LaMP-Cap：个性化图像说明生成结合多模态图像概况

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]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

图例对于帮助读者理解和记住图形的关键信息至关重要。许多模型已被开发用于生成这些图例，帮助作者更容易地撰写质量更好的图例。然而，作者几乎总是需要修改通用的人工智能生成的图例以匹配他们的写作风格和领域风格，这突出了个性化的需求。尽管语言模型的个性化（LaMP）取得了进展，这些技术通常专注于纯文本环境，很少涉及输入和配置文件均为多模态的场景。本文介绍了 LaMP-Cap，[1]，一个用于个性化图例生成的数据集，具有多模态图形配置文件。对于每个目标图形，LaMP-Cap 不仅提供了所需的输入，例如图形图像，还提供了来自同一文档的另外三个图形，每个图形都有其图像、图例和提及该图形的段落，作为配置文件来表征其上下文。四个大型语言模型的实验表明，使用配置文件信息一致有助于生成更接近原作者撰写的图例。消融研究揭示，配置文件中的图像比提及图形的段落更有帮助，突出多模态配置文件相对于仅文本配置文件的优势。

1 相关工作

图注生成。 图注生成要求模型理解视觉内容和更广泛的上下文 (Kantharaj et al., 2022; Wang et al., 2024; Hu et al., 2024; Obeid and Hoque, 2020)。早期的方法，如 FigCAP 和 SciCap 的初始版本，仅依赖于图像作为输入 (Chen et al., 2020; Hsu et al., 2021)。研究人员很快意识到这不足够，并开始加入额外的上下文，如提到图表的段落，甚至是文档的标题或摘要 (Huang et al., 2023; Yang et al., 2024; Stokes et al., 2022)。尽管有这些进展，之前的工作常常忽视个性化。虽然研究指出用户常常需要定制化到他们风格或领域的图注 (Hsu et al., 2025; Huang et al., 2023)，但这些方法没有明确提供能够让模型学习个性化风格所需的具体生成上下文的源-目标对。一些研究探索了图像注释的创造性个性化 (Shuster et al., 2019; Anantha Ramakrishnan et al., 2025)，但这些方法依赖于显式的风格输入，因此依赖于用户提供的风格描述。

个性化大型语言模型。 大型语言模型的个性化引起了关注，主要有两个方向 (Zhang et al., 2024)：(*i*) 个性化文本生成，它根据特定情境调整生成的文本，和 *(ii)* 下游任务个性化，它增强了推荐系统等目标应用。我们的工作专注于第一个方向。在这个领域之前的大多数工作都集中在仅使用文本的环境中（§**??**）。例如，LaMP 涉及任务包括新闻标题生成和电子邮件主题创建——完全依赖于基于文本的输入和轮廓 (Salemi et al., 2024)。这些方法如何扩展到多模态场景仍然是一个悬而未决的问题。

我们整理了 SciCap 挑战数据集来构建 LaMP-Cap (Hsu et al., 2025)。SciCap 挑战数据集包括 476,389 个图形，每个图形都有图像、标题和提到该图的段落，从 231,675 篇 arXiv 论文中提取出来。为了创建 LaMP-Cap，我们专注于包含至少两个图形的论文。对于每篇符合条件的论文，我们随机选择一个图形作为目标图形，模型必须为其生成标题。同一篇论文中的其他图形作为资料，提供个性化标题生成的背景。由于 SciCap 挑战数据集限制每篇 arXiv 论文最多包含四个图形，以减少数据大小并支持计算资源有限的小实验室参与，LaMP-Cap 中的每个目标图形最多可以有三个资料图形。

根据 SciCap 挑战数据集的分割（*i.e.*，80/10/10 训练/验证/测试），LaMP-Cap 包括 110,828 个目标图：其中 86,197 用于训练，12,361 用于验证，12,270 用于测试。这些目标图中，54,680 个（49.3 %）有一个轮廓图，26,193 个（23.6 %）有两个，30,027 个（27.1 %）有三个，总计 197,075 个轮廓图。只有一个图的论文被排除，因为至少需要两个图来形成目标-轮廓对。按照图类型和轮廓数的详细数据分割见 Appendix A。

