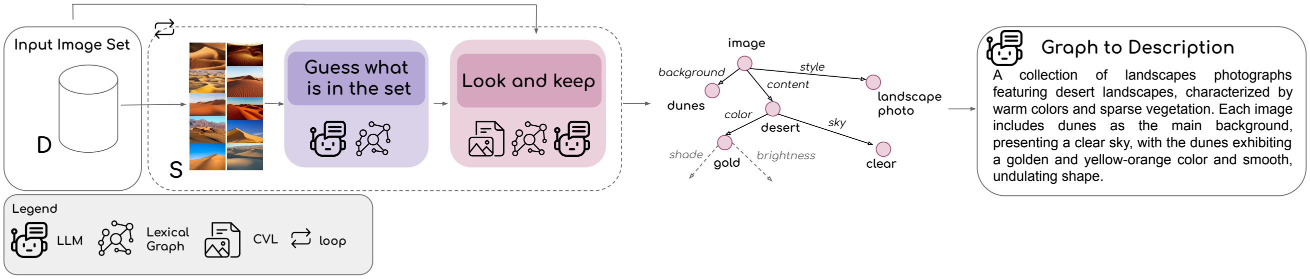
Towards Accessible Image Set Descriptions with `ImageSet2Text`

Piera Riccio
ELLIS Alicante
Alicante, Spain
piera@ellisalicante.org

Luis Domene García
Universitat Autonoma de Barcelona
Barcelona, Spain

Nuria Oliver
ELLIS Alicante
Alicante, Spain



A collection of landscapes photographs featuring desert landscapes, characterized by warm colors and sparse vegetation. Each image includes dunes as the main background, presenting a clear sky, with the dunes exhibiting a golden and yellow-orange color and smooth, undulating shape.

Figure 1. `ImageSet2Text` generates detailed and nuanced descriptions from large sets of images. We explore its capabilities in the context of accessibility.

Abstract

Recent advances in generative AI have enabled new tasks in Computer Vision, such as generating textual summaries of entire image collections. While single-image and group captioning have been widely explored for accessibility applications, in this work, we present `ImageSet2Text`, a system designed to produce high-level descriptions of image sets, and investigate its applicability for visually impaired users. After pilot interviews with members of the visually impaired community, we adapt `ImageSet2Text`'s pipeline by integrating the NCAM principles of accessibility and perform preliminary evaluations through an LLM-as-a-judge. Finally, we outline key future directions, including broader evaluation strategies.

1. Introduction

Single-image captioning and group captioning are well-established tasks in Computer Vision, with their applications for accessibility thoroughly explored in prior work [24]. Recently, however, researchers have begun to investigate the potential of generative AI and foundation models to produce summaries of large image collections, giving rise to novel tasks in the field [8, 10, 22]. The accessibility implications of these emerging tasks remain largely unexplored and represent a promising new research direction that we address in this extended abstract.

As the volume of visual content continues to grow across digital archives, the ability to interpret and navigate large-scale image collections becomes increasingly important. Image collections often contain rich patterns that are not apparent when examining single images in isolation [8], and making sense of these patterns is a predominantly visual task, leaving blind and low vision (BLV) individuals excluded from understanding the broader narratives embedded in large image datasets. In this extended abstract, we present the potential of `ImageSet2Text` [22] —a recently developed method that generates textual descriptions of entire image collections— for accessibility. We argue that, by transforming visual patterns into concise textual descriptions, `ImageSet2Text` can offer non-visual access to image sets of hundreds or thousands of samples, empowering BLV users, and offering insights that would otherwise be inaccessible.

2. Related Work

Accessibility is a key application of image captioning [11, 23, 24], enabling BLV individuals to access and understand visual content. The Computer Vision community has made significant progress in generating accurate captions for single images [12, 31, 33]. In the context of accessibility, Safiya and Pandal emphasize the importance of real-time captioning systems and propose methods to address the technical and computational challenges these systems face [23]. In parallel, Wibowo et al. explore the role of image captioning in enhancing the social media experience for blind users on platforms like X (formerly Twitter), illustrating how such technologies can foster more inclusive digital communication [32].

