LAMP-CAP: Personalized Figure Caption Generation With Multimodal Figure Profiles

Ho Yin 'Sam' Ng[1] **Ting-Yao Hsu**[1] **Aashish Anantha Ramakrishnan**[1]
Branislav Kveton[2] **Nedim Lipka**[2] **Franck Dernoncourt**[2] **Dongwon Lee**[1]
Tong Yu[2] **Sungchul Kim**[2] **Ryan A. Rossi**[2] **Ting-Hao 'Kenneth' Huang**[1]
[1]The Pennsylvania State University [2]Adobe Research
[1]{sam.ng,txh357,aashish,dongwon,txh710}@psu.edu
[2]{kveton,lipka,dernonco,tyu,sukim,ryrossi}@adobe.com

Abstract

Figure captions are crucial for helping readers understand and remember a figure's key message. Many models have been developed to generate these captions, helping authors compose better quality captions more easily. Yet, authors almost always need to revise generic AI-generated captions to match their writing style and the domain's style, highlighting the need for personalization. Despite language models' personalization (LaMP) advances, these technologies often focus on text-only settings and rarely address scenarios where both inputs and profiles are multimodal. This paper introduces **LAMP-CAP**,[1] a dataset for personalized figure caption generation with multimodal figure profiles. For each target figure, LAMP-CAP provides not only the needed inputs, such as figure images, but also up to three other figures from the same document—each with its image, caption, and figure-mentioning paragraphs—as a *profile* to characterize the context. Experiments with four LLMs show that using profile information consistently helps generate captions closer to the original author-written ones. Ablation studies reveal that images in the profile are more helpful than figure-mentioning paragraphs, highlighting the advantage of using multimodal profiles over text-only ones.

1 Introduction

Figures like bar charts or line charts are widely used by scientists, companies, and governments to communicate key insights (Kim et al., 2021; Farahani et al., 2023). Captions—text placed next to these figures—are known to be crucial for helping readers understand and remember the figure's message (Tang et al., 2023; Kantharaj et al., 2022a; Meng et al., 2024). Many models have been developed to generate high-quality captions to help authors compose captions more easily (Hsu et al., 2021; Huang et al., 2023; Liu et al., 2023; Masry et al., 2023). For example, the SCICAP Challenges in 2023 and 2024 invited global teams to generate captions for scientific figures in arXiv papers (Hsu et al., 2025; Kim et al., 2025). Systems like SCICAPENTER also emerged to assist authors by providing AI-generated captions (Hsu et al., 2024). Despite these advances, studies show that authors almost always need to revise generic AI-generated captions to match their style and the domain's style, with one participant noting, "I need to revise the facade because this is *not the right way* to present (the concept)" (Ng et al., 2025a,b). This highlights the need for personalized caption generation.

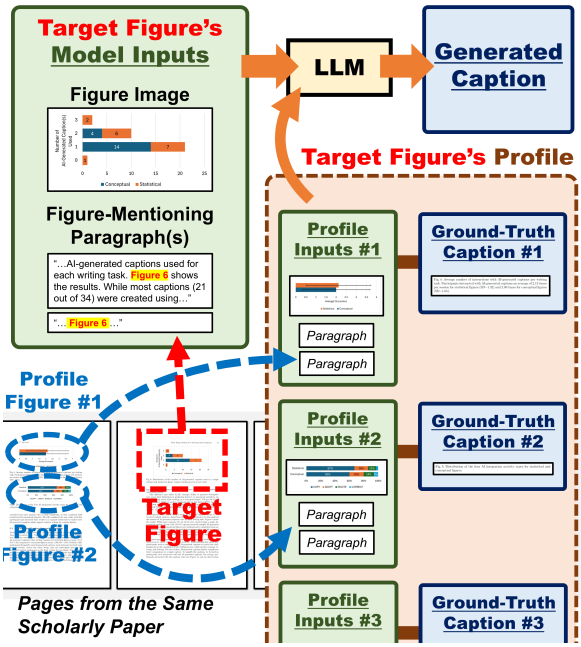


Figure 1: Overview of LAMP-CAP. For each target figure, the dataset provides multimodal *inputs*—the figure image and figure-mentioning paragraphs—and a multimodal *profile* of up to three other figures (*i.e.*, profile figures) from the same paper, each with its image, caption, and related paragraphs. The model generates a caption for the target figure using the inputs and profile.

Meanwhile, the rise of large language models (LLMs) has recently fueled interest in personal-

[1]Data: https://github.com/Crowd-AI-Lab/lamp-cap

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9818–9832
November 4-9, 2025

ized text generation (Zhang et al., 2024; Woźniak et al., 2024). Benchmarks like LaMP (Salemi et al., 2024) (LAnguage Models Personalization) and LongLaMP (Kumar et al., 2024) were created to study how LLMs can tailor text for specific contexts. However, most of these explorations focused on text-only settings, where both the input (used for generation) and profile (used for personalization) were text-based. How these text-only approaches apply to multimodal scenarios—such as figure caption generation—remains unclear.

This paper introduces **LaMP-Cap**, a dataset for **personalized figure caption generation with multimodal figure profiles** (§3). LaMP-Cap includes 110,828 target figures—scientific figures for which models aim to generate captions for—each from a distinct arXiv paper. For each target figure, LaMP-Cap provides the needed inputs (*source*)—figure images and figure-mentioning paragraphs (*e.g.*, "Figure 3 shows...")—along with up to three other figures from the same paper, each with its image, caption, and figure-mentioning paragraphs, as a *profile* to capture context. Models are then tasked with generating captions for the target figure using its image and figure-mentioning paragraphs (multimodal *source*), given a figure profile of source-caption pairs from the same paper (multimodal *profile* for personalization). We used LaMP-Cap to test caption generation with four LLMs and found that profile information consistently improved the similarity of generated captions to ground-truth captions (§4). Ablation studies revealed that captions are the most critical profile element, followed by images, with figure-mentioning paragraphs being the least important (§4.1). Our work provides a new benchmark for personalized text generation and demonstrates the effectiveness of using multimodal profiles beyond text-only approaches.

2 Related Work

Figure Caption Generation. Figure caption generation requires models to understand both the visual content and the broader context (Kantharaj et al., 2022b; Wang et al., 2024; Hu et al., 2024; Obeid and Hoque, 2020). Early approaches, like FigCAP and the initial version of SciCap, relied solely on figure images as input (Chen et al., 2020; Hsu et al., 2021). Researchers soon realized this was insufficient and began incorporating additional context, such as figure-mentioning paragraphs and even the document's title or abstract (Huang et al., 2023; Yang et al., 2024; Stokes et al., 2022). Despite this progress, prior work often overlooked personalization. Although studies noted that users often need captions tailored to their style or domain (Hsu et al., 2025; Huang et al., 2023), none of these approaches explicitly provided source-target pairs that capture the specific generation context needed for models to learn personalized styles. A few studies have explored creative personalization of image captions (Shuster et al., 2019; Anantha Ramakrishnan et al., 2025), but these approaches relied on explicit style inputs, making them dependent on user-provided style descriptions.

