Understanding How Paper Writers Use AI-Generated Captions in Figure Caption Writing

Ho Yin (Sam) Ng[1][0000−0001−9316−6146], Ting-Yao Hsu[1][0009−0008−9082−6039], Jiyoo Min[2][0009−0002−2991−6875], Sungchul Kim[3][0000−0003−3580−5290], Ryan A. Rossi[3], Tong Yu[3][0000−0002−5991−2050], Hyunggu Jung[4][0000−0002−2967−4370], and Ting-Hao 'Kenneth' Huang[1][0000−0001−7021−4627]

[1] Pennsylvania State University, University Park PA 16802, USA
{hzn5135,txh357,txh710}@psu.edu
[2] University of Seoul, Seoul, Republic of Korea
jessy0618@uos.ac.kr
[3] Adobe Research
{sukim,ryrossi,tyu}@adobe.com
[4] Seoul National University, Seoul, Republic of Korea
hyunggu@snu.ac.kr

Abstract. Figures and their captions play a key role in scientific publications. However, despite their importance, many captions in published papers are poorly crafted, largely due to a lack of attention by paper authors. While prior AI research has explored caption generation, it has mainly focused on reader-centered use cases, where users evaluate generated captions rather than actively integrating them into their writing. This paper addresses this gap by investigating how paper authors incorporate AI-generated captions into their writing process through a user study involving 18 participants. Each participant rewrote captions for two figures from their own recently published work, using captions generated by state-of-the-art AI models as a resource. By analyzing video recordings of the writing process through interaction analysis, we observed that participants often began by copying and refining AI-generated captions. Paper writers favored longer, detail-rich captions that integrated textual and visual elements but found current AI models less effective for complex figures. These findings highlight the nuanced and diverse nature of figure caption composition, revealing design opportunities for AI systems to better support the challenges of academic writing.

Keywords: Figure caption · Scientific figure · Writing assistant · Scientific document · Human-AI interaction.

1 Introduction

Scientific publications rely on figures and their captions to convey complex information, making captions crucial for enhancing figure interpretability and accessibility. Research shows that captions improve comprehension of visuals in
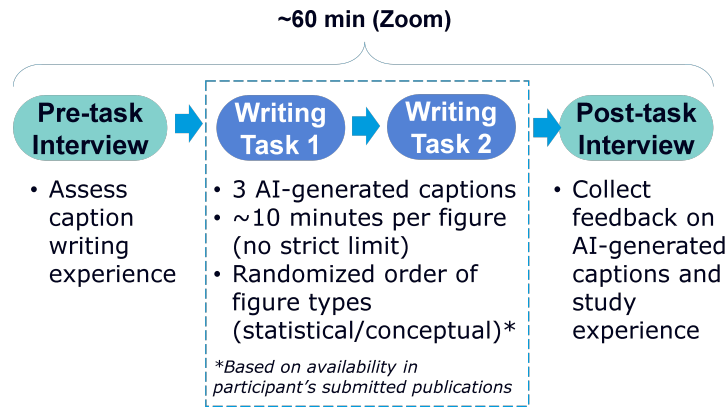
~60 min (Zoom)

Pre-task Interview
→ **Writing Task 1** → **Writing Task 2** → **Post-task Interview**

- Assess caption writing experience

- 3 AI-generated captions
- ~10 minutes per figure (no strict limit)
- Randomized order of figure types (statistical/conceptual)*

Based on availability in participant's submitted publications

- Collect feedback on AI-generated captions and study experience

Fig. 1: Overview of the user study procedure, which included interviews and two writing tasks. In each task, participants rewrote figure captions from their recently published papers. Participants were provided with the original paper's PDF (with the original caption redacted) and three AI-generated captions to assist in the rewriting process.

educational materials [25], aid students across expertise levels, and enhance accessibility by guiding readers through complex visualizations [36]. Given the importance of captions, researchers have explored using AI, particularly Large Language Models (LLMs), to generate captions for scientific figures [24,31,19,16]. Most prior work focused on *reader*-centered use cases, assuming users would only read and evaluate pre-generated captions without modifying them. However, the prevalence of poor captions in published papers—studies have shown that a significant portion of captions in arXiv papers are unhelpful [19]—stems from paper *writers*, not readers. Many paper writers do not pay attention to crafting effective captions for their figures, resulting in many low-quality captions, even in published papers.

This paper examines how **paper authors integrate captions generated by state-of-the-art figure captioning models into their writing process**. Studying the academic writing process presents unique challenges distinct from those involved in supporting academic reading. Paper authors are domain experts who have devoted substantial time to their projects, developing a deep understanding of both the content and the broader scientific context. This level of expertise often surpasses that of readers of their papers. This knowledge gap also complicates participant recruitment, as it is challenging to find participants actively writing new papers who are willing to work on their real papers during a user study. As a result, prior user studies on AI-generated captions often involved participants evaluating captions from papers written by others [50,19]. Even SciCAPENTER, the only prior work focusing on supporting figure caption writing [17], asked participants to write captions for others' papers. While this approach simplifies participant recruitment, it does not accurately reflect the

real-world scenario where researchers write captions for their own work. Additionally, scholarly authors face the unique challenge of addressing a broader and more diverse academic audience [51]. They must consider subtle and nuanced aspects of communication to effectively convey their ideas. These considerations are inherently difficult for automated systems to capture and support, making them particularly valuable to study and understand.

This paper presents a user study involving 18 participants, each tasked with rewriting captions for two figures from their own recently published work (Figure 1). For each figure, participants were provided with captions generated by state-of-the-art AI models, which they could freely use to aid in their caption-writing process. During the study, participants had access to the full content of their papers, except that the original captions for the target figures were redacted. To study their writing behavior, we recorded the entire caption-writing process on video. These recordings were then analyzed through interaction analysis [21], which involved manually coding the videos to capture participants' writing behaviors. We observed that most participants began by copying substantial portions of the AI-generated captions, subsequently refining or building upon them. Our findings also suggest that AI-generated captions were more helpful for **statistical** figures than for **conceptual**, non-statistical ones. Additionally, authors showed a preference for longer AI-generated captions that integrated both textual and visual details. However, the current AI models struggle to assist effectively with caption writing for complex figures.