Evaluation is also a crucial aspect of the research efforts in this field. For example, Nganji et al. proposed the Image Description Assessment Tool (IDAT) to evaluate the quality of image descriptions and support accessibility for BLV [18]. Anwar et al. proposes benchmark datasets to evaluate such solutions in real-world scenarios, including pedestrian crossings, currency recognition, bus stops, and stairs detection [3]. Other approaches introduce different machine learning based tools to automatically assess the accessibility of image descriptions in terms of their compliance with the NCAM guidelines [9, 26, 27].[1] These guidelines, developed by the National Center for Accessible Media (NCAM), emphasize concise and essential descriptions, clear and consistent language, and a structured format that allows users to access more detail progressively. Qualitative evaluations are also valued in the literature. Early studies highlighted how preferences of BLV users vary across digital contexts [28, 29], while others emphasized the need for evaluation metrics aligned with the users' needs, particularly regarding context [14].

While the literature has explored image captioning beyond single instances, such as multi-image descriptions and temporally coherent narratives [2, 6, 15], the task of summarizing large image collections has only recently gained attention [10, 22]. As a result, its potential for accessibility remains largely unexplored. This extended abstract investigates that open question by focusing on the recently introduced `ImageSet2Text` method, described next.

3. `ImageSet2Text`

To generate textual descriptions of image collections, the `ImageSet2Text` pipeline (depicted in Fig. 1) generates descriptions through an iterative two-phase process: (a) *Guess what is in the set*, and (b) *Look and keep*. To ensure scalability, each iteration operates on a randomly sampled subset of images. In *Guess what is in the set*, an LLM-based visual question answering (VQA) module identifies prominent visual elements within the subset. These elements are then used, in conjunction with an external lexical graph, to formulate hypotheses about the content of the entire image set. In *Look and keep*, these hypotheses are validated using contrastive vision-language (CVL) embeddings to assess their consistency across the full collection. Verified hypotheses are added to a *concept graph*, which informs the next iteration. The process continues iteratively, expanding the concept graph using information obtained via the VQA. Once no further expansion is possible, a final description is generated through an LLM call, leveraging the accumulated insights within the concept graph. This graph is initialized with three core nodes, *i.e. content*, *background*, and *style*, which represent the primary dimensions of analysis based on verified visual information from the image collection. The authors of `ImageSet2Text` have used GPT-4o-mini [1] as the language model, CLIP [21] as the CVL model, and WordNet [17] as the external lexical resource in their implementation.

In this extended abstract, we explore the potential of `ImageSet2Text` to enable non-visual interaction with large collections of visual content, making it accessible to BLV individuals. We analyze possible use-cases through preliminary interviews with BLV participants, highlight current limitations, and outline future design considerations needed to ensure its effectiveness for this community.

4. `ImageSet2Text` for Accessibility

The automatic generation of descriptions of image sets is a novel challenge that is recently addressed by the Computer Vision community. Hence, its potential use for accessibility in the case of visually impaired individuals remains unexplored. In this section, we outline our ongoing efforts and key considerations to take into account when tackling this emerging task in the context of accessibility. Specifically, we present preliminary findings from interviews with members of the visually impaired community, followed by a set of straightforward technical adaptations applied to `ImageSet2Text`. These adaptations aim to enhance the accessibility of the generated descriptions, whose effectiveness is assessed by means of a large language model (LLM) as an evaluative tool.

4.1. Pilot Interviews

We conducted interviews with three collaborators from Fundación ONCE[2], a Spanish national foundation for universal accessibility, to gather community insights on the usefulness and possible improvements of our approach [7]. Carried out in Spanish and later translated to English, the interviews followed a structured format with the following parts: (1) an assessment of the collaborator's familiarity with automatic alt-text generative tools; (2) an exploration of the potential value of textual descriptions for image collections in everyday tasks, where the collaborators were asked to evaluate a sample description generated by `ImageSet2Text`; and (3) open-ended questions for additional feedback.

[1] Image Description Guidelines, http://diagramcenter.org/table-of-contents-2.html, last access: 6th of August 2025.

