180* (2020).

[8] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[9] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, 15–29.

[10] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing interactive systems conference*. 1002–1019.

[11] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*. 549–559.

[12] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. 2020. Twitter A11y: A browser extension to make Twitter images accessible. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–12.

[13] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).

[14] Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14953–14962.

[15] 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.

[16] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. SciCap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624* (2021).

[17] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2023. Do LVLMs Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning. *arXiv preprint arXiv:2312.10160* (2023).

[18] Alyssa Hwang, Andrew Head, and Chris Callison-Burch. 2023. Grounded Intuition of GPT-Vision's Abilities with Scientific Images. *arXiv preprint arXiv:2311.02069* (2023).

[19] KV Jobin, Ajoy Mondal, and CV Jawahar. 2019. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 1. IEEE, 74–79.

[20] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

[21] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[22] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence* 35, 12 (2013), 2891–2903.

[23] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[24] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[26] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2022. DePlot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505* (2022).

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[28] Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*. Springer, 473–474.

[29] Alan 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.

[30] Kelly 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.

[31] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. *arXiv preprint arXiv:2305.14761* (2023).

[32] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.

[33] Valerie S Morash, Yue-Ting Siu, Joshua A Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)* 7, 4 (2015), 1–21.

[34] Mickey Nguyen, Matthew Crane, John Romley, and Yannis M Paulus. 2023. Accessibility of Figures in Leading Biomedical and Ophthalmology Journals: Analysis of Alternative Text Use. *Investigative Ophthalmology & Visual Science* 64, 8 (2023), 2806–2806.

[35] Eric Nichols, Leo Gao, and Randy Gomez. 2020. Collaborative storytelling with large-scale neural language models. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 1–10.

[36] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[38] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[39] Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. 2020. apricot: Submodular selection for data summarization in Python. *The Journal of Machine Learning Research* 21, 1 (2020), 6474–6479.

[40] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580* (2023).

[41] Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond Summarization: Designing AI Support for Real-World Expository Writing Tasks. *arXiv preprint arXiv:2304.02623* (2023).

[42] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. 2023. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–57.

[43] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128* (2023).

[44] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. *arXiv preprint arXiv:2307.05356* (2023).

[45] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677* (2023).

[46] Candace Williams, Lilian de Greef, Ed Harris III, Leah Findlater, Amy Pavel, and Cynthia Bennett. 2022. Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference*. 1–12.

[47] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

[48] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1180–1192.

[49] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv*

preprint arXiv:2302.08081 (2023).

[50] Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2023. SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning. *arXiv preprint arXiv:2306.03491* (2023).

[51] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381* (2023).

[52] Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. *arXiv preprint arXiv:2305.14985* (2023).

[53] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[54] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848* (2023).

[55] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594* (2023).

A SURVEYS

A.1 Draft+Revise Survey

Please rate the following factors (Very Low to Very High) based on your experience with the Figure Description Writing Assistant.

- Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
- Own Performance: How successful were you in performing the task? How satisfied were you with your performance?
- Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- Frustration Level: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The tool helped me produce alt text more efficiently.
- The tool helped me think to describe figure elements I would not have thought to describe otherwise.
- The tool helped me produce better alt text.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The draft alt texts were helpful
- The generated draft for the summarized figure description was helpful

Please explain your ratings for each of the above statements.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- I would use the tool if it were available.
- I would recommend the tool to my friends and colleagues.
- I found the tool to be helpful.
- I found the tool to be able to improve my productivity.
- I found the tool to be annoying or distracting.

A.2 Interactive Assistance Survey

Please rate the following factors (Very Low to Very High) based on your experience with the Figure Description Writing Assistant.

- Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
- Own Performance: How successful were you in performing the task? How satisfied were you with your performance?
- Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- Frustration Level: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The tool helped me produce alt text more efficiently.
- The tool helped me think to describe figure elements I would not have thought to describe otherwise.
- The tool helped me produce better alt text.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The draft alt texts were helpful
- The Potential User Question type suggestions were helpful
- The Generate at Cursor type suggestions were helpful
- The generated draft for the summarized figure description was helpful

Please explain your ratings for each of the above statements.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- I would use the tool if it were available.
- I would recommend the tool to my friends and colleagues.
- I found the tool to be helpful.
- I found the tool to be able to improve my productivity.
- I found the tool to be annoying or distracting.

A.3 Comparison Survey

Please reflect back on both interfaces.

What aspects of each interface did you like, and why?

Please explain any situations where the tool was especially helpful: *For example, if suggestions drew your attention to specific visual elements of the figure or ways to describe them, or provided text that did so which you were able to incorporate directly.*

Do you have any other feedback about problems, bugs, or areas for improvement with regard to the interfaces?