Despite being a seemingly narrow aspect of academic writing, figure caption composition involves subtle nuances and exhibits significant diversity across authors. These findings highlight opportunities for designing future writing systems that better support the complex and nuanced process of caption writing.

2   Related Work

Figure Caption Generation and Evaluation. Despite the growing focus on evaluating AI models for generating captions of scientific figures, the evaluation of such models's output often overlooked the writer's perspective.

Previous studies have widely used automatic metrics to evaluate the quality of AI-generated figure descriptions. For instance, BLEU measures word overlap between the generated and reference texts [7,40,32,5,6,1,34,52,8] and, in a similar vein, ROUGE focuses on how much content from the reference text is included in the generated text [7,43,5,6,52,8]. METEOR evaluates the generated output by measuring its semantic similarity to reference translations [7,1,8], considering factors such as synonyms and morphological variations. However, automatic evaluation metrics have significant limitations. Human-written captions in scientific papers are often of low quality, leading to unreliable comparisons between machine output and manual captions [23]. Automatic evaluations do not always align well with human judgments, revealing a gap between machine scoring and actual human comprehension [11]. Finally, these comparisons miss what was not explicitly stated, including the author's intentions and considerations [18].

To address these limitations, human evaluations have been incorporated alongside automatic metrics. However, these human evaluations often fail to adequately capture the writer's intent behind the captions [22,27,33,37,34,52,8]. Instead, they focus on aspects like matching degree and reasoning correctness [33], informativeness, conciseness, coherence, and fluency [37] to understand how well the generated text describes the figures or aids reader comprehension. These methods, while valuable, miss the unique intentions and considerations of the writers throughout the writing process, including brainstorming, planning, and refinement. Motivated by this gap, our work engages real caption authors to rewrite their captions using AI-generated captions, allowing us to understand the AI tool's utility.

AI Writing Assistants. AI writing assistants have been explored in various writing contexts, from creative [42] to legal [48] and medical practice [13], highlighting the potential for Human-AI collaboration in specialized tasks. In academic writing, AI has been utilized for sentence-level suggestions [15], literature review writing [9,2], and support in critical evaluation during peer reviews [46]. The integration of AI writing assistants in academic tasks raises questions about accuracy and trust, with previous studies highlighting both potential benefits and challenges [4]. While many studies have explored AI in academic writing, specific tasks like caption generation remain understudied, particularly from the writer's perspective. Our work aims to address this gap by examining how researchers interact with and perceive AI-generated captions in scientific writing.

Human-AI Collaboration. Human-AI collaboration has gained significant attention, particularly in academic contexts where AI tools assist in various research and writing tasks. For example, Shen et al. [39] introduced the concept of bidirectional human-AI alignment, focusing on aligning AI systems with human values while also fostering the adaptation of human practices to AI capabilities. Lee et al. [29] also emphasized the importance of considering both AI capabilities and user writing processes, providing a foundation for our investigation into researchers' perspectives on AI-assisted caption writing for scientific publications. These works underscore the need for writing tools that not only generate content but also support the user's cognitive processes and decision-making.

3   User Study

We conducted a study where 18 participants rewrote captions from their published work using AI-generated captions. This section provides a detailed overview of the study's design and procedure.

3.1   Participants

Participants were recruited through the online questionnaire assessing their academic writing experience. The questionnaire collected data on research areas,

years of experience in academic writing, English language proficiency, and publication history. To ensure the quality and relevance of our study, our inclusion criteria targeted researchers with experience in scientific caption writing and at least one recent publication (Details in the 'Study Material Preparation' section), including preprints. Participants were required to upload up to three of their most recent published papers through the online questionnaire. We also set a constraint that participants' recent publications should collectively contain at least two figures. We did not impose restrictions on author order to preserve participant anonymity.

We recruited 18 participants representing diverse research areas, with the majority in Computer Science/Informatics (28%) and Human-Computer Interaction (22%), other also includes Artificial Intelligence/Robotics (17%), Energy and Minerals Engineering (6%), Mechanical Engineering (6%), Environmental Engineering (6%), Chemistry/Biochemistry (6%), Materials Science (6%), Cybersecurity (6%). Participants' age ranged from 22 to 44, with 78% between 26-29 years old. 72% of participants reported that English is not their first language. All participants had published at least one paper (including preprints) within the past three years.

3.2   Study Material Preparation

Participants provided up to three of their own papers published within the last three years, but at least one month ago. This timeframe ensured participants were familiar with their work while simulating realistic writing scenarios.