Personalized LLMs. Personalization of LLMs has gained attention (Zhang et al., 2024), primarily in two directions: *(i)* personalized text generation (tailoring generated text for specific contexts) and *(ii)* downstream task personalization (enhancing targeted applications like recommendation systems). We focus on the first direction, defining the personalization target as a *group* of users—all co-authors of a paper—rather than individuals. Prior work, such as the LaMP-5 task (Salemi et al., 2024) on Personalized Scholarly Title Generation, has also treated a paper's author group as a single entity for personalization. Most prior work in this space has been centered on text-only settings (§1). For example, LaMP included tasks such as news headline generation and email subject creation—relying exclusively on text-based inputs and profiles (Salemi et al., 2024). How these approaches extend to multimodal scenarios remains an open question.

3 LaMP-Cap Dataset

We constructed LaMP-Cap by curating the SciCap Challenge Dataset (Hsu et al., 2025). We first selected all papers containing at least two figures. From each paper, we then randomly designated one figure as the **target figure** (the one needing a caption) and used the remaining figures from that paper (up to a maximum of three, since the SciCap Challenge allowed at most four figures per paper) as the **profile** to provide personalization context.

Following the SciCap Challenge Dataset's split (*i.e.*, 80/10/10 train/val/test), LaMP-Cap includes 110,828 target figures: 86,197 for training, 12,361 for validation, and 12,270 for testing. Among these, 54,680 (49.3%) had one profile figure, 26,193 (23.6%) had two, and 30,027 (27.1%) had three, totaling 197,075 profile figures. Papers with only one figure were excluded. See Appendix A for details.

4 Experimental Results

Experiment Setups. We evaluated four LLMs on personalized caption generation using LAMP-CAP:[2] *(i)* GPT-4o (Hurst et al., 2024), *(ii)* Llama 4 Scout (MetaAI, 2025), *(iii)* Gemini 2.5 Flash Preview (DeepMind, 2024), and *(iv)* GPT-4.1 Mini (OpenAI, 2024). The first three are larger models, while the last one is smaller. We used OpenAI's API for GPT-4o and OpenRouter (openrouter.ai) for the others. We focused on LLMs because large-scale human evaluations from the SCICAP Challenge showed a clear performance gap between model classes (Hsu et al., 2025): only LLMs like GPT-4V consistently generate captions matching or exceeding those by human authors, while smaller or specialized models such as PEGASUS (Zhang et al., 2020a) and UniChart (Masry et al., 2023) perform poorly.

Building on prior work showing that more profile information improves performance (Tan et al., 2024), we tested four caption generation settings with varying amounts and sources of profile input: **(1) No Profile:** The model generated captions using only the target figure's image and figure-mentioning paragraphs. **(2) One Profile:** The model used the same source as in (1) but additionally used *one* randomly selected profile figure from the *same* paper for personalization. **(3) All Profile:** The model used the same source as in (1) but additionally used *all* profile figures from the *same* paper for personalization. **(4) Other Profile:** The model used the same source as in (1) but additionally used one or three randomly selected profile figure(s) from random *other* papers. The setup (4) tests whether performance gains come from paper-specific context or generic in-domain examples. See Appendix B for the full prompt.

We cleaned the output by removing unnecessary reasoning steps or explanations. We also removed cases (56 out of 12,259) where models failed to generate valid output. See Appendix C and Appendix D for details.

Using profile information (from the same paper) makes captions more similar to ground truth, especially with all profile figures. Table 1 shows the personalized caption generation results of four LLMs, evaluated using BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004).[3] We used reference-based metrics to measure how closely the generated captions matched the original author-written captions, following a standard evaluation approach for personalized text generation used in well-known work like LongLaMP (Kumar et al., 2024). The results show that incorporating profile information consistently improves caption quality across all four models. Additionally, using all profile figures provides better results than using just one. See Appendix E for details.

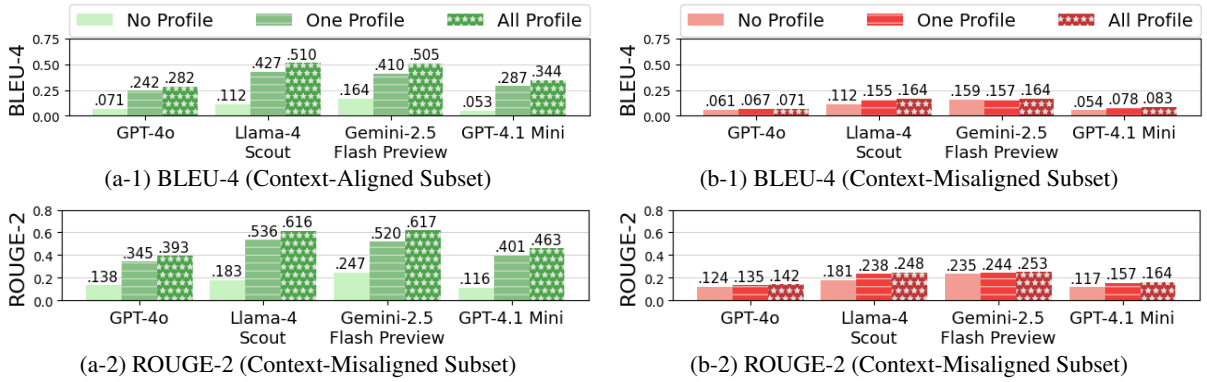
LLM	Profile: Same Paper	Profile: # ≤ 3	BLEU: B-1	BLEU: B-2	BLEU: B-3	BLEU: B-4	ROUGE: R-1	ROUGE: R-2	ROUGE: R-L
GPT-4o	N/A	0	.219	.133	.091	.063	.321	.127	.248
	N	1	.189	.108	.070	.046	.289	.101	.218
		3	.206	.118	.076	.051	.300	.107	.223
	Y	1	.279	.186	.137	.103	.384	.178	.313
		All	.292	.200	.150	.115	.397	.194	.328
Llama-4 Scout	N/A	0	.254	.178	.138	.112	.357	.182	.293
	N	1	.223	.167	.132	.107	.348	.180	.295
		3	.251	.187	.152	.129	.368	.202	.317
	Y	1	.372	.293	.246	.211	.481	.300	.423
		All	.396	.318	.270	.235	.503	.324	.447
Gemini-2.5 Flash Preview	N/A	0	.305	.230	.188	.160	.417	.237	.361
	N	1	.268	.198	.159	.132	.388	.213	.335
		3	.279	.205	.164	.137	.398	.217	.343
	Y	1	.370	.291	.244	.209	.482	.301	.426
		All	.395	.317	.270	.234	.504	.328	.449
GPT-4.1 Mini	N/A	0	.209	.124	.081	.054	.305	.117	.225
	N	1	.216	.133	.091	.063	.325	.132	.250
		3	.220	.136	.093	.065	.328	.134	.252
	Y	1	.286	.202	.155	.121	.398	.207	.326
		All	.300	.218	.171	.137	.412	.225	.342

Table 1: Performance of LLMs on caption generation across profile settings. The highest scores are achieved by using **all** available profile(s) from the **same** paper.