[1]LaMP-Cap 数据集和代码可在 GitHub 上获取：https://github.com/Crowd-AI-Lab/lamp-cap。

www.xueshuxiangzi.com

LLM	Profile Used	BLEU				ROUGE		
		B-1	B-2	B-3	B-4	R-1	R-2	R-L
GPT-4o	No	.219	.133	.091	.063	.321	.127	.248
	One	.279	.186	.137	.103	.384	.178	.313
	All	.292	.200	.150	.115	.397	.194	.328
Llama-4 Scout	No	.254	.178	.138	.112	.357	.182	.293
	One	.372	.293	.246	.211	.481	.300	.423
	All	.396	.318	.270	.235	.503	.324	.447
Gemini-2.5 Flash Preview	No	.305	.230	.188	.160	.417	.237	.361
	One	.370	.291	.244	.209	.482	.301	.426
	All	.395	.317	.270	.234	.504	.328	.449
GPT-4.1 Mini	No	.209	.124	.081	.054	.305	.117	.225
	One	.286	.202	.155	.121	.398	.207	.326
	All	.300	.218	.171	.137	.412	.225	.342

Table 1: LLM 在不同配置下的字幕生成性能。使用配置文件图——尤其是所有可用图——在所有模型上一致提升了性能。

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 在只有一个轮廓图的图形上的表现。当轮廓图和目标图属于同一类型时（n=8,083），个性化效果比它们属于不同类型时（n=4,120）更加显著。

2 实验结果

我们使用 LAMP-CAP 评估了四个 LLM 用于个性化字幕生成：*(i)* GPT-4o (Hurst et al., 2024)，(*ii*) Llama 4 Scout (MetaAI, 2025)，*(三)* Gemini 2.5 Flash Preview (DeepMind, 2024)，以及 *(iv)* GPT-4.1 Mini (OpenAI, 2024)。前三个是大型模型，而最后一个则较小。我们使用 OpenAI 的 API 来进行 GPT-4o，其他的则使用 OpenRouter（openrouter.ai）。基于之前的研究表明更多的个人资料信息可以改善性能 (Tan et al., 2024)，我们测试了三种不同的字幕生成设置，输入的个人资料量不同：(1) 无个人资料：模型仅使用目标图像及提到图像的段落生成字幕。(2) 一个个人资料：模型使用 (1) 中的相同来源，但另外使用一个随机选择的个人资料图进行个性化。(3) 所有个人资料：模型使用 (1) 中的相同来源，但另外使用所有个人资料图进行个性化。[2]（完整的提示见 **??**）。生成后，我们通过去除不必要的推理步骤或不属于实际字幕的说明来清理结果。我们还去除了模型未能生成任何输出的情况（在 12,259 个样例中有 56 个）。Appendix B 详细描述了数据清理程序，而 **??** 描述了本研究中使用的评估包。

使用概况信息使标题更接近真实值，尤其是在所有概况数据的情况下。 Table 1 显示了四个大型语言模型（LLMs）的个性化字幕生成结果，并使用 BLEU 和 ROUGE 进行评估。我们使用基于参考的指标来衡量生成的字幕与原作者撰写的字幕的相似程度，这种方法遵循了在例如 LongLaMP (Kumar et al., 2024) 等知名工作中使用的个性化文本生成的标准评估方法。

[2]我们避免将我们的方法称为“零样本”或“少样本”，因为概要图和标题作为隐含的例子。

结果表明，结合个人资料信息始终可以改善所有四个模型的字幕质量。此外，使用所有个人资料数据比仅使用一个提供更好的结果。详细的模型性能分布及个人资料配置将在 **??** 中展示。