Selecting Statistical and Conceptual Figures from Participants' Papers. The selection of figure types aimed to encompass diverse caption writing requirements. While prior works have established detailed classifications spanning tables, photos, diagrams, maps, and plots [12], we simplified the figure classification into two broader categories based on previous work about qualitative and quantitative representation [3]: **statistical** and **conceptual** figures. **Statistical** figures present quantitative data through graphs or charts, while **conceptual** figures illustrate theoretical models or processes. Figure 2 shows representative examples. This binary classification offered a systematic framework for comparing caption-writing strategies across different types of visual information. For the study, we aimed to select one statistical and one conceptual figure from each participant's provided papers. However, the selection was constrained by participants' research fields, as some areas or venues predominantly feature statistical figures. While achieving an equal distribution of figure types across all participants was challenging, we made every effort to balance them. We ended up including 16 statistical figures (from 15 participants) and 18 conceptual figures (from 16 participants). Two figures from one participant were excluded due to their complexity, as they combined both statistical and conceptual elements; we still completed the user study with this participant and discussed their results in a separate subsection.
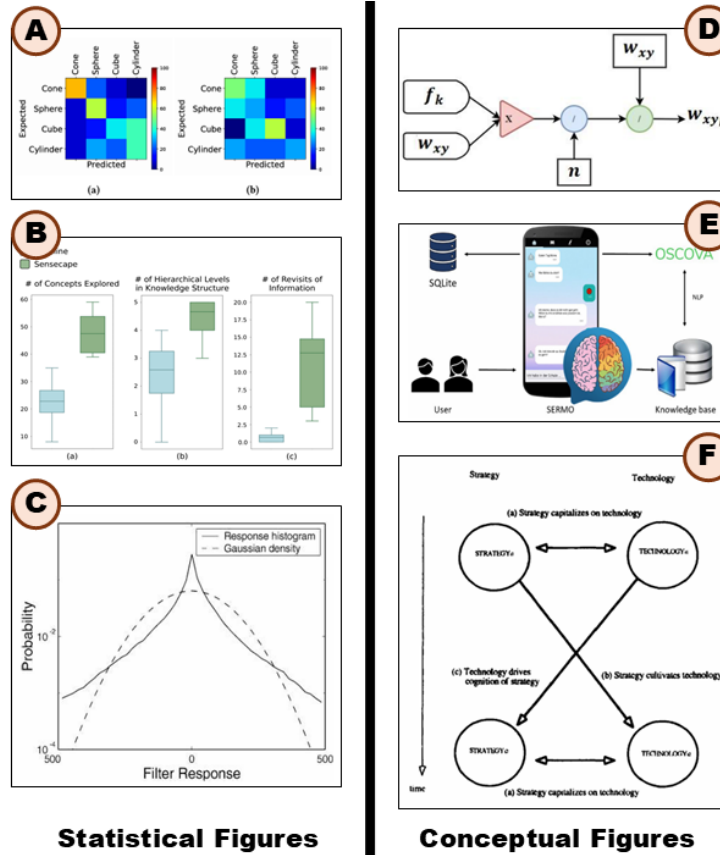
Fig. 2: Representative examples of statistical and conceptual figures, with their sources detailed in the Appendix.

Generating Captions for Selected Figures. The goal of our study is to understand how paper writers interact with AI-generated captions when composing captions for scientific figures. For this purpose, we employed state-of-the-art approaches to generate captions for the selected figures in participants' papers. Prior work has pointed out that scientific figure caption generation is not a straightforward image-to-text task but a generation task that relies heavily on the intertwined scientific context [19,50]. The context is so crucial that the figure caption generation task can be tackled as a pure text summarization task that "summarizes" all the figure-mentioning paragraphs in a paper—without even considering the figure image—into a caption [19]. In this study, we selected GPT-4o, the most recent large vision-language model, as our base model and prompted the model to take both the target figure image and figure-mentioning paragraphs as input to generate figure captions. We used a prompt similar to that of SciCapenter [17]; approaches have been shown to be effective in the SciCap Challenge
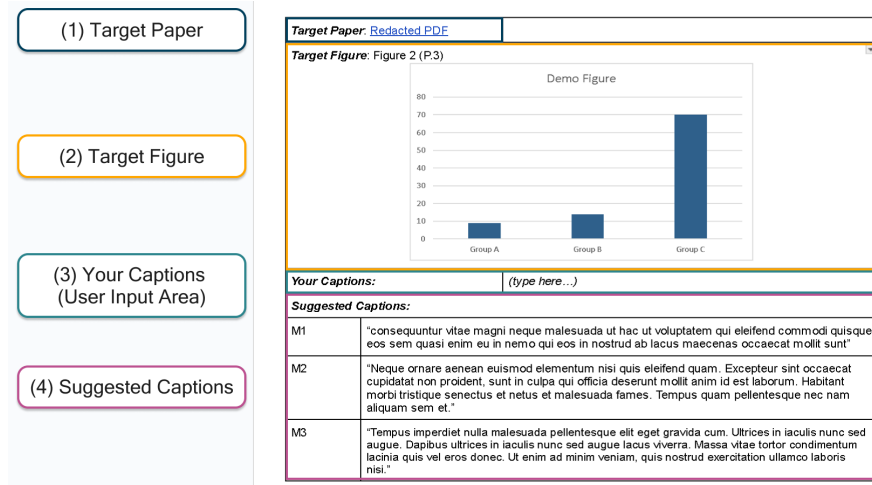
Fig. 3: User interface for the figure caption writing task, showing: (1) 'Target Paper' - the hyperlink to the redacted PDF, (2) 'Target Figure' - displaying the figure image for the writing task, with the figure number and the page number in the redacted PDF, (3) 'Your Captions' - User input area for caption writing, and (4) 'Suggested Captions' - AI-generated captions from 3 different configurations (UNLIMITED, TEXT-ONLY, *30-Word*), presented in randomized order for each caption item.

2024 [28,45,30]. We prompted GPT-4o in a zero-shot manner (no example provided) to generate captions under three distinct conditions:

1. **UNLIMITED:** In this setting, we prompted GPT-4o to generate captions using both the figure image and the paragraphs referencing the target figure (e.g., "Figure 3 shows...") as input. No length constraints were imposed, allowing GPT-4o to produce captions of any desired length.
2. **30-WORD:** This setting is identical to UNLIMITED, except that a length constraint of no more than 30 words (tokens) was applied. This specific length constraint has also been used in prior work on figure caption generation [19].
3. **TEXT-ONLY:** In this setting, we prompted GPT-4o to generate captions using only the paragraphs referencing the target figure as input, without using the figure image. Previous research has demonstrated that figure captions can be generated using textual context alone [19]. No length constraints were applied in this setting.

To determine which captions were most preferred, all AI-generated captions were presented to participants in a randomized order, without revealing the model that generated each caption. Figure 2 in the Appendix shows examples.