Using profiles from other papers often lowered BLEU and ROUGE scores, though there were exceptions. Table 1 suggests that performance gains primarily came from paper-specific context, rather than generic in-domain examples. Profiles from other papers generally hurt performance, but some models, such as GPT-4.1 Mini, showed slight

[2] Qwen-2.5-VL-7B-Instruct was excluded from our main analysis due to significantly higher failure rate (2.2%, 269/12,259) compared to other models. See Appendix C.

[3] We also explored BERTScore (Zhang et al., 2020b), which correlated highly with BLEU and ROUGE. See Appendix E.



(a-1) BLEU-4 (Context-Aligned Subset)

(b-1) BLEU-4 (Context-Misaligned Subset)

(a-2) ROUGE-2 (Context-Misaligned Subset)

(b-2) ROUGE-2 (Context-Misaligned Subset)

Figure 2: BLEU-4 and ROUGE-2 scores on LAMP-CAP's Context-Aligned and Context-Misaligned subsets, showing that personalization is most effective when profile caption are similar to the target (a-1, a-2).

LLM	Same Type	BLEU				ROUGE		
		B-1	B-2	B-3	B-4	R-1	R-2	R-L
GPT-4o	No	.233	.142	.097	.069	.340	.134	.270
	Yes	.302	.208	.157	.121	.406	.201	.336
Llama-4 Scout	No	.325	.245	.199	.167	.436	.250	.377
	Yes	.396	.317	.269	.233	.504	.325	.447
Gemini-2.5 Flash Preview	No	.326	.246	.201	.169	.440	.254	.384
	Yes	.393	.314	.266	.229	.503	.325	.447
GPT-4.1 Mini	No	.237	.154	.109	.079	.349	.155	.276
	Yes	.311	.227	.178	.142	.422	.234	.351

Table 2: LLM performance on figures with one profile figure. Personalization is more effective when the single profile figure shares the same type as the target.

improvements. Furthermore, using more *other* profiles tended to reduce the performance drop or, occasionally, provide minor gains.

When profile figures shared the same type as the target figure, personalization works better. To examine how figure type affects personalization, we analyzed cases with a single profile figure, splitting them into two groups: those where the profile and target figure types matched (n=8,083) and those where they did not (n=4,120). Table 2 shows that matching the figure type resulted in captions that were significantly closer to the gold caption.

Personalization is more effective when profile captions are highly similar to the target caption. To test if personalization is more effective when profiles are similar to the target, we split our test set into two groups. We calculated the similarity (using BERTScore and ROUGE-L) between each target caption and its available profile captions. The top 25% of examples with the most similar profiles formed our Context-Aligned set (n=2,513); the remainder formed the Context-Misaligned set. The results in Figure 2 confirm our hypothesis. Performance gains from using profiles were substantially larger for the Context-Aligned group (Figures 2a-1 and 2a-2), while the impact was noticeably smaller for the Context-Misaligned set ((Figures 2b-1 and 2b-2). (See Appendix F.)

Profile #	BLEU				ROUGE		
	B-1	B-2	B-3	B-4	R-1	R-2	R-L
0	.212	.124	.082	.055	.302	.113	.223
1	.289	.198	.145	.109	.390	.181	.312
2	.319	.215	.159	.121	.411	.198	.331
3	.332	.231	.175	.136	.424	.215	.345

Table 3: GPT-4o performance with varying numbers of profile figures. Scores improve as more profiles are added, with the largest gain from 0 to 1 profile.

Adding more profile figures consistently improved performance, but the largest gain came from the first profile. We examined how the number of profiles affected caption quality using GPT-4o. A subset figure with exactly three profiles from the same paper was used (n=3,424). For the 1-Profile and 2-Profile settings, profiles were randomly sampled. Table 3 showed clear diminishing returns as more profiles were added.

4.1 Ablation Study

Captions are the most important profile element, while images are more influential than paragraphs. To assess the importance of each profile element, we conducted an ablation study on the test set using the GPT-4o model with the One Profile setting. We tested three conditions by removing one profile element at a time: *(i)* figure captions (No Caption), *(ii)* figure images (No Image), and *(iii)* figure-mentioning paragraphs (No Paragraph). Figure 3 shows the results. Removing captions had the most significant impact, as captions directly

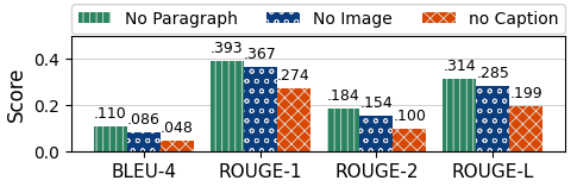


Figure 3: Ablation study on the impact of profile elements using GPT-4o. Results show a clear hierarchy of importance: caption > image > paragraph.

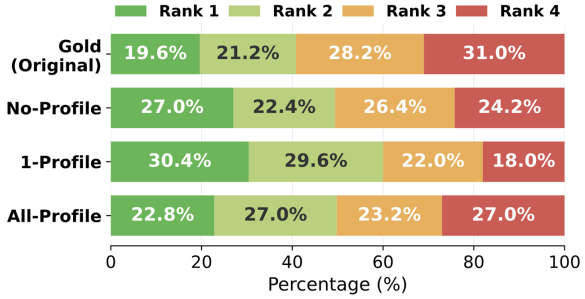


Figure 4: Distribution of human preference rankings (lower is better). A Friedman's test confirmed a statistically significant difference between the model configurations ($\chi^2 = 28.48$, $p < 0.001$).

guide generation. Removing images also reduced performance more than removing paragraphs, highlighting the greater influence of visual information. Appendix G shows the detailed results.

4.2 Human Evaluation

We recruited 10 US-based NLP researchers (PhD students with publication or review experience) for our human evaluation. The study involved ranking captions for 50 figures randomly selected from the arXiv *cs.CL* domain. For each figure, we provided the figure's image, the paper's title, and abstract as context. Participants then ranked four corresponding captions based on how well each one helped them understand the figure. The four captions were: *(i)* the original *Gold* caption, and *(ii) No-Profile*, *(iii) 1-Profile*, and *(iv) All-Profile* settings generated by our best-performing model (**Gemini**, based on automatic metrics). See Appendix H for details.

1-Profile was the most preferred condition. Our human evaluation result shows a clear preference for *1-Profile* setting, which achieved the best average rank of 2.27. The other models followed in order: *No-Profile* (2.48) and *All-Profile* (2.54). Interestingly, the author-written *Gold* captions were ranked last overall with an average rank of 2.71. Such outcome is reflected in the preference distribution (Figure 4): The *1-Profile* configuration received the most first-place votes (30.6%) and the fewest last-place votes (17.8%). On contrast, the *Gold* captions were ranked last most frequently (31.2%). Post-hoc Friedman-Nemenyi tests showed that 1-Profile and No-Profile captions were significantly preferred over gold ($p < 0.001$ and $p = 0.03$, respectively).

Trade-offs between human-perceived quality and similarity to gold captions. While the All-Profile setting generated captions with higher reference similarity (Table 1), human judges significantly preferred captions from 1-Profile ($p = 0.006$). This suggests that optimizing for similarity may also reproducing flaws from the inconsistently quality's caption from arXiv reference (Huang et al., 2023), reducing the perceived quality.