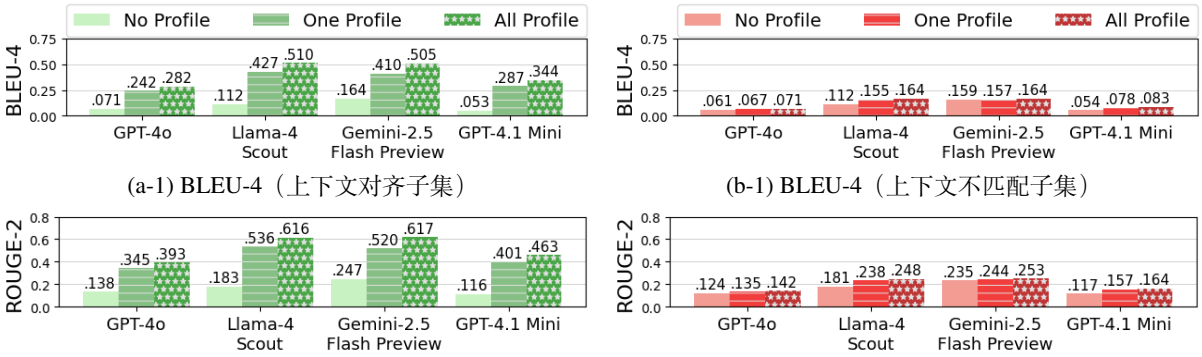
当个人档案与目标图形类型相同，个性化效果更佳。 SCICAP Challenge 数据集提供了图表类型（*e.g.*，图表图，散点图），使我们能够研究图表类型如何影响个性化标题生成。我们关注于目标图表仅有一个参考图表的情况，将其分为两组：参考图表类型与目标图表类型相同（是）和不同（否）。Table 2 显示，当参考图表类型与目标图表类型相同时，生成的标题更接近于真实标题。

当个人资料说明与目标说明高度相似时，个性化效果更佳。 受人物类型研究（Table 2）启发，我们调查了类似目标标题的个人资料标题是否能改善个性化。对于测试集中每个目标图，我们使用 BERTScore (Zhang et al., 2020) 计算语义相似性，并使用 ROUGE-L (Lin, 2004) 计算词汇相似性，在个人资料标题与目标标题之间。（**??** 显示了详细的得分分布。）然后我们按每个目标图的最高相似性得分对测试集进行排序，并选择在两个指标中相似性都高的前 25 % 个数据点，创建了包含 2,513 个目标图的上下文对齐集。测试集中剩下的 9,690 个目标图形成了上下文不对齐集。Figure 1 显示，当至少一个个人资料标题与真实标题高度相似时，个性化更加有效（图 1a-1 和 1a-2）。尽管在上下文不对齐集（图 1b-1 和 1b-2）中，个人资料仍然有帮助，但它们的影响明显较小。详细结果见 **??**。

2.1 消融研究

为了评估每个配置文件元素的重要性，我们使用 GPT-4o 模型在" 单一配置文件" 设

www.xueshuxiangzi.com



(a-1) BLEU-4（上下文对齐子集）

(b-1) BLEU-4（上下文不匹配子集）

(a-2) ROUGE-2（上下文不匹配子集）

(b-2) ROUGE-2（上下文未对齐子集）

Figure 1: 在 LAMP-CAP 的上下文对齐和上下文错位子集上，不同大型语言模型（LLMs）和配置设置对应的 BLEU-4 和 ROUGE-2 分数。当至少一个配置说明与目标说明紧密匹配时，个性化更为有效（a-1, a-2）。

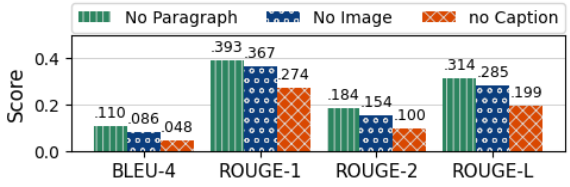


Figure 2: 逐步移除一个概要元素（标题、图片或图形提及段落）进行标题生成的消融研究。标题贡献最大，其次是图片，最后是段落。

置下，对测试集进行了消融研究。我们通过每次去掉一个配置文件元素来测试三种情况：去掉图示标题（无标题），去掉图示图像（无图像），以及去掉提到图示的段落（无段落）。Figure 2 展示了结果。去掉标题对性能的影响最大，因为标题直接指导生成。去掉图像对性能的降低也超过去掉段落，凸显了视觉信息的更大影响。?? 展示了详细结果。