3.3 Study Procedure and Setups

Figure 1 shows the procedure of our user study, which contained semi-structured interviews and two writing tasks. The study was conducted remotely via Zoom and lasted approximately 60 minutes. The entire Zoom session, including participants' computer screens, video, and audio, was recorded for further analysis. The writing task was conducted in a prepared Google Doc, ensuring a consistent writing environment across all participants while allowing for a natural, unrestricted writing flow. For each figure, participants were provided with the following to work on the caption writing: *(i)* three AI-generated captions presented in randomized order, and *(ii)* the corresponding PDF file with the original caption redacted. Participants engaged in the writing task using a *think-aloud* protocol [14], in which the participants verbalize their thought processes and decision-making in real time as they perform the task. This enabled us to capture participants' cognitive processes directly, providing insight into their reasoning rather than just their final caption output.

While we estimated that each caption would take approximately 10 minutes to complete, no strict time limit was imposed to ensure a comfortable and natural writing environment. After completing each caption, participants were asked to rank the three AI-generated captions based on their perceived usefulness. They were then asked to compare their rewritten caption to the original (revealed after writing) by responding to the statement, *"My rewritten caption is better than the original caption,"* using a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

4 Analysis Methods

To understand participants' behaviors, we followed the practice of *interaction analysis* [21], a systematic method for examining video recording of human interactions. Our analysis involved manual video coding of participants' writing behaviors and their engagement with AI-generated captions. Using Dovetail video analysis software,[5] we reviewed and coded the recordings by marking specific time segments where participants demonstrated distinct writing activities. Each video segment was analyzed for verbal think-aloud protocols and writing behaviors, with segments typically defined by natural transitions in participants' activities or explicit shifts in their focus. We describe our video coding scheme in the following subsection.

4.1 Code Scheme for AI Integration Activities

Table 1 shows the coding scheme—**Copy, Adapt, Delete,** and **Correct**—used in our analysis, which was formed as follows. Following an inductive coding approach [47], our coding scheme for AI integration activities emerged

---

[5] Dovetail video analysis software: dovetail.com

Table 1: Coding scheme for AI integration activities in AI-assisted caption writing. Examples for each code are provided in the Appendix.

Code	Description
COPY	Participant copies text from an AI-generated caption without any modifications, either by directly copying and pasting or by explicitly referencing it while typing.
ADAPT	Participant significantly modifies an AI-generated caption by rewording, restructuring, or incorporating their perspective while preserving the original meaning. This may involve editing previously copied text or explicitly referencing an AI-generated caption while typing. The adaptation behavior indicates that participants perceive the underlying content of the AI-generated caption as fundamentally accurate, though requiring stylistic or structural refinement to better suit their needs.
DELETE	Participant removes portions of text from an AI-generated caption that they previously copied. These deletions may address redundancy, incorrect information, or unnecessary details.
CORRECT	Participant modified text from an AI-generated caption that they previously copied to address inaccuracies. These corrections are typically informed by the participant's understanding of the topic and the conventions of the publication venue.

through iterative analysis of the video data. We initially conducted open coding [49] to examine how participants engaged with AI-generated captions during their writing tasks, allowing patterns to emerge naturally rather than imposing predetermined categories. Through this process, we identified that participants' behaviors could be characterized by their level of reliance on and trust in AI-generated captions. These behaviors ranged from direct text copying to slight modifications of AI-generated captions, text corrections, and content deletion, which informed the final coding scheme.

4.2   Video Coding Procedure

The AI integration activities code is assigned by tracking each participant's sequence of actions in the draft area. Each subsequent action is then coded chronologically as the participant continues to develop their draft. The coding decisions are informed by both the participants' written output and their think-aloud protocol data. We systematically coded all sequential actions taken by participants during each writing task, creating a timeline of their writing process. This analysis provides a comprehensive understanding of how participants interact with and build upon AI-generated captions in scientific writing.

5   Results

In our study, we analyzed 16 statistical figures from 15 participants and 18 conceptual figures from 16 participants. Two figures from one participant were
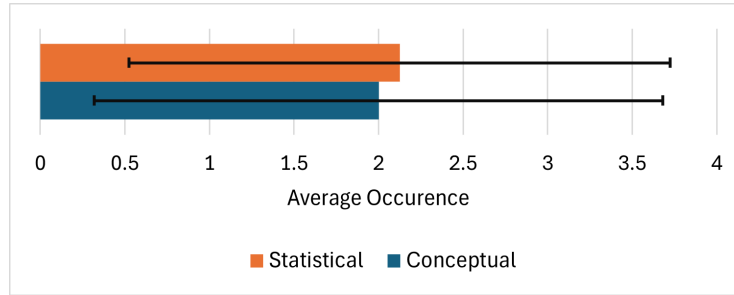
Fig. 4: Average number of interactions with AI-generated captions per writing task. Participants interacted with AI-generated captions an average of 2.13 times per session for statistical figures (SD=1.32) and 2.00 times for conceptual figures (SD=1.65).
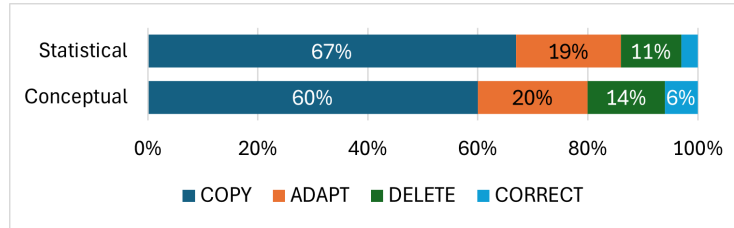

Fig. 5: Distribution of the four AI integration activity types for statistical and conceptual figures.

excluded from most analyses due to their complexity, as they combined both statistical and conceptual elements. We still completed the user study with this participant and discussed their results in a separate subsection to explore how AI-generated captions might support caption writing for complex figures.