5 Discussion

Our results with LAMP-CAP show that including figure images in profiles improves personalized caption generation, and that more profile information makes captions closer to the original author-written captions. Although we focused on personalized text generation, Zhang et al. noted strong links between LLM-based personalized text generation and downstream applications such as recommendation systems, suggesting that multimodal profiles could also benefit tasks like multimodal recommendation. Our findings also echo challenges noted by Zhang et al., such as reduced LLM effectiveness when profiles lack similarity—a problem linked to cold-start scenarios in low-resource settings.

We further highlight the limitations of automatic metrics for evaluating personalized text generation. As shown in prior work (Salemi et al., 2025), n-gram-based scores, such as BLEU and ROUGE, often fail to reflect human judgments of quality accurately. We hope our work, along with the LAMP-CAP dataset, motivates the community to explore multimodal profiles and broaden the scope of LLM personalization.

6 Conclusion and Future Work

We introduced LAMP-CAP, a new dataset for personalized caption generation for scientific figures using multimodal profiles, and showed that profiles make captions more personalized across four language models. Future work includes expanding profile components, exploring cross-domain generalization, and developing writer-centric evaluation metrics. We are also developing a caption writing assistant that generates personalized captions by analyzing users' local document context.

7 Limitations

We acknowledge several limitations in this work.

- First, our approach assumes that each figure has profile figures from the same arXiv paper, but this is not always true, especially for papers with only one figure, which we excluded. This assumption also limits the method's usefulness in early-stage writing, when context for personalization is sparse—a classic example of the "cold start" problem in personalization. However, because authors often write captions late in the process (Ng et al., 2025a), our method is well-suited to assist at this critical stage.

- Second, our work only used basic figure selection strategies, such as random choice or matching by the same type, rather than more advanced strategies to further optimize the outcomes. Our primary goal was to introduce the concept of the dataset and to encourage further research on personalized text generation with multi-modal profiles by demonstrating that even basic strategies yield promising results.

- Third, we did not include individual author information in personalization profiles because most papers are co-authored, and different figures and captions may be written by different authors. Although author-based personalization could be explored using their past works, the collaborative nature of academic writing makes this difficult.

- Fourth, we recognize the risk of data contamination when testing LLMs on public datasets. Personalized text generation tasks, including our own, have historically relied on existing datasets as their data source. For this study, we built LAMP-CAP on the well-established and widely used SCICAP Challenge dataset for figure caption generation. Because this dataset is derived from publicly available arXiv papers, eliminating contamination risks entirely is difficult. That said, our work is the first to explore multimodal profiles for scientific figure captioning, and we believe the trade-off is justified. Supporting this view, the SCICAP Challenge's human evaluation paper (Hsu et al., 2025) ran a small study on newly published arXiv papers to test contamination effects. Their findings showed that model preference rankings remained consistent on unseen data, suggesting contamination does not undermine the validity of results.

- Finally, our automatic evaluation focused on caption similarity to original captions, which does not guarantee caption quality. As suggested by our human evaluation results (§4.2), high similarity indicates that profiles capture context and style, but it does not ensure the captions are useful for readers. Future work could additionally explore automatic evaluation approaches, such as LLMs-as-judges methods, to assess caption quality and usefulness more meaningfully. Scaling the reliable but expensive human evaluation is also an interesting direction.

8 Ethics Statements

Using LLMs to generate text inherently carries risks, including producing inaccurate or misleading information. In scholarly contexts, such errors could mislead readers. Our approach minimizes this risk by involving paper authors, who should review and revise generated captions. If captions are presented to readers without human validation—contrary to our intent—the system should clearly indicate that the captions are AI-generated, not written by the original authors.

Acknowledgments

We thank the participants of our caption evaluation study for their time and effort. We are also grateful to the anonymous reviewers for their constructive feedback and to the Alfred P. Sloan Foundation for their generous support of this research (Grant Number: 2024-22721).

References

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. 2025. RONA: Pragmatically diverse image captioning with coherence relations. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 74–86. Association for Computational Linguistics.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. 2020. Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1545.

Google DeepMind. 2024. Gemini 2.5 flash preview. https://deepmind.google/technologies/gemini/. Large language model preview by Google.

Ali Mazraeh Farahani, Peyman Adibi, Mohammad Saeed Ehsani, Hans-Peter Hutter, and Alireza Darvishy. 2023. Automatic chart understanding: a review. *IEEE Access*, 11:76202–76221.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting-Yao Hsu, Chieh-Yang Huang, Shih-Hong Huang, Ryan Rossi, Sungchul Kim, Tong Yu, C Lee Giles, and Ting-Hao Kenneth Huang. 2024. 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*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Ting-Yao E Hsu, Yi-Li Hsu, Shaurya Rohatgi, Chieh-Yang Huang, Ho Yin Sam Ng, Ryan Rossi, Sungchul Kim, Tong Yu, Lun-Wei Ku, C Lee Giles, and 1 others. 2025. Do large multimodal models solve caption generation for scientific figures? lessons learned from scicap challenge 2023. *arXiv preprint arXiv:2501.19353*.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6929–6938.

Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunyee Koh, C Lee Giles, and Ting-Hao Huang. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 80–92, Prague, Czechia. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022a. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Dae Hyun Kim, Vidya Setlur, and Maneesh Agrawala. 2021. Towards understanding how readers integrate charts and captions: A case study with line charts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Jaeyoung Kim, Jongho Lee, Hong-Jun Choi, Ting-Yao Hsu, Chieh-Yang Huang, Sungchul Kim, Ryan Rossi, Tong Yu, Clyde Lee Giles, Ting-Hao'Kenneth' Huang, and 1 others. 2025. Multi-llm collaborative caption generation in scientific documents. In *International Workshop on AI for Transportation*, pages 142–160. Springer.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, and 1 others. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7775–7803, Bangkok, Thailand. Association for Computational Linguistics.

MetaAI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: May 2025.

Ho Yin Sam Ng, Ting-Yao Hsu, Jiyoo Min, Sungchul Kim, Ryan A Rossi, Tong Yu, Hyunggu Jung, and Ting-Hao Huang. 2025a. Understanding writing assistants for scientific figure captions: A thematic analysis. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 1–10.

Ho Yin Sam Ng, Ting-Yao Hsu, Jiyoo Min, Sungchul Kim, Ryan A Rossi, Tong Yu, Hyunggu Jung, and Ting-Hao Kenneth Huang. 2025b. Understanding how paper writers use ai-generated captions in figure caption writing. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4.1 mini. `https://openai.com/index/gpt-4-1/`. Large language model.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. Expert: Effective and explainable evaluation of personalized long-form text generation. *Preprint*, arXiv:2501.14956.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12508–12518. IEEE.

Chase Stokes, Vidya Setlur, Bridget Cogley, Arvind Satyanarayan, and Marti A Hearst. 2022. Striking a balance: Reader takeaways and preferences when integrating text and charts. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1233–1243.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520. IEEE.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2024. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

A LaMP-Cap Dataset Details

Figure 5 provides a detailed breakdown of figure type across each data split. Figure 6 provides a detailed distribution across each data split.