3 讨论

我们的 LAMP-CAP 结果表明，在个人档案中加入视觉元素（如图像）可以提升个性化字幕生成的效果，并且更多的档案信息进一步改善了性能。虽然我们的研究仅专注于个性化文本生成，但 Zhang et al. 指出基于 LLM 的个性化文本生成与下游应用（如推荐系统）之间有深刻联系，这表明多模态档案可能会对包括多模态推荐在内的超文本生成以外的任务带来益处。我们的发现也呼应了 Zhang et al. 指出的挑战，例如在档案缺乏相似性时 LLM 的有效性降低，这个问题与低资源环境中的冷启动场景有关。我们希望我们的结果能鼓励研究界探索多模态档案以实现更广泛的 LLM 个性化。

4 结论和未来工作

我们引入了 LAMP-CAP，这是一个新的数据集，用于使用多模态配置文件生成科学图表的个性化标题，并表明配置文件可以使标题在四种语言模型中变得更加个性化。未来的工作包括扩展配置文件组件、探索跨领域泛化以及进行人工评估。我们还在开发一个标题写作助手，通过分析用户自己的文档上下文生成个性化标题。

我们承认这项工作存在若干限制。

- 首先，我们的方法假设每个图都有来自同一篇 arXiv 论文的配置文件图，但这并不总是成立，特别是对于只有一个图的论文，我们将其排除。这个假设也限制了在论文写作早期阶段的实际使用，因为此时个性化的上下文可能很少。
- 其次，我们没有在个性化资料中包含个人作者信息，因为大多数论文是合作撰写的，不同的图表和说明可能由不同的作者撰写。尽管可以通过他们的过去作品来探索基于作者的个性化，但学术写作的协作性质使这变得困难。
- 第三，尽管使用了一个较小的 LLM（GPT-4.1 Mini）来降低数据污染风险，但由于使用已发布的数据（SCICAP 挑战数据集），仍然存在一些风险。
- 最后，我们的评估集中在与原始标题的相似性上，这并不保证标题质量。高相似性表明这些简介抓住了上下文和风格，但这并不能确保标题对读者有用。未来的工作应包括人工评估，以评估标题的质量和实用性。

使用

5 伦理声明

生成文本固有地带有风险，包括产生不准确或误导性的信息。在学术背景下，这样的错误可能会误导读者。我们的方法通过让论文作者参与进来以审查和修改生成的说明文字，来尽量减少这种风险。如果这些说明文字在没有人类验证的情况下展示给读者——这与我们的意图相违背——系统应清楚地表明这些说明文字是由 AI 生成的，而非由原作者撰写的。

www.xueshuxiangzi.com

致谢 我们感谢阿尔弗雷德 · P · 斯隆基金会对本研究的慷慨支持（资助编号：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.

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 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. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

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.

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

www.xueshuxiangzi.com

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6476–6491.

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.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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 数据集详情

Figure 3 提供了每个数据分割的图表类型的详细分类。Figure 4 提供了每个数据分割的详细分布。

在本节中，我们提供了我们在 **??** 中使用的提示。[IMG-TARGET] 和 [PARA-TARGET] 表示来自目标图的编码图像和提及图的段落。[num_profiles] 表示使用的配置文件数量，而 [profile_index] 表示特定配置文件的索引。[IMG-PROFILE]、[PARA-PROFILE] 和 [CAP-PROFILE] 分别对应于来自配置文件图的编码图像、提及图的段落和图的说明。

没有个人资料的提示。 以下提示用于没有个人信息的基线条件：

```
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].
```