Please explain any situations where the tool was unhelpful or detrimental:
Can you provide an example where a suggestion was unhelpful or misleading? If so, why?
Did any suggestions make your alt text worse in a significant way? Please explain.

What changes to each tool would make it more helpful?

Anything else that you would like to share with us?

Which version of the system did you prefer? (Without vs. With Suggestions)

B PROMPT DESIGN

The overall prompt structure is given as follows:

- **Instruction Prompt**
- **Metadata Prompt**
- **Description Content**

B.1 Instruction Prompt

We defined several different versions of the instruction prompt, toward different goals. The first two below form part of the *Generate at Cursor* feature, while the third is used for pre-generating drafts.

B.1.1 Initial High-Level Summary. Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible, based on metadata provided to you.

Some of this data is automatically extracted from the figure, and may contain errors. Infer as much detail as possible from the information given.

Respond with only a brief and high-level overview (1-2 sentences), with no additional content. In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.

B.1.2 Text Continuation and Infilling. Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible, based on metadata provided to you.

Some of this data is automatically extracted from the figure, and may contain errors. Infer as much detail as possible from the information given. Only include clear and helpful statements for understanding the figure. Do not make explicit reference to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.

Respond with only a continuation of the given description itself (1-4 sentences), with no additional content. Add as much detail as possible. You may also be given a DESCRIPTION CONTEXT, which contains text after your response. In this case, provide text that bridges the gap between the description, and additional text the user has already written.

In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.

B.1.3 Full Draft. Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible, based on metadata provided to you.

Some of this data is automatically extracted from the figure, and may contain errors. Infer as much detail as possible from the information given.

Respond with a full description of the figure, with no additional content. In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.

B.1.4 Potential User Questions. Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible. Infer as much detail as possible from the information given.

What visual aspects of the figure are unclear from the given alt text description? Ask a series of questions to elicit all the necessary information about the figure to describe these elements. Based on the type of figure, focus on essential visual aspects that someone who cannot see the figure would need to know. Based on the guidelines and metadata you have access to, suggest an answer for each question. In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only. Do not repeat any existing questions.

B.2 Metadata Prompt

We define the **Metadata Prompt** as:

```
---CAPTION
        <Caption Text>

---FIGURE MENTIONS FROM PAPER
        <Mentioning Paragraphs>

---OCR TEXT RECOGNIZED FROM FIGURE (MAY CONTAIN ERRORS)
        <Layout Preserving OCR Text>

---DATA TABLE EXTRACTED FROM FIGURE (MAY CONTAIN ERRORS)
        <Automatically Extracted Data Table>

---Please refer to the following guidelines
    when writing your description:
        <Selected Guidelines>
---
```

C EVENT TRACES

Fig. 9 shows event traces for all logged participants in our study (i.e. P6 through P14). We provide them here to give a broader sense of the diversity of strategies we observed.