B Prompts

In this section, we provide the prompt we used in Section 4. [IMG-TARGET] and [PARA-TARGET] represent encoded images and figure-mentioning paragraphs from target figures. [num_profiles] indicates the number of profiles used, while [profile_index] denotes a specific profile's index. [IMG-PROFILE], [PARA-PROFILE], and [CAP-PROFILE] correspond to encoded images, figure-mentioning paragraphs, and captions from profile figures, respectively.

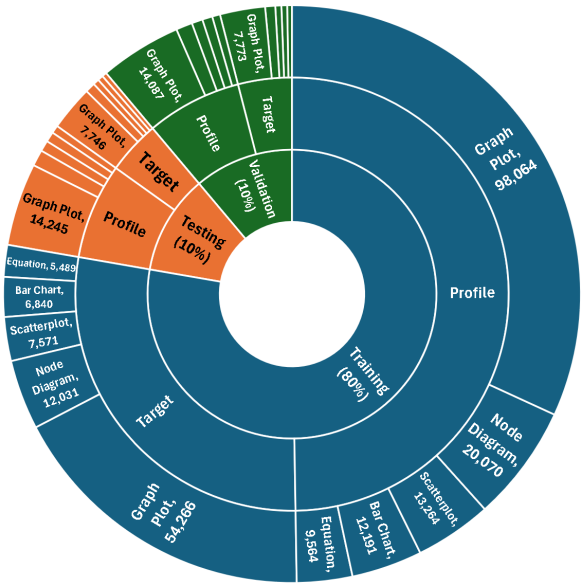


Figure 5: Data split of LAMP-CAP by figure type. The dataset contains 307,903 figures from 110,828 scientific papers, split into training (80%), validation (10%), and testing (10%) sets. Each set includes target and profile figures. The five main figure types are a) Graph Plot, b) Node Diagram, c) Equation, d) Bar Chart, and e) Scatterplot. Graph plots are the most common figure type across all splits.

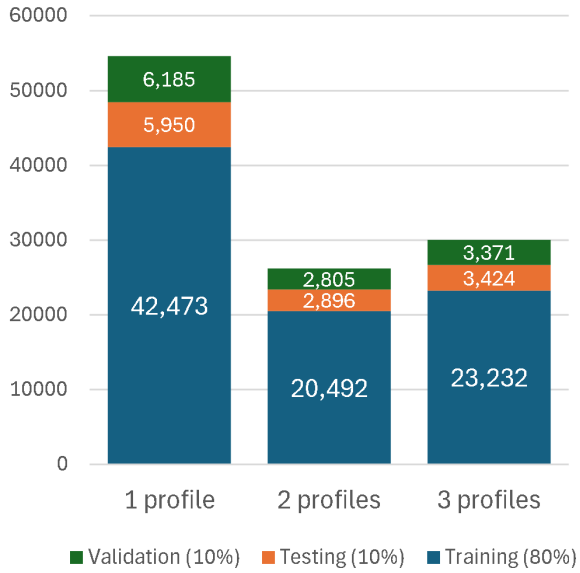


Figure 6: Profile distribution in LAMP-CAP, showing the number of target figures with 1, 2, or 3 profile figures.

Prompt with No Profile. The following prompt was used for the baseline condition without profile information:

```
Your task is to generate a caption for
    the Target Figure. We will provide
    you with the image of the Target
    Figure, labeled as 'Target Figure
    Image', and the paragraphs that
    mention the Target Figure, labeled
    as 'Target Figure Paragraph(s)',
    from the same paper.

The elements for the Target Figure will
    be labeled as follows:
- Target Figure Image:[IMG-TARGET],
- Target Figure Paragraph(s): [PARA-
    TARGET].
```

Prompt with Profile. The following prompt was used with profile information:

```
We will present you with the captions,
    images, and paragraphs referencing [
    num_profiles] scientific figures
    from the same paper. These elements
    will be labeled as follows:
- Profile Figure [profile_index]:
-- Image [profile_index]: [IMG-PROFILE],
-- Paragraph [profile_index]: [PARA-
    PROFILE],
-- Caption [profile_index]: [CAP-PROFILE
    ].
```

```
Your task is to carefully analyze the
    content, tone, structure, and
    stylistic elements of these captions
     and associated text. Based on this
    analysis, generate a caption for the
     Target Figure, maintaining the same
     writing style. We will provide you
    with the image of the Target Figure,
     labeled as 'Target Figure Image',
    and the paragraphs that mention the
    Target Figure, labeled as 'Target
    Figure Paragraph(s)', from the same
    paper. The elements for the Target
    Figure will be labeled as follows:
- Target Figure Image:[IMG-TARGET],
- Target Figure Paragraph(s): [PARA-
    TARGET].
```

C Generation Output Cleaning Procedure

We performed data output cleaning in three steps.

1. We manually examined cases with BLEU or ROUGE scores of 0 to identify data issues. We identified 11 cases (out of 12,270) where the original captions were incorrectly captured due to parsing errors—either missing the real caption content or capturing the wrong text. In one instance, the parser mistakenly captured the paper authors name from the bottom-right corner instead of the caption beneath the figure on the same page. These cases were excluded from evaluation.

2. We used GPT-4o-mini to clean the generated captions, removing irrelevant text such as rea-

soning steps, figure indexes, or any pre- and suffix that did not belong to the actual caption. Our cleaning instructions explicitly specified not to add any additional text or explanations to the raw output. The following prompt was used:

```
You are a helpful assistant that
    extracts the actual caption from
     text. Ignore any reasoning or
    explanation text and only return
     the actual caption content.
    Also, remove any prefix that
    indicates figure index such as '
    Caption:', 'Fig.', 'Figure', or
    any numbering before the main
    caption text. Extract ONLY the
    actual caption from this text,
    without any additional text or
    explanation: [GENERATED-CAPTION]
```

Below is an example from arXiv paper 1105.0392, showing the response from the "Llama-4 Scout" model with "All-Profile" configuration, before and after cleaning:

RAW OUTPUT The following is the raw output before cleaning:

```
Based on the provided elements, I
↪ will generate a caption for the
↪ Target Figure.

Target Figure Caption: A
↪ deterministic algorithm for
↪ the 1-dimensional case chooses
↪ a good sensor at each step.

This caption maintains the same
↪ writing style as Caption 1:
↪ "Four similar rhombi form a set
↪ of regions for which no
↪ stateless algorithm can be
↪ competitive." It is concise,
↪ informative, and provides a
↪ brief description of the Target
↪ Figure.

The tone and structure of the
↪ caption are consistent with the
↪ provided caption, and it
↪ effectively conveys the main
↪ idea of the Target Figure.

Please let me know if you need
↪ further adjustments!
```

EXTRACTED CAPTION The following is the extracted caption after cleaning:

```
A deterministic algorithm for the
↪ 1-dimensional case chooses a
↪ good sensor at each step.
```

After GPT-4o cleaning, we randomly sampled 100 outputs for human evaluation. Using binary labels (GOOD/BAD), we assessed whether the extracted captions were correct. All 100 sampled extractions were labeled as GOOD, confirming the cleaning procedure's effectiveness.