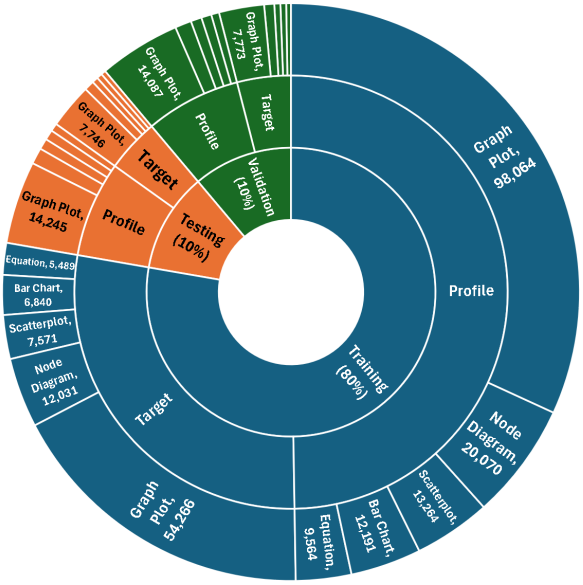


Figure 3: 按图形类型划分 LaMP-Cap 数据。该数据集包含来自 110,828 篇科学论文的 307,903 个图形，划分为训练集（80 %）、验证集（10 %）和测试集（10 %）。每个集合都包括目标图和轮廓图。五种主要的图形类型是：a) 图形绘图，b) 节点图，c) 方程，d) 条形图，e) 散点图。在所有划分中，图形绘图是最常见的图形类型。

带有简介的提示。 使用了以下提示和档案信息：

```
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
    ],
```

www.xueshuxiangzi.com

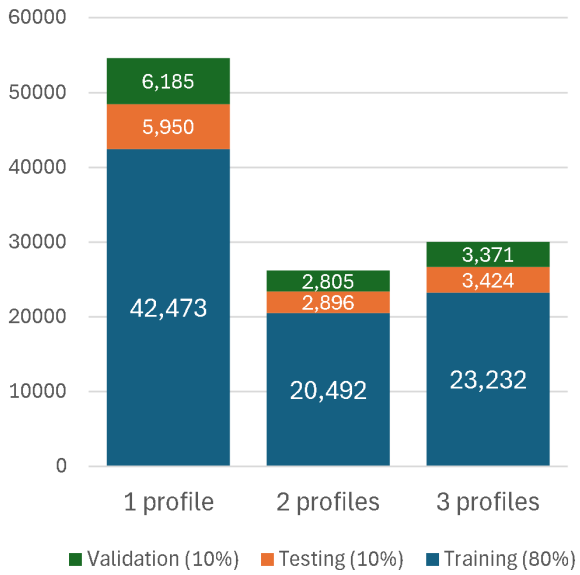


Figure 4: LAMP-CAP 中的剖面分布，显示具有 1、2 或 3 个剖面图的目标图数量。

```
- Target Figure Paragraph(s): [
    PARA-TARGET].
```

B 生成输出清理程序

我们分三步进行了数据输出清理。

1. 我们手动检查了 BLEU 或 ROUGE 得分为 0 的案例，以识别数据问题。我们发现 11 个案例（总计 12,270 个）中由于解析错误，原始标题被错误捕获——要么是缺少了真实的标题内容，要么是捕获了错误的文本。在一个实例中，解析器错误地从右下角捕获了作者的名字，而不是从同一页图表下方的标题。这些案例被排除在评估之外。
2. 我们使用 GPT-4o-mini 来清理生成的图片说明，去除无关的文本，如推理步骤、图形索引或任何不属于实际图片说明的前后缀。我们的清理说明明确规定不添加任何额外的文本或解释到原始输出中。使用了以下提示：
   下面是来自 arXiv 论文 1105.0392 的一个例子，展示了"Llama-4 Scout" 模型在"All Profile" 配置下清理前后的响应：
   以下是清理前的原始输出：
   以下是清理后的提取图片说明：
   在 GPT-4o 清理后，我们随机抽取了 100 个输出进行人工评估。使用二元标签（GOOD/BAD），我们评估提取的图片说明是否正确。所有 100 个抽取的提取结果都被标记为 GOOD，确认了清理过程的有效性。
3. 我们采用关键词筛选并辅以人工验证来过滤掉失败的生成结果，包括空白响应。这些错误案例的详细示例记录在 Table 3 中。清理后，我们在所有模型和配置中（共 12,259 个）识别出了 56 个独特的有问题案例，它们被排除在进一步分析之外。