[2] Fundación ONCE, https://www.fundaciononce.es/es, Last Access: 22.08.2025

Our collaborators responded positively to the concept of set-level image descriptions, noting their usefulness in contexts where understanding the overall scene is more important than focusing on individual details, such as during **events**, **travel**, **entertainment**, or **organizing digital photo folders**. However, they stressed that these summaries should complement, not replace, individual image descriptions, as each serves a different purpose. When reviewing a sample output from `ImageSet2Text`, participants were generally satisfied with its coherence, clarity, and level of detail. They particularly valued the explicit description of relationships among entities across images, an aspect they often find missing in commercial alt-text tools. This strength stems from `ImageSet2Text`'s use of graph representations [19, 20, 22].

If further developed for accessibility, our collaborators emphasized the need for simple, direct language to reduce ambiguity and suggested adapting descriptions based on users' visual experience: for example, including references to color and light for those with prior sight, and offering alternative cues for those without. Overall, their feedback highlights the potential of `ImageSet2Text` to enhance the accessibility and inclusion for BLV users in both personal and professional settings.

4.2. Integration of NCAM principles

Based on the feedback gathered during our preliminary interviews, we made an adjustment to the `ImageSet2Text` pipeline by incorporating the NCAM principles of accessible media into the prompt used to generate the final description of the image set from the concept graph constructed by `ImageSet2Text`. Since these principles are originally designed for describing individual images, they address a different use case than that of `ImageSet2Text`, which aims to describe entire collections of images. However, by identifying and generalizing the most relevant elements of these guidelines, we adapted them to our task and modified accordingly the prompt used to generate the descriptions, namely *Graph to Description* (in Fig. 1). Below, we present the portion of the prompt that explicitly incorporates these principles:

```
[...]
**ACCESSIBILITY GUIDELINES**
- **Avoid visual-only cues**: Do not describe
     information in a way that requires sight to
     understand.
- **Accessible Rephrasing**:
    - You **may change** visual-centric terms to words
        that describe meaning, function, or non-visual
        properties, **only if this does not add or
        remove factual information**.
    - If no non-visual alternative exists, keep the
        term but clarify it in accessible language.
    - **Do not invent** attributes not in the GRAPH.
**Examples:**
[...]
```

4.3. Preliminary Evaluation

Through manual inspection, we qualitatively observe the positive impact of incorporating NCAM guidelines into `ImageSet2Text`'s pipeline. To complement this, we introduce a quantitative evaluation, described next.

Datasets. Using the updated prompt, we generate descriptions for all image sets in the GroupConceptualCaptions and GroupWikiArt datasets, both originally introduced and released by the authors of `ImageSet2Text` [22]. These datasets are derived from the ConceptualCaptions [25] and WikiArt [30] image collections, respectively.

Evaluation Methodology. As previously mentioned, various approaches have been proposed in the literature to assess a text's adherence to the NCAM principles [4, 26, 27]. In our work, we leverage recent advances in large language models (LLMs) for text evaluation, we adopt Prometheus [13], an LLM specifically designed to evaluate outputs from other LLMs. We instruct Prometheus to rate each description across multiple accessibility-related dimensions on a scale from 1 to 5. These evaluation dimensions are detailed in Table 1.

Table 1. Dimensions analyzed through the LLM-as-a-judge evaluation

Evaluation Questions
(1) Is the text adapted for blind or visually impaired users, avoiding reliance on visual-only cues?
(2) Is the text concise and free of unnecessary repetition or filler?
(3) Does the text maintain objectivity and avoid personal opinions?
(4) Does the text follow a clear structure from general context to specific details?
(5) Is the tone and language accessible for blind or visually impaired users and screen-reader friendly?