5.1   AI-Generated Captions Integration Activities

Participants rely slightly more on AI-generated captions for statistical figures than for conceptual ones. Figure 4 shows the average frequency of AI integration activities across figure types. Participants engaged more frequently with AI generated captions when writing captions for statistical figures (mean=2.13, SD=1.32) compared to conceptual figures (mean=2.00, SD=1.65). Notably, while participants frequently used AI-generated captions, most instances involved some level of revision. Across the entire study, only one participant used an AI-generated caption unchanged as the final version for one figure. Furthermore, we also analyzed the type of integration behavior for each figure type (Figure 5). COPY was the predominant behavior for both types of figures, with marginally higher frequency in statistical figures. Additionally, DELETE and CORRECT occurred more frequently with conceptual figures than statistical figures, suggest-
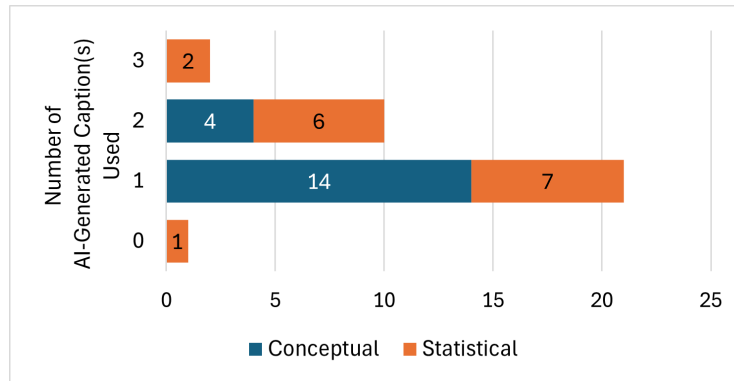
Fig. 6: Distribution of the number of AI-generated captions used in a single writing task during the figure caption writing process in our study.

ing participants were more likely to reject portions of AI-generated captions for conceptual content.

This difference may reflect LLMs' stronger ability to generate descriptive text versus their limitations in producing abstract or conceptual narratives. As one participant noted while working on a conceptual figure, "I'll get rid of this because I don't think we need to emphasize this in this figure" (P17). Another participant emphasized the need for concept revision, stating "I need to revise the facade because this is not the right way to [present] the dynamic self-design concept." (P18)

Most captions were created using a single AI-generated caption, though 35.3% involved multiple captions. Apart from the integration behavior, we also analyzed the number of AI-generated captions used for each writing task. Figure 6 shows the results. While most captions (21 out of 34) were created using a single AI-generated caption, 12 out of 34 (35.3%) captions involved multiple AI-generated captions. Meanwhile, conceptual figures were predominantly completed using one AI-generated caption (14 out of 18), with few cases integrating two AI-generated captions. In contrast, for statistical figures, half of the caption items (8 out of 16) were completed using multiple AI-generated captions during the writing process.

Most participants' primary (first) interaction with AI captions involves using or referencing an entire caption or a full sentence. Understanding how much text participants took or referenced from AI-generated captions is crucial but challenging due to the complexity of the writing process, which involves copying, revising, and deleting. The use of three AI-generated captions further complicates direct comparisons to a single caption. To simplify the analysis, we focused on participants' *first interaction* with the AI-generated captions. On average, participants interacted with the captions twice (see Figure 4), and our observations
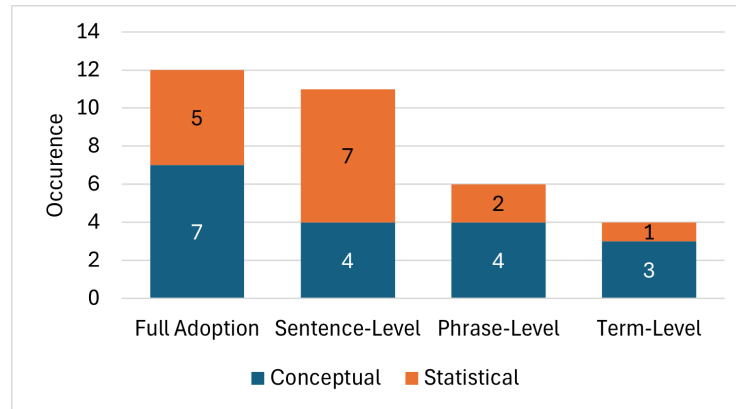
Fig. 7: Distribution of AI text usage degrees—FULL, SENTENCE, PHRASE, and TERM—in participants' initial interactions with AI-generated captions. One session (out of 34) was excluded from this analysis as the participant did not use AI-generated captions.

showed that the first interaction often involved the main caption they used in their writing.

This analysis was based on video coding developed using the coding scheme described in Table 1. Among the 34 figure-writing sessions, 26 sessions' first interaction with the AI-generated captions was COPY, 7 sessions was ADAPT, and only 1 session did not involve any interaction with AI captions. None of the sessions' first interaction with AI was CORRECT or DELETE. For each of the 33 COPY and ADAPT activities, we further coded the degree of AI text used during the interaction. The degree of use was categorized into four levels: FULL, where the entire text of an AI-generated caption was used; SENTENCE, where a complete sentence from the AI-generated caption was used; PHRASE, where a phrase or text span was taken; and TERM, where only a single word from the AI-generated caption was used. Figure 7 shows the results. Most participants' first interaction with AI captions involved using or referencing an entire caption or a full sentence, which they often revised later to improve its quality. Notably, there was only one instance where a participant used an AI-generated caption as-is without any revisions for the final outcome. The distribution of AI-generated text usage levels was similar for statistical and conceptual figures.

Participants typically start by using AI-generated captions and modify them afterward. We conducted a separate analysis of the 34 writing sessions and found that in most cases (31 out of 34), participants began their writing by copying or referencing AI-generated captions before revising them. In only 3 sessions did participants start by typing without using or referring to any AI content.
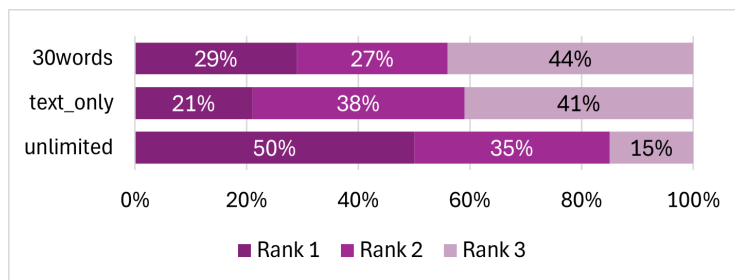
Fig. 8: Distribution of GPT-4o configurations ranked by participants, with Rank 1 indicating the best and Rank 3 the worst (N=34).