在评估之前的文本规范化中，我们使用标准 Python 库实现了一个自定义预处理管道，该管道：(1) 将文本转换为小写，(2) 去除所有标点符号，以及 (3) 规范化空白字符。

在评估中，我们使用了在 Python 包中实现的标准 NLP 指标：`NLTK`（版本 3.9.1）用于 BLEU 评分（与 `SmoothingFunction` 结合用于平滑），以及 Google 的 `rouge_scorer`（版本 0.1.2）用于 ROUGE 指标。我们为两个包使用默认参数。具体的实现是直接从 `nltk.translate.bleu_score` 和 `rouge_score` 模块中导入的。

这部分附录是为了补充 Section 2 中有关使用四种模型不同配置进行字幕生成主要研究的结果。

Figure 6 显示了在不同语言模型和配置文件设置下的 BLEU-4 分布。

Figure 7 显示了不同语言模型和配置文件下的 ROUGE-2 分布。

本附录提供了关于 Section 2 中描述的情境对齐和情境不对齐的子集划分和评估的更多详细信息。

Figure 5 显示了 LAMP-CAP 测试集中目标字幕和个人简介字幕之间的 BERTScore 和 ROUGE-L 的分布。

Table 4 显示了在不同的大语言模型和配置文件配置下，上下文对齐子集的性能指标。

Table 5 展示了不同 LLM 和配置文件设置下的上下文不对齐子集的性能指标。

本附录部分是对 subsection 2.1 中关于消融研究发现的补充。Table 6 展示了消融研究的详细结果。

我们使用 Perplexity 来促进校对和文本润色。

www.xueshuxiangzi.com

Model and Config	Cases	Examples of Invalid Generations
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 3: 不同语言模型和配置文件中无效生成的示例

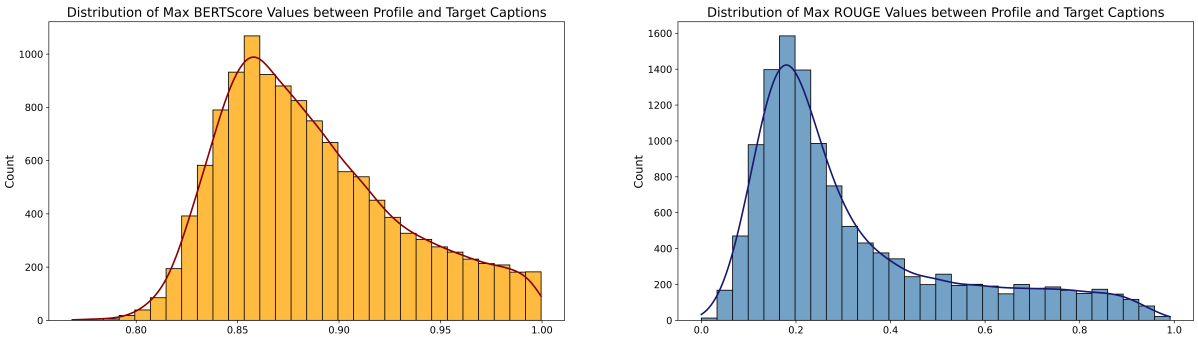


Figure 5: 在 LAMP-CAP 测试集中目标和配置文件说明之间的 BERTScore（左）和 ROUGE-L（右）指标的分布。这两个分数都呈现出向左偏的单峰分布。BERTScore 图显示，每个目标的提供的配置文件说明在语义上是非常相关的。另一方面，ROUGE-L 分数的更广的分布显示配置文件说明在词汇重叠方面较少。高语义相关性和词汇多样性激励我们将配置文件说明用作个性化的关键风格指标。