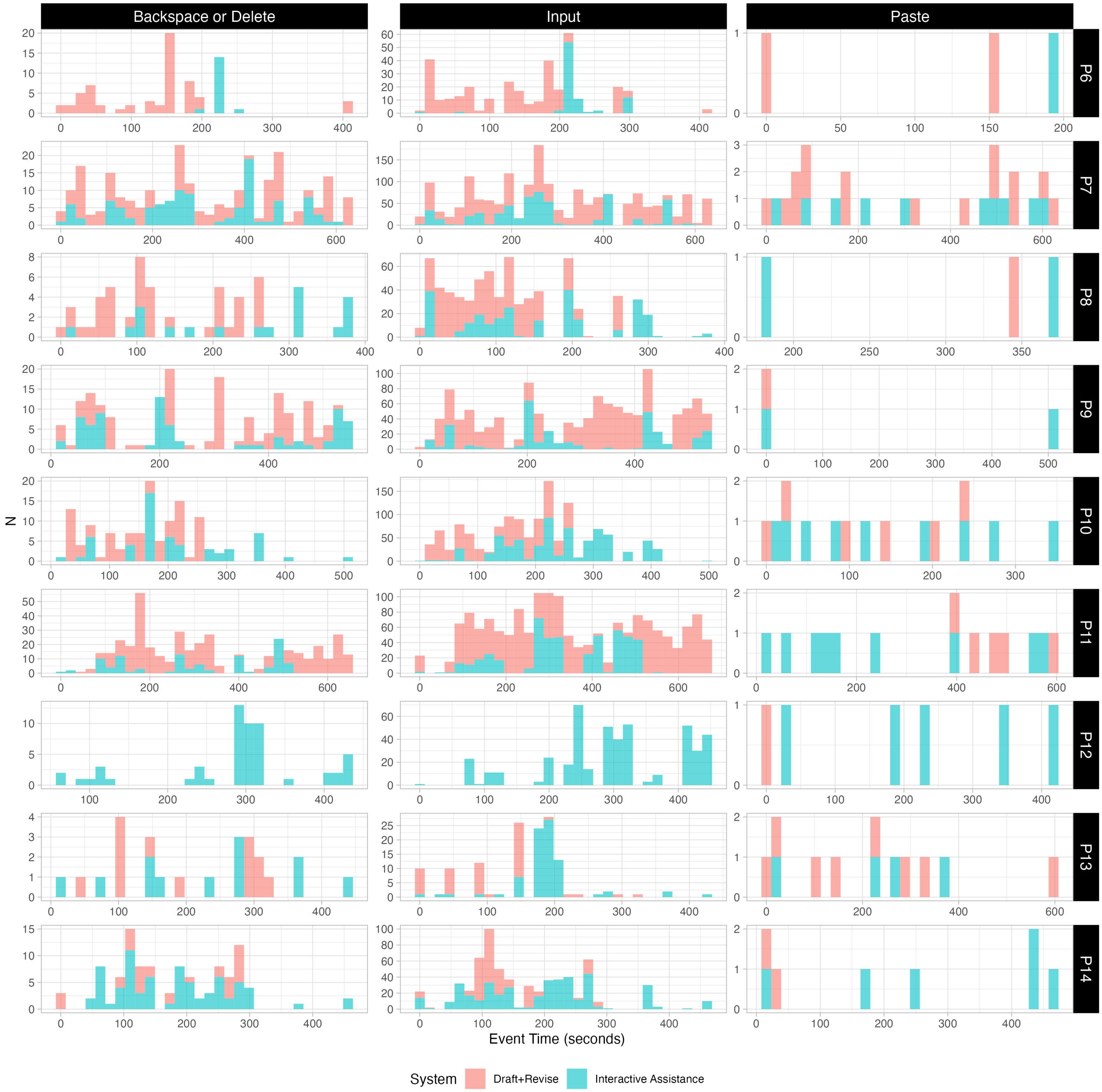
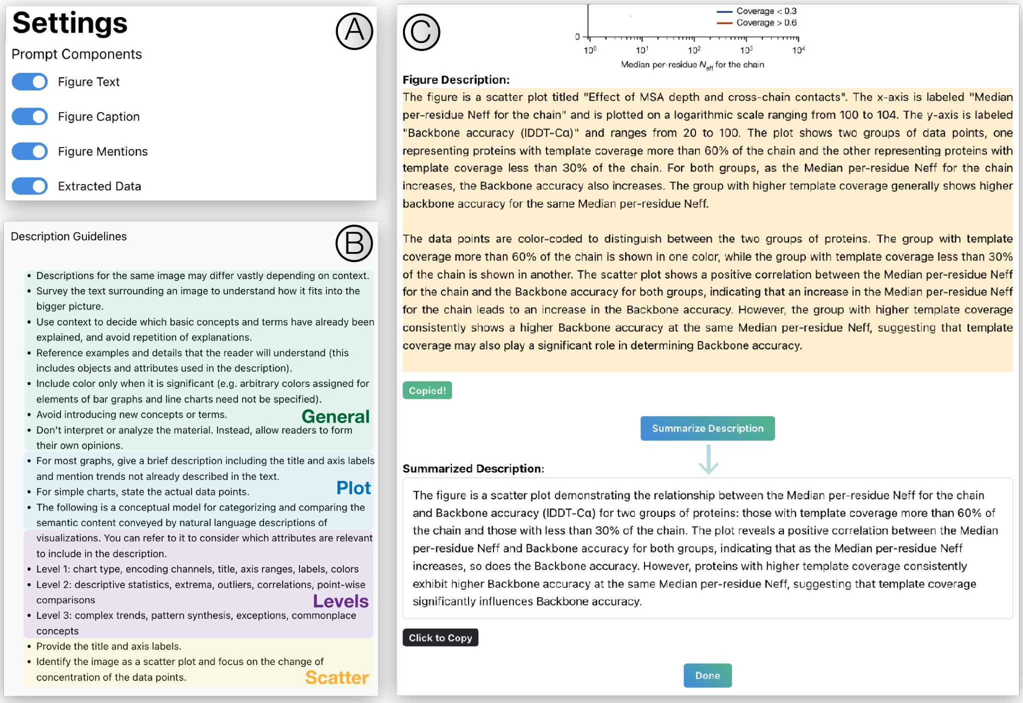


Figure 9: Event traces for all logged participants (N=9) in our study. Different patterns show a wide range of strategies for using our systems' features to produce detailed alt text.