5.2   Model Preferences

Participants were asked to rank each AI-generated caption based on its usefulness in helping them compose the final caption. Figure 8 summarizes how often captions generated under each of the three conditions were ranked 1st (Best), 2nd, or 3rd (Worst) across all 34 writing sessions. Captions generated using both figure images and figure-mentioning paragraphs without a length constraint (UNLIMITED) were ranked first in 50% of cases. This suggests that participants may prefer AI-generated captions with more information, as these captions were longer and incorporated both textual and visual details. In contrast, captions generated with a 30-word length constraint (30-WORD), which also used images and text, were ranked first more often (29%) than captions generated using only text paragraphs (TEXT-ONLY) (21%). However, 30-WORD captions were also ranked worst (third) the most frequently (44%) among the three conditions. This indicates that while including both visual and textual information is important, imposing a strict length constraint may significantly reduce the usefulness of generated captions.

5.3   Perceived Caption Quality Improvement

At the end of each writing task, participants compared the quality of their rewritten captions to the original captions from their published papers (*"My rewritten caption is better than the original caption"*). As a result, 68% of the captions were rated 4 or 5, with an overall mean rating of 3.82 (SD = 1.24). However, these results should be interpreted with caution, as the assessment inherently favored the rewritten captions. Participants had access to both their original work and AI-generated suggestions, enabling them to refine their captions further. Additionally, participants were likely more experienced at the time of the study than when they initially wrote the original captions, introducing another potential bias.

5.4   Special Case: Hybrid, Complex Figures

One participant's published paper included figures that were difficult to classify as either statistical or conceptual, as they contained multiple subfigures combining both statistical and conceptual elements. While we completed the user study and analysis for this participant, their results were excluded from the main findings discussed above. In this subsection, we examine these two unique cases.

In both writing tasks, the participant relied minimally on AI-generated captions, choosing to write most of the content independently. They noted that the AI-generated captions provided limited utility, as the participant noted "One thing I realized is that it might be described as hallucination because for the first paper [...] the summary was totally different from what I discussed in the main text. So it might not be as accurate" (P04). Addressing this challenge presents a critical avenue for future research in developing AI systems capable of handling multi-component academic figures more effectively.

6   Discussion

Our results indicate that writers generally preferred captions generated from multimodal inputs with unconstrained lengths, as the additional information provided greater flexibility to choose, add, delete, and modify content. This preference for rich information suggests that future AI writing assistants should prioritize interactive suggestions and embrace a human-in-the-loop approach, given that direct adoption of AI-generated content without revision was rare.

Limitations. While our paper presents several interesting findings, we acknowledge a few limitations. First, our study was not interactive; participants could not iterate with or prompt AI models to revise their captions, which limits the generalizability of our findings. Second, participants were asked to rewrite captions for their own published papers. While this is closer to real-world caption-writing scenarios than most prior studies—a reasonable trade-off in study design—it still differs from writing captions for new, unpublished work. Finally, our generation process relied on the full context of completed and polished papers, whereas in real-world scenarios, AI-generated captions may need to be based on incomplete or even empty drafts.

7   Conclusion

This study examined how paper writers use AI-generated captions in scientific caption writing, analyzing the writing processes of 18 participants. The results showed that paper writers favored longer, detail-rich captions that integrated textual and visual elements but found current AI models less effective for complex, conceptual figures. These insights enhance our understanding of human-AI collaboration in academic writing and lay the groundwork for developing AI tools tailored to the unique demands of different figure types. Future work should focus

on specialized AI systems that adapt to researchers' diverse integration preferences, improving the efficiency and quality of scientific communication while maintaining the human element.

References

1. Agarwal, V., Ghosh, S., BSS, H., Arora, H., Raja, B.R.K.: Tricy: Trigger-guided data-to-text generation with intent aware attention-copy. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **32**, 1173–1184 (jan 2024). https://doi.org/10.1109/TASLP.2024.3353574, https://doi.org/10.1109/TASLP.2024.3353574
2. Aydın, Ö., Karaarslan, E.: Openai chatgpt generated literature review: Digital twin in healthcare. Aydın, Ö., Karaarslan, E.(2022). OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. In Ö. Aydın (Ed.), Emerging Computer Technologies **2** (2022)
3. Borgstede, M., Scholz, M.: Quantitative and qualitative approaches to generalization and replication–a representationalist view. Frontiers in psychology **12**, 605191 (2021)
4. Buruk, O.: Academic writing with gpt-3.5 (chatgpt): reflections on practices, efficacy and transparency. In: Proceedings of the 26th International Academic Mindtrek Conference. pp. 144–153 (2023)
5. Cao, J., Gong, J., Zhang, P.: Open-domain table-to-text generation based on seq2seq. In: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. ACAI '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3302425.3302484, https://doi.org/10.1145/3302425.3302484
6. Cao, J., Gong, J., Zhang, P.: Two-level model for table-to-text generation. In: Proceedings of the 2019 International Symposium on Signal Processing Systems. p. 121–124. SSPS '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3364908.3365287, https://doi.org/10.1145/3364908.3365287
7. Chen, C., Zhang, R., Kim, S., Cohen, S., Yu, T., Rossi, R., Bunescu, R.: Neural caption generation over figures. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. p. 482–485. UbiComp/ISWC '19 Adjunct, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3341162.3345601, https://doi.org/10.1145/3341162.3345601
8. Chen, M., Lu, X., Xu, T., Li, Y., Jingbo, Z., Dou, D., Xiong, H.: Towards table-to-text generation with pretrained language model: A table structure understanding and text deliberating approach. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 8199–8210. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.562, https://aclanthology.org/2022.emnlp-main.562