Results. In Fig. 2, we present the evaluation results comparing the original (vanilla) version of `ImageSet2Text` with the adapted version that incorporates accessibility guidelines into the prompt. Overall, the results indicate that the vanilla system is already reasonably well-aligned with accessibility standards, particularly in terms of objectivity and structural clarity. It also performs well on conciseness and tone, but receives more average scores (around 3 out of 5) on the "audience" dimension, which assesses suitability for BLV users. The adapted version shows slightly higher scores across all evaluated dimensions, suggesting that the prompt modifications have had a positive impact. However,

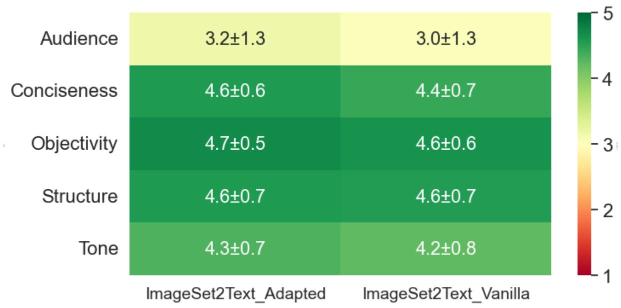


Figure 2. Comparison between two versions of `ImageSet2Text`: the original **vanilla** model as described in [22], and the **adapted** model incorporating NCAM principles into the final prompt. The plot displays the mean scores and standard deviations across four evaluation runs for five different accessibility-related dimensions (summarized in Tab. 1).

the overall improvements are modest and may be negligible in practical settings.

To strengthen this evaluation, we are also exploring alternative LLM-as-a-judge paradigms to investigate whether different models produce more varied scores across evaluation dimensions. With Prometheus [13], the results tend to be relatively uniform across most dimensions, with the only exception of the "audience" dimension. These scores raise the question of whether they reflect a limitation in Prometheus's sensitivity to fine-grained distinctions in the context of our study, or whether the descriptions genuinely exhibit similar levels of quality across those dimensions. Comparing evaluations from different models will help clarify this and further validate the observed improvements in the adapted version.

5. Discussion and Future Work

In ongoing work, we are conducting semi-structured interviews with participants recruited through our collaborators with the ONCE Foundation. For these sessions, we consider three image sets from the GroupConceptualCaptions dataset and three image sets from the GroupWikiArt dataset, chosen to contain different levels of visual homogeneity. This choice reflects a shared understanding with our collaborators that real-world image collections are often heterogeneous. The current version of `ImageSet2Text` is designed to identify and summarize common visual elements within an image set; as a result, it generates less detailed descriptions when fewer elements are shared among the images, *i.e.*, as the heterogeneity of the image sets increases. Therefore, visual homogeneity is a key factor in evaluating the system's performance and its perceived usefulness. The aim of these interviews is to evaluate whether the descriptions generated by means of the accessible adaptation of `ImageSet2Text` are perceived by community members as more useful than the descriptions produced by vanilla `ImageSet2Text`. To provide a more comprehensive evaluation of the descriptions, we also include large-scale group captions generated by two baseline models: Qwen-2.5-VL [5] for image sets from GroupConceptualCaptions, and BLIP-2 [16] for those from GroupWikiArt. This selection is based on experimental results reported by the authors of `ImageSet2Text` [22], from which we chose the best-performing baseline for each dataset.

Our preliminary evaluation through the LLM-as-a-judge reveals a key limitation of the current approach: the descriptions of the image sets are concise by design. If the system is intended to support BLV individuals, it may be insufficient to apply accessibility principles only at the final stage of the pipeline, as it is difficult to significantly alter the existing descriptions to make them more accessible. Therefore, a promising direction for future work would consist of integrating the accessibility principles earlier in the process, influencing both the VQA module and the construction of the concept graph in `ImageSet2Text`. As a result, the pipeline would be tailored for accessibility *by design*, potentially generating a concept graph that captures more relevant and meaningful details for the target audience.

Such a shift in design, however, would require a different evaluation strategy given that the generated descriptions in the accessible version of `ImageSet2Text` would potentially contain very different content when compared to the descriptions created by the vanilla version. Thus, feedback from BLV users alone may be insufficient, as they are unable to assess the relevance of unseen visual information. In this scenario, input from sighted experts with deep experience in accessibility would be essential to ensure that the system captures and communicates the most important visual details effectively.