3. We employed keyword filtering with manual verification to filtered out failed generation, including blank responses. Detailed examples of these error cases are documented in Table 4. After cleaning, we identified a total of 56 unique problematic cases across all models and configurations (out of 12,259), which were excluded from further analysis.

   The Qwen-2.5-VL-7B-Instruct model was excluded from the main comparative analysis due to its high rate of generation failures, particularly when using the profile-based configurations. Specifically, this model failed in 2.19% of All-Profile cases and 0.83% of One-Profile cases, whereas the highest failure rate for any other model was just 0.07%. For completeness, we nevertheless report its baseline performance in Table Table 5.

D Text Preprocessing and Evaluation

For text normalization before evaluation, we implemented a custom preprocessing pipeline using standard Python libraries that: (1) converts text to lowercase, (2) removes all punctuation, and (3) normalizes whitespace.

For evaluation, we used standard NLP metrics implemented in Python packages: `NLTK` (version 3.9.1) for BLEU scores (with `SmoothingFunction` for smoothing) and Google's `rouge_scorer` (version 0.1.2) for ROUGE metrics. We used the default parameters for both packages. The specific implementations were imported directly from `nltk.translate.bleu_score` and `rouge_score` modules.

Model and Config	Cases	Examples of Invalid Generations
Qwen_All-Profile	269	Null output "---" "Caption for Target Figure:", "**Caption for the Target Figure:**" "PLEASE, provide the image of the Target Figure, so that I can..." "Based on the analysis of the provided captions, images, and paragraphs, your task..."
Qwen_One-Profile	102	Null output "Could you give me the image of the Target Figure labeled 'Target Figure Image'?" "Your analysis shows us your own comprehensive and detailed interpretation..."
Qwen_No-Profile	31	Null output
Gemini_No-Profile	24	"The provided paragraphs do not mention the Target Figure." "nan", "None" "Sorry, I lack the necessary information to generate a caption..." "Please provide the Target Figure Image and the Target Figure Paragraph(s)..." "no caption found" "we are unable to generate a caption for this figure..."
Gemini_One-Profile	9	"image 1", "Target Figure Image" "there is no caption to extract"
Gemini_All-Profile	9	"image 1", "image"
Llama_No-Profile	6	"There is no caption provided in the text."
Llama_One-Profile	8	"target", "target figure" "There is no caption provided in the text." "Since the Target Figure Image does not contain any specific data or information..."
Llama_All-Profile	4	"target", "target figure" "not applicable"
4.1 Mini_One-Profile	1	"no caption provided"

Table 4: Examples of invalid generation across different language models and profile configurations

LLM	Profile		BLEU				ROUGE		
	Same Paper	# ≤ 3	B-1	B-2	B-3	B-4	R-1	R-2	R-L
Qwen-2.5-VL-7B-Instruct	N/A	0	.198	.117	.079	.056	.295	.110	.228
	Same	1	.257	.174	.133	.105	.348	.168	.285
		All	.262	.181	.140	.112	.353	.175	.290

Table 5: Performance of the Qwen model on caption generation with varying profile settings[a]

[a] The cross-paper source analysis was not performed for this model due to its high failure rate in initial experiments compared to other LLMs

E Detail about Caption Evaluation

This appendix session is to supplement the result in Section 4 regarding the main study of caption generation using different profile configuration across 4 models.

Figure 9 shows the BLEU-4 distribution across different language models and profile configurations.

Figure 10 shows the ROUGE-2 distribution across different language models and profile configurations.

Table 9 shows BERTScore semantic similarity between the generated and gold captions. The results correlate strongly with our BLEU and ROUGE metrics, confirming the same performance trends. The *All-Profile* setting achieves the highest score, followed closely by *1-Profile*. Notably, this analysis reveals a pattern of diminishing returns. The performance leap from *No-Profile* to *1-Profile* is substantially larger than the subsequent gain from *1-Profile* to *All-Profile*, indicating that while the first profile provides a significant semantic boost, the marginal benefit of additional profiles is less substantial.

F Context-Alignment Data Partition

This appendix provides additional details on the Context-Aligned and Context-Misaligned subset partitioning and evaluation described in Section 4.

Figure 7 presents the distribution of BERTScore and ROUGE-L between target and profile captions in the **LAMP-CAP** Test Set.

Table 6 shows performance metrics for the Context-Aligned Subset across different LLMs and profile configurations.

Table 7 presents performance metrics for the Context-Misaligned Subset across different LLMs and profile configurations.

G Detailed Result of Ablation Study

This appendix session is to supplement the finding in subsection 4.1 regarding the ablation study. Table 8 shows the detailed result of ablation study.

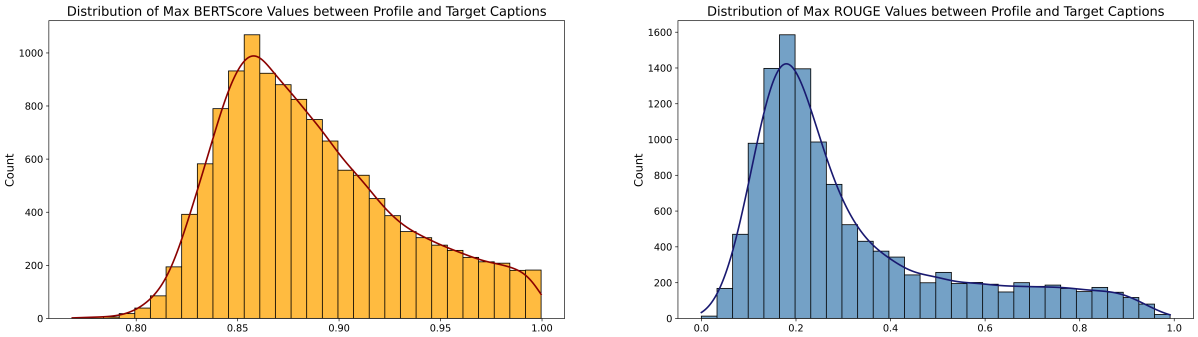


Figure 7: Distribution of BERTScore (left) and ROUGE-L (right) metrics between Target and Profile captions in the **LaMP-Cap** Test Set. Both these scores share a left-shifted skewed unimodal distribution. The BERTScore plot shows that the provided profile captions for each target are *very semantically related*. On the other hand, the broader spread of ROUGE-L scores shows that profile captions exhibit *low lexical overlap*. High semantic relatedness and lexical variety motivates our use of profile captions as key style indicators for personalization.