9. Choe, K., Park, S., Jung, S., Kim, H., Yang, J.W., Hong, H., Seo, J.: Supporting novice researchers to write literature review using language models. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–9 (2024)

10. Denecke, K., Vaaheesan, S., Arulnathan, A.: A mental health chatbot for regulating emotions (sermo)-concept and usability test. IEEE Transactions on Emerging Topics in Computing **9**(3), 1170–1182 (2020)

11. Deutsch, D., Foster, G., Freitag, M.: Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)

12. Dhote, A., Javed, M., Doermann, D.S.: A survey on figure classification techniques in scientific documents. arXiv preprint arXiv:2307.05694 (2023)

13. Escartín, C.P., O'Brien, S., Goulet, M.J., Simard, M.: Machine translation as an academic writing aid for medical practitioners. In: Proceedings of Machine Translation Summit XVI: Research Track. pp. 254–267 (2017)

14. Fonteyn, M.E., Kuipers, B., Grobe, S.J.: A description of think aloud method and protocol analysis. Qualitative health research **3**(4), 430–441 (1993)

15. Gero, K.I., Liu, V., Chilton, L.: Sparks: Inspiration for science writing using language models. In: Proceedings of the 2022 ACM Designing Interactive Systems Conference. pp. 1002–1019 (2022)

16. Hsu, T.Y., Giles, C.L., Huang, T.H.: Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624 (2021)

17. Hsu, T.Y., Huang, C.Y., Huang, S.H., Rossi, R., Kim, S., Yu, T., Giles, C.L., Huang, T.H.K.: Scicapenter: Supporting caption composition for scientific figures with machine-generated captions and ratings. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–9 (2024)

18. Hsu, T.Y., Huang, C.Y., Rossi, R., Kim, S., Giles, C., Huang, T.H.: Gpt-4 as an effective zero-shot evaluator for scientific figure captions. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5464–5474 (2023)

19. Huang, C.Y., Hsu, T.Y., Rossi, R., Nenkova, A., Kim, S., Chan, G.Y.Y., Koh, E., Giles, C.L., Huang, T.H.: Summaries as captions: Generating figure captions for scientific documents with automated text summarization. arXiv preprint arXiv:2302.12324 (2023)

20. Itami, H., Numagami, T.: Dynamic interaction between strategy and technology. Strategic Management Journal **13**(S2), 119–135 (1992)

21. Jordan, B., Henderson, A.: Interaction analysis: Foundations and practice. The journal of the learning sciences **4**(1), 39–103 (1995)

22. Kantharaj, S., Leong, R.T., Lin, X., Masry, A., Thakkar, M., Hoque, E., Joty, S.: Chart-to-text: A large-scale benchmark for chart summarization. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4005–4023. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.277, https://aclanthology.org/2022.acl-long.277

23. Kasai, J., Sakaguchi, K., Dunagan, L., Morrison, J., Le Bras, R., Choi, Y., Smith, N.A.: Transparent human evaluation for image captioning. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3464–3478 (2022)

24. Kim, D.H., Hoque, E., Agrawala, M.: Answering questions about charts and generating visual explanations. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–13 (2020)

25. Koutsikou, M., Bonoti, F., Christidou, V.: The effect of explanatory captions on understanding a scientific explanation. Int. J. Res. Educ. Methodol **7**, 1127–1138 (2015)
26. Kreimeier, J., Hammer, S., Friedmann, D., Karg, P., Bühner, C., Bankel, L., Götzelmann, T.: Evaluation of different types of haptic feedback influencing the task-based presence and performance in virtual reality. In: Proceedings of the 12th acm international conference on pervasive technologies related to assistive environments. pp. 289–298 (2019)
27. Laha, A., Jain, P., Mishra, A., Sankaranarayanan, K.: Scalable micro-planned generation of discourse from structured data. Comput. Linguist. **45**(4), 737–763 (jan 2020). https://doi.org/10.1162/coli_a_00363, https://doi.org/10.1162/coli_a_00363
28. Lee, J., Choi, H., Kwon, G., Kim, E., Lee, S., Kim, J.: Self-assessment and enhancement framework for scientific figure caption generation using llms. Tech. rep., The Second Scientific Figure Captioning Challenge (SCICAP Challenge 2024), Jeju, South Korea (August 2024), https://www.dropbox.com/scl/fi/rw7ll3aqb9p8x4dreztqs/IJCAI_2024_paper-1.pdf?rlkey=e7s47mpnldpe3b7c44zy9r2ct&dl=0, co-located with IJCAI 2024
29. Lee, M., Gero, K.I., Chung, J.J.Y., Shum, S.B., Raheja, V., Shen, H., Venugopalan, S., Wambsganss, T., Zhou, D., Alghamdi, E.A., et al.: A design space for intelligent and interactive writing assistants. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–35 (2024)
30. Li, P., Li, T., Wang, J., Wang, B., Yang, Y.: Proposal report for the 2nd scicap competition 2024. Tech. rep., 2nd Scientific Figure Captioning Challenge (SCICAP Challenge 2024), Jeju, South Korea (August 2024), https://www.dropbox.com/scl/fi/report-LI-rabbit-1.pdf?rlkey=Ly8auuerMvUP2wGEapFwvC&dl=0, co-located with IJCAI 2024
31. Liew, A., Mueller, K.: Using large language models to generate engaging captions for data visualizations. arXiv preprint arXiv:2212.14047 (2022)
32. Liu, C., Xu, Y., Yin, W., Zheng, D.: Structure-aware table-to-text generation with prefix-tuning. In: Proceedings of the 2023 4th International Conference on Control, Robotics and Intelligent System. p. 135–140. CCRIS '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3622896.3622919, https://doi.org/10.1145/3622896.3622919
33. Liu, M., Chen, D., Li, Y., Fang, G., Shen, Y.: ChartThinker: A contextual chain-of-thought approach to optimized chart summarization. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 3057–3074. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.273
34. Masry, A., Kavehzadeh, P., Do, X.L., Hoque, E., Joty, S.: UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 14662–14684. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.906, https://aclanthology.org/2023.emnlp-main.906
35. Niu, Y., Ying, L., Yang, J., Bao, M., Sivaparthipan, C.: Organizational business intelligence and decision making using big data analytics. Information Processing & Management **58**(6), 102725 (2021)