6. Conclusion

We introduce the potential of `ImageSet2Text`, a method designed to generate textual summaries of image collections. Through a pilot study conducted with members of the ONCE Foundation, we gathered positive feedback on the usefulness of set-level descriptions, along with suggestions for improvement. In response, we integrated NCAM principles into the final prompt of the pipeline, aiming to enhance the clarity, structure, and audience appropriateness of the generated descriptions for visually impaired users. A preliminary assessments through LLM-as-a-judge paradigm show modest but consistent improvements across key accessibility dimensions. However, these gains remain limited, suggesting that prompt-level adaptation alone may not be sufficient. As future work, we plan to conduct more extensive quantitative and qualitative evaluations (via semi-structured interviews) of the adapted version, with the goal of identifying further improvements across the full pipeline.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[3] Qazi Mohd Iqbal Hussain Anwar, Ch VS Satyamurty, and Rakesh Kumar Godi. Enhancing accessibility: Image captioning for visually impaired individuals in the realm of ece advancements. In *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 317–321. IEEE, 2024. 2

[4] Jinat Ara and Cecilia Sik-Lanyi. Automated evaluation of accessibility issues of webpage content: tool and evaluation. *Scientific Reports*, 15(1):9516, 2025. 3

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv*, 2502.13923, 2025. 4

[6] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[7] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020. 2

[8] Boyang Deng, Songyou Peng, Kyle Genova, Gordon Wetzstein, Noah Snavely, Leonidas Guibas, and Thomas Funkhouser. Visual chronicles: Using multimodal llms to analyze massive collections of images. *arXiv preprint arXiv:2504.08727*, 2025. 1

[9] Himmat Dogra. A framework for an automatic evaluation of image description based on an image accessibility guideline. Master's thesis, OsloMet-storbyuniversitetet, 2020. 2

[10] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024. 1, 2

[11] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer, 2020. 1

[12] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1

[13] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024. 3, 4

[14] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*, 2022. 2

[15] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 2

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[17] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[18] Julius T Nganji, Mike Brayshaw, and Brian Tompsett. Describing and assessing image descriptions for visually impaired web users with idat. In *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011*, pages 27–37. Springer, 2012. 2

[19] Itthisak Phueaksri, Marc A Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. Towards captioning an image collection from a combined scene graph representation approach. In *International Conference on Multimedia Modeling*, pages 178–190. Springer, 2023. 3

[20] Itthisak Phueaksri, Marc A Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. Image-collection summarization using scene-graph generation with external knowledge. *IEEE Access*, 2024. 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[22] Piera Riccio, Francesco Galati, Kajetan Schweighofer, Noa Garcia, and Nuria Oliver. Imageset2text: Describing sets of images through text. *arXiv preprint arXiv:2503.19361*, 2025. 1, 2, 3, 4

[23] KM Safiya and R Pandian. A real-time image captioning framework using computer vision to help the visually impaired. *Multimedia Tools and Applications*, 83(20):59413–59438, 2024. 1, 2

[24] Henry Senior, Gregory Slabaugh, Shanxin Yuan, and Luca Rossi. Graph neural networks in vision-language image understanding: a survey. *The Visual Computer*, 41(1):491–516, 2025. 1

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 3

[26] Raju Shrestha. A neural network model and framework for an automatic evaluation of image descriptions based on ncam image accessibility guidelines. In *Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference*, pages 68–73, 2021. 2, 3

[27] Raju Shrestha. A transformer-based deep learning model for evaluation of accessibility of image descriptions. In *Proceedings of the 2022 14th International Conference on Machine Learning and Computing*, pages 28–33, 2022. 2, 3

[28] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. " person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020. 2

[29] Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*, pages 1–15, 2021. 2

[30] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. 3

[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1

[32] Jessica Lynn Wibowo, Gabriel Seemore Gunawan, and Ivan Sebastian Edbert. Enhancing social media accessibility: Automatic alternative text generation in x by image captioning. In *2024 14th International Conference on System Engineering and Technology (ICSET)*, pages 123–128. IEEE, 2024. 2

[33] Kelvin Xu. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 1