LLM	Profile Used	BLEU				ROUGE		
		B-1	B-2	B-3	B-4	R-1	R-2	R-L
GPT-4o	No	.226	.144	.101	.071	.321	.138	.259
	One	.437	.347	.289	.242	.533	.345	.484
	All	.478	.391	.332	.282	.573	.393	.530
Llama-4 Scout	No	.249	.178	.139	.112	.347	.183	.296
	One	.589	.523	.473	.427	.676	.536	.646
	All	.666	.605	.556	.510	.744	.616	.720
Gemini-2.5 Flash Preview	No	.319	.241	.195	.164	.459	.247	.379
	One	.576	.507	.456	.410	.664	.520	.635
	All	.659	.600	.551	.505	.742	.617	.717
GPT-4.1 Mini	No	.188	.116	.078	.053	.282	.116	.220
	One	.449	.379	.330	.287	.554	.401	.511
	All	.496	.433	.387	.344	.600	.463	.565

Table 6: Performance on LaMP-Cap Context-aligned Subset (n=2,513) across LLMs and profile configurations.

LLM	Profile Used	BLEU				ROUGE		
		B-1	B-2	B-3	B-4	R-1	R-2	R-L
GPT-4o	No	.217	.131	.088	.061	.320	.124	.245
	One	.238	.144	.097	.067	.345	.135	.269
	All	.244	.150	.102	.071	.351	.142	.275
Llama-4 Scout	No	.255	.178	.138	.112	.360	.181	.292
	One	.316	.233	.187	.155	.430	.238	.366
	All	.326	.243	.196	.164	.440	.248	.376
Gemini-2.5 Flash Preview	No	.301	.227	.186	.159	.414	.235	.357
	One	.317	.235	.189	.157	.434	.244	.371
	All	.326	.244	.197	.164	.443	.253	.380
GPT-4.1 Mini	No	.215	.127	.082	.054	.312	.117	.226
	One	.243	.156	.109	.078	.357	.157	.278
	All	.249	.162	.114	.083	.364	.164	.284

Table 7: Performance on LaMP-Cap Context-Misaligned Subset (n=9,690) across LLMs and profile configurations.

H Setup of Human Evaluation Study

This appendix session provides supplementary materials for the human evaluation study described in subsection 4.2. Figure 8 shows the interface of the user study in Microsoft Form.

To supplement the

Our protocol was reviewed and approved by the Institutional Review Board (IRB) of The Pennsylvania State University (STUDY00025214), which granted a waiver of written documentation of consent. Consent was implied by participants voluntary action of proceeding with and completing the survey. Each participant received $20 cash compensation upon completion of the study. The following is the instruction shown to the human participant before they start the study.

```
We are conducting a human evaluation
    study on the quality of captions
    generated for scientific figures.
For each question, you will be shown the
     paper's title and abstract to
    provide context, followed by a
    figure and four caption options.
Your task is to rank the captions based
    on how well they help you understand
     the figure.

Some captions may be generated with the
    assistance of AI. However, the goal
    is not to identify which captions
    are human- or AI-written. Please
    focus only on the clarity and
    usefulness of each caption in
    conveying the figure's message.

There are no right or wrong answers;
    please use your own judgment. The
    survey includes 50 figures and takes
```

Elements	BLEU				ROUGE		
	B-1	B-2	B-3	B-4	R-1	R-2	R-L
No Paragraph	.299	.199	.146	.110	.393	.184	.314
No Image	.273	.171	.119	.086	.367	.154	.285
No Caption	.189	.109	.071	.048	.274	.100	.199

Table 8: Result from Ablation Study.

Model	No-Profile	1-Profile	All-Profile
GPT-4o	.844	.860	.863
Llama-4 Scout	.856	.873	.876
Gemini-2.5 Flash Preview	.865	.874	.877
GPT-4.1 Mini	.844	.860	.863

Table 9: Performance of various LLMs on figure caption generation, as measured by BERTScore.

```
    about 1.5 hours to complete. Please
    use a desktop computer only. To
   ensure fair evaluation, please avoid
    searching for the original papers
   online while completing the task.
```

And for each figure, we ask them to do the ranking of the captions with the following prompt:

```
Please rank the four captions below
    based on how well they help you
    understand the figure.
Drag and drop to reorder them from 1 (
    best) to 4 (worst) using your mouse.
```

I Disclosure of AI Assistance

We used Perplexity and Gemini to facilitate proofreading and text refinement.

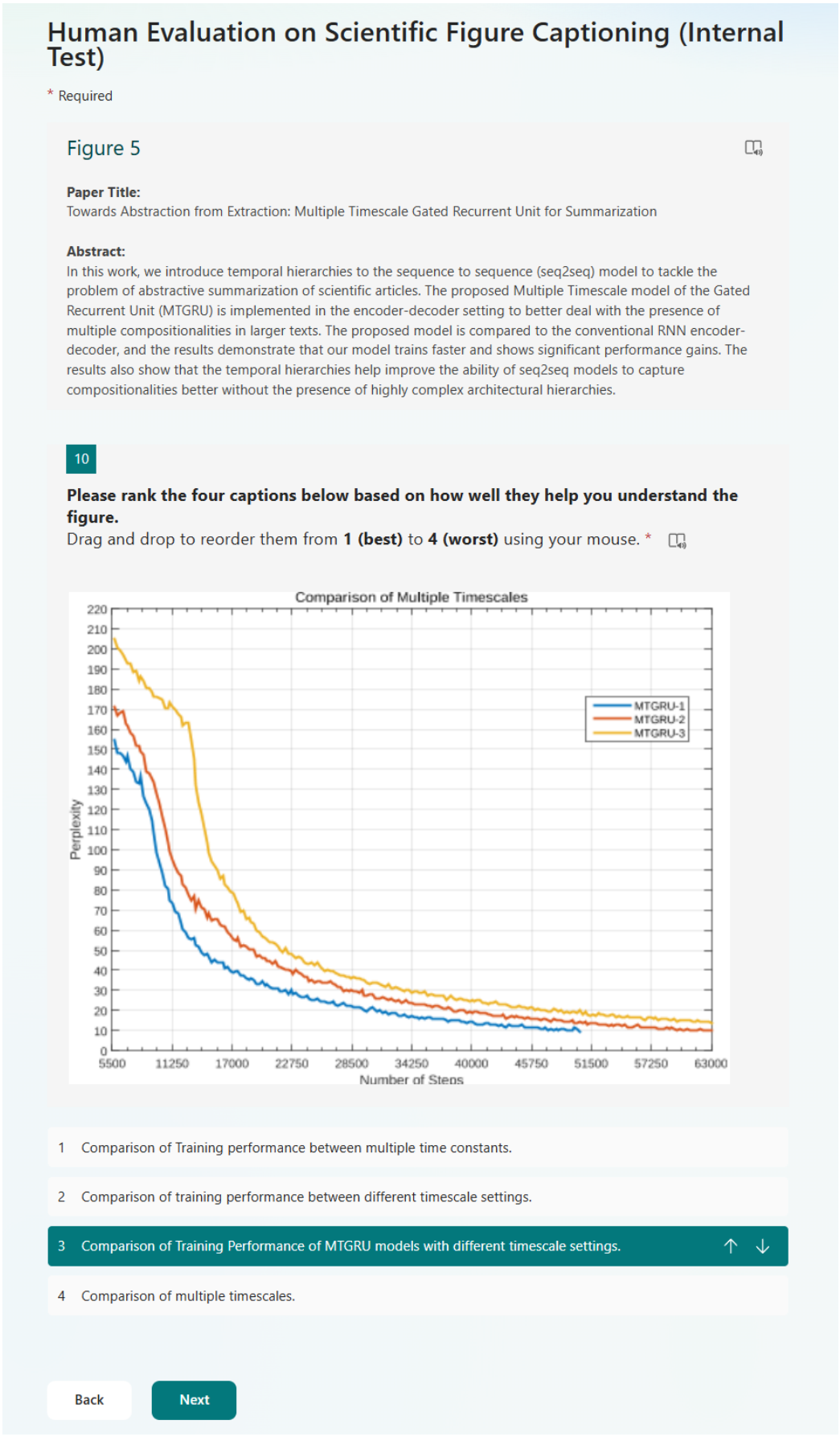
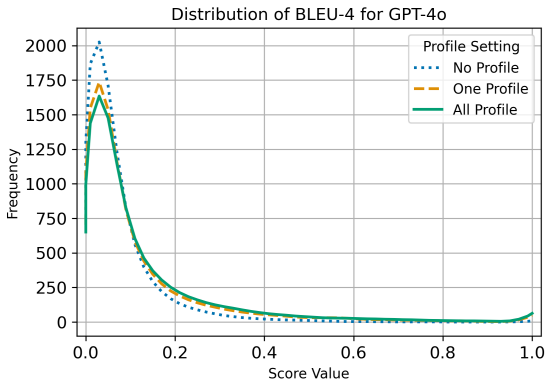
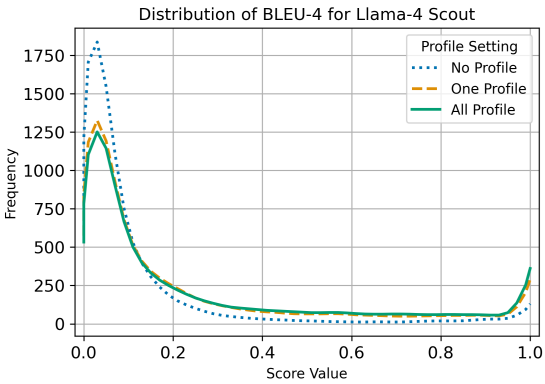