36. Nolan, D., Stoudt, S.: Captions: The unsung heroes of data communication. In: Statistics Education in the Era of Data Science Proceedings of the Satellite conference of the International Association for Statistical Education (IASE)(R. Helenius and E. Falck, eds.) (2021)

37. Obeid, J., Hoque, E.: Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In: Davis, B., Graham, Y., Kelleher, J., Sripada, Y. (eds.) Proceedings of the 13th International Conference on Natural Language Generation. pp. 138–147. Association for Computational Linguistics, Dublin, Ireland (Dec 2020). https://doi.org/10.18653/v1/2020.inlg-1.20, https://aclanthology.org/2020.inlg-1.20

38. Ramanan, K.: Interacting stochastic processes on sparse random graphs. arXiv preprint arXiv:2401.00082 (2023)

39. Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.H., Qiwei, L., et al.: Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. arXiv preprint arXiv:2406.09264 (2024)

40. Shi, W., Liu, Y., Wu, J., Liao, J.: Three-stage logical table-to-text generation based on type control. In: Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence. ACAI '22, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3579654.3579667, https://doi.org/10.1145/3579654.3579667

41. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual review of neuroscience **24**(1), 1193–1216 (2001)

42. Singh, N., Bernal, G., Savchenko, D., Glassman, E.L.: Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. ACM Transactions on Computer-Human Interaction **30**(5), 1–57 (2023)

43. Spreafico, A., Carenini, G.: Neural data-driven captioning of time-series line charts. In: Proceedings of the 2020 International Conference on Advanced Visual Interfaces. AVI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3399715.3399829, https://doi.org/10.1145/3399715.3399829

44. Suh, S., Min, B., Palani, S., Xia, H.: Sensecape: Enabling multilevel exploration and sensemaking with large language models. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–18 (2023)

45. Sun, H.L., Chen, Q.G., Zhou, D.W., Zhan, D.C., Ye, H.J.: Length-adaptive caption generation with llms. Tech. rep., 2nd Scientific Figure Captioning Challenge (SCICAP Challenge 2024), Jeju, South Korea (August 2024), https://www.dropbox.com/scl/fi/IJCAI_2024_Competition-Hailong-Sun-1.pdf?rlkey=7qqYaqAQWEqEugxKvpxe8L&dl=0, co-located with IJCAI 2024

46. Sun, L., Tao, S., Hu, J., Dow, S.P.: Metawriter: Exploring the potential and perils of ai writing support in scientific peer review. Proceedings of the ACM on Human-Computer Interaction **8**(CSCW1), 1–32 (2024)

47. Thomas, D.R.: A general inductive approach for qualitative data analysis (2003)

48. Weber, F., Wambsganss, T., Neshaei, S.P., Soellner, M.: Legalwriter: An intelligent writing support system for structured and persuasive legal case writing for novice law students. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–23 (2024)

49. Williams, M., Moser, T.: The art of coding and thematic exploration in qualitative research. International management review **15**(1), 45–55 (2019)

50. Yang, Z., Dabre, R., Tanaka, H., Okazaki, N.: Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. Journal of Natural Language Processing **31**(3), 1140–1165 (2024)
51. Yore, L.D., Hand, B.M., Prain, V.: The desired image of a science writer. (2000)
52. Zhao, Y., Zhang, H., Si, S., Nan, L., Tang, X., Cohan, A.: Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In: Wang, M., Zitouni, I. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 160–175. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-industry.17, https://aclanthology.org/2023.emnlp-industry.17

A   Examples of AI-Generated Captions

Model	Generated Caption
IMG+TEXT (30-WORD)	Comparison of mean-field (MF) and Markovian approximation (MA) models with simulation for voter model agreement probability on a 3-regular tree, showing improved accuracy of the Markovian approach.
TEXT-ONLY (UNLIMITED)	Comparison of the ad hoc Markovianized local equation and the original local equation for the voter model on a 3-regular tree truncated after 9 generations, demonstrating improved accuracy in predicting the probability of root agreement with two neighbors.
IMG+TEXT (UNLIMITED)	Comparison of Mean-Field (MF), Markovian Approximation (MA), and simulation results for the probability of agreement with two neighbors in a voter model on a 3-regular tree.

Fig. 9: GPT-4o generated captions in three different configurations: (i) GPT-4o (image+text) with a 30-word limit, (ii) GPT-4o (text-only) with unlimited length, and (iii) GPT-4o (image+text) with unlimited length. The figure is sourced from Ramanan [38].

B   Code Scheme for AI Integration Activities: Examples.

(a) Example of Copy - Retain AI suggestion in its original form

(b) Example of Adapt - Incorporates AI suggestion's core concept while modifying its linguistic expression

(c) Example of Delete - Discards AI suggestion in draft area

(d) Example of Correct - Modifies AI suggestion to ensure accuracy

Fig. 10: AI Suggestion Integration Activities in Caption Writing: Examples

C   References for Example Figures in Figure 2

- **Figure 2A:** (Denecke, Vaaheesan, and Arulnathan 2020) [10]
- **Figure 2B:** (Itami and Numagami 1992) [20]
- **Figure 2C:** (Kreimeier et al. 2019) [26]
- **Figure 2D:** (Niu et al. 2021) [35]
- **Figure 2E:** (Simoncelli and Olshausen 2001) [41]
- **Figure 2F:** (Suh et al. 2023) [44]