D ADDITIONAL INTERFACE FEATURES

Authors can selectively ablate certain metadata they deem irrelevant or erroneous via interface settings (Fig. 10A). Also present in this menu is a set of guidelines which our pipeline selects based on the figure type (expanded in Fig. 10B), incorporating general figure description guidelines along with domain-specific items (e.g. for general plots), and figure type-specific ones as well (e.g. describing the change of concentration of datapoints for a scatter plot). The summarization workflow is shown in Fig. 10C).



Settings

Prompt Components

- Figure Text
- Figure Caption
- Figure Mentions
- Extracted Data

Description Guidelines

- Descriptions for the same image may differ vastly depending on context.
- Survey the text surrounding an image to understand how it fits into the bigger picture.
- Use context to decide which basic concepts and terms have already been explained, and avoid repetition of explanations.
- Reference examples and details that the reader will understand (this includes objects and attributes used in the description).
- Include color only when it is significant (e.g. arbitrary colors assigned for elements of bar graphs and line charts need not be specified).
- Avoid introducing new concepts or terms.
- Don't interpret or analyze the material. Instead, allow readers to form their own opinions.
- For most graphs, give a brief description including the title and axis labels and mention trends not already described in the text.
- For simple charts, state the actual data points.
- The following is a conceptual model for categorizing and comparing the semantic content conveyed by natural language descriptions of visualizations. You can refer to it to consider which attributes are relevant to include in the description.
- Level 1: chart type, encoding channels, title, axis ranges, labels, colors
- Level 2: descriptive statistics, extrema, outliers, correlations, point-wise comparisons
- Level 3: complex trends, pattern synthesis, exceptions, commonplace concepts
- Provide the title and axis labels.
- Identify the image as a scatter plot and focus on the change of concentration of the data points.

Figure Description:

The figure is a scatter plot titled "Effect of MSA depth and cross-chain contacts". The x-axis is labeled "Median per-residue Neff for the chain" and is plotted on a logarithmic scale ranging from 100 to 104. The y-axis is labeled "Backbone accuracy (lDDT-Cα)" and ranges from 20 to 100. The plot shows two groups of data points, one representing proteins with template coverage more than 60% of the chain and the other representing proteins with template coverage less than 30% of the chain. For both groups, as the Median per-residue Neff for the chain increases, the Backbone accuracy also increases. The group with higher template coverage generally shows higher backbone accuracy for the same Median per-residue Neff.

The data points are color-coded to distinguish between the two groups of proteins. The group with template coverage more than 60% of the chain is shown in one color, while the group with template coverage less than 30% of the chain is shown in another. The scatter plot shows a positive correlation between the Median per-residue Neff for the chain and the Backbone accuracy for both groups, indicating that an increase in the Median per-residue Neff for the chain leads to an increase in the Backbone accuracy. However, the group with higher template coverage consistently shows a higher Backbone accuracy at the same Median per-residue Neff, suggesting that template coverage may also play a significant role in determining Backbone accuracy.

Summarized Description:

The figure is a scatter plot demonstrating the relationship between the Median per-residue Neff for the chain and Backbone accuracy (lDDT-Cα) for two groups of proteins: those with template coverage more than 60% of the chain and those with less than 30% of the chain. The plot reveals a positive correlation between the Median per-residue Neff and Backbone accuracy for both groups, indicating that as the Median per-residue Neff increases, so does the Backbone accuracy. However, proteins with higher template coverage consistently exhibit higher Backbone accuracy at the same Median per-residue Neff, suggesting that template coverage significantly influences Backbone accuracy.

Figure 10: Additional features that our system versions implement. (A) Prompt ablation settings (in Interactive Assistance), wherein the user can de-select metadata components for use in suggestion and question generation, to account for highly erroneous extractions or irrelevant information. (B) Figure description guidelines (both versions). These begin with general guidelines for descriptions, then plot-specific guidelines, then the semantic level framework introduced by Lundgard and Satyanarayanan [29] for data visualizations, then scatterplot-specific items, to construct a full set of guidelines for both prompting and user review. A link to the DIAGRAM Center's original guidelines is also provided. (C) After writing the full description, we implement a summarization workflow to produce more concise descriptions (both versions; one paragraph long by default). This also serves as a description review stage.