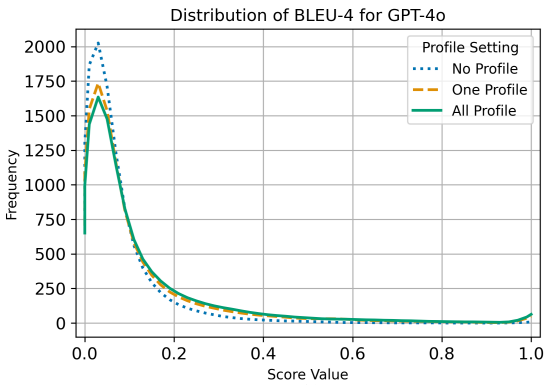
LLM	Profile Used	BLEU				ROUGE		
		B-1	B-2	B-3	B-4	R-1	R-2	R-L
GPT-4o	No	.226	.144	.101	.071	.321	.138	.259
	One	.437	.347	.289	.242	.533	.345	.484
	All	.478	.391	.332	.282	.573	.393	.530
Llama-4 Scout	No	.249	.178	.139	.112	.347	.183	.296
	One	.589	.523	.473	.427	.676	.536	.646
	All	.666	.605	.556	.510	.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 4: 在不同的 LLMs 和配置文件设置中，LAMP-CAP 上下文对齐子集（n=2,513）的表现。

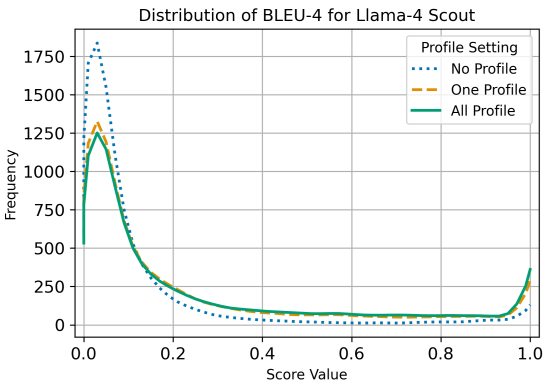
LLM	Profile Used	BLEU				ROUGE		
		B-1	B-2	B-3	B-4	R-1	R-2	R-L
GPT-4o	No	.217	.131	.088	.061	.320	.124	.245
	One	.238	.144	.097	.067	.345	.135	.269
	All	.244	.150	.102	.071	.351	.142	.275
Llama-4 Scout	No	.255	.178	.138	.112	.360	.181	.292
	One	.316	.233	.187	.155	.430	.238	.366
	All	.326	.243	.196	.164	.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 5: 各种大型语言模型和配置文件在 LAMP-CAP 上下文不匹配子集（n=9,690）上的表现。