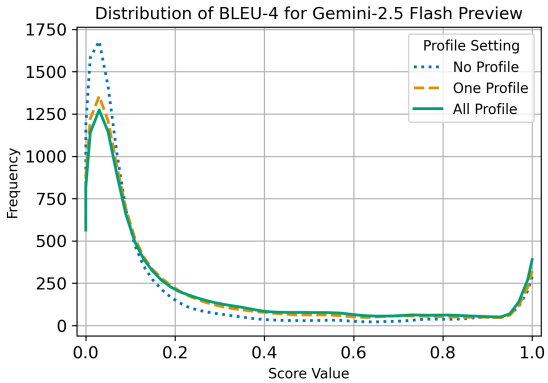
Figure 8: The user interface for our human evaluation study.



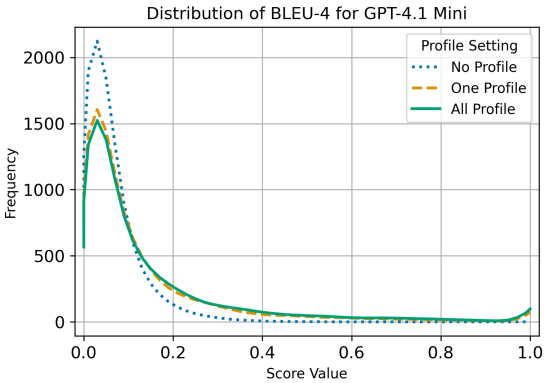
(a) BLEU-4 for GPT-4o



(b) BLEU-4 for Llama-4 Scout

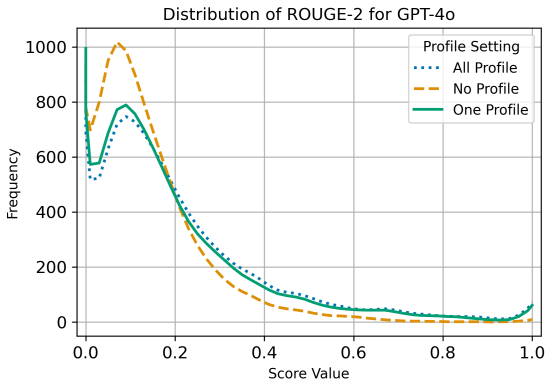


(c) BLEU-4 for Gemini-2.5 Flash Preview

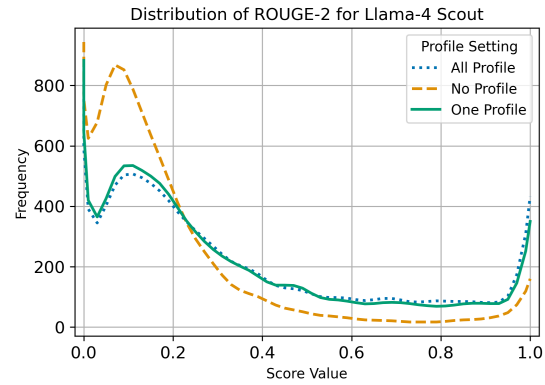


(d) BLEU-4 for GPT-4.1 Mini

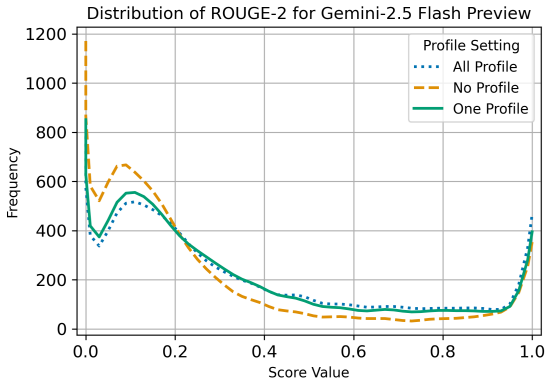
Figure 9: Distribution of the BLEU-4 across different LLMs and profile configuration.



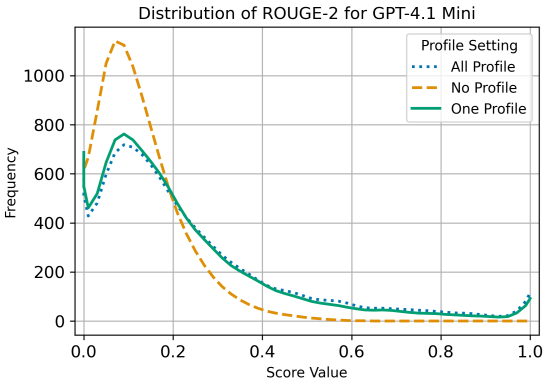
(a) ROUGE-2 for GPT-4o



(b) ROUGE-2 for Llama-4 Scout



(c) ROUGE-2 for Gemini-2.5 Flash Preview



(d) ROUGE-2 for GPT-4.1 Mini

Figure 10: Distribution of the ROUGE-2 across different LLMs and profile configuration.