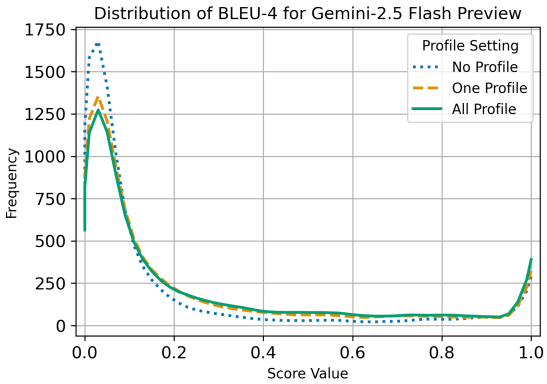
www.xueshuxiangzi.com



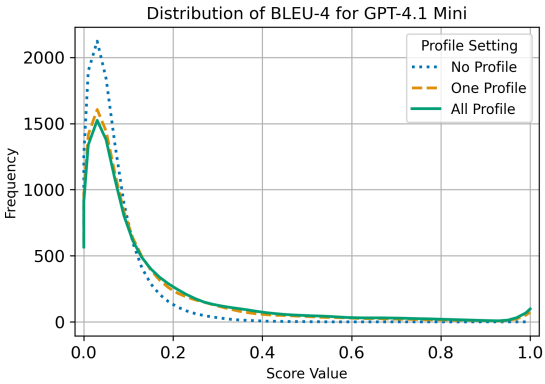
(a) GPT-4o 的 BLEU-4



(b) Llama-4 Scout 的 BLEU-4



(c) Gemini-2.5 Flash Preview 的 BLEU-4



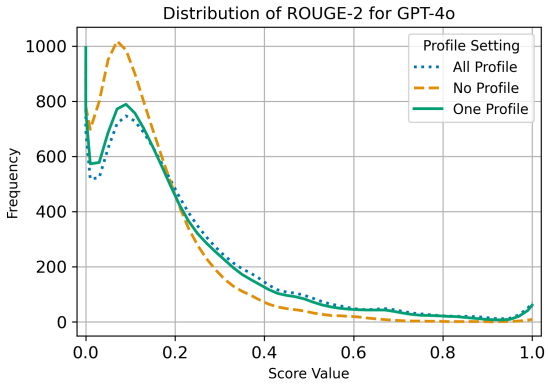
(d) GPT-4.1 Mini 的 BLEU-4

Figure 6: 不同大语言模型和配置文件中的 BLEU-4 分布。

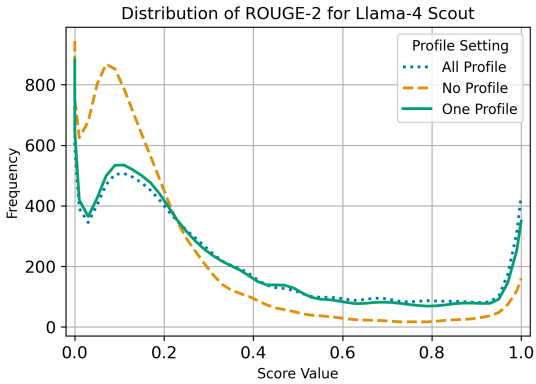
Elements	BLEU				ROUGE		
	B-1	B-2	B-3	B-4	R-1	R-2	R-L
No Paragraph	.299	.199	.146	.110	.393	.184	.314
No Image	.273	.171	.119	.086	.367	.154	.285
No Caption	.189	.109	.071	.048	.274	.100	.199

Table 6: 消融研究结果。

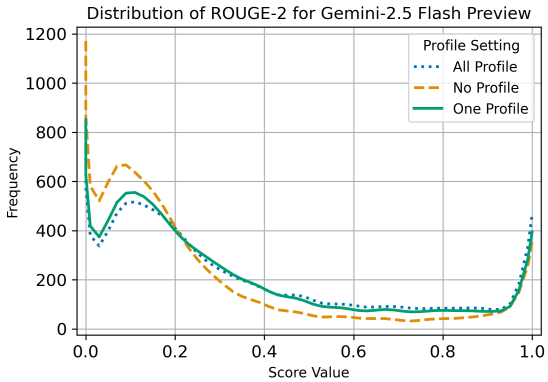
www.xueshuxiangzi.com



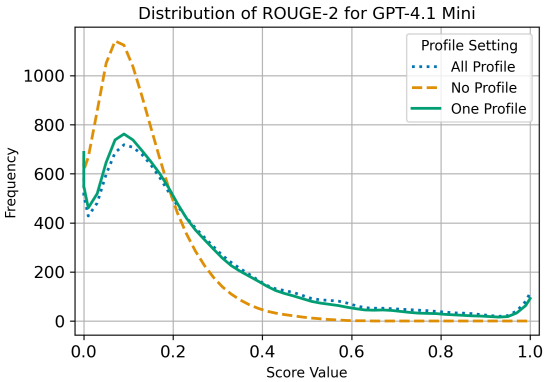
(a) GPT-4o 的 ROUGE-2



(b) Llama-4 Scout 的 ROUGE-2



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



(d) GPT-4.1 Mini 的 ROUGE-2

Figure 7: 在不同的大型语言模型和配置文件中 ROUGE-2 的分布。

www.xueshuxiangzi.com