**Image-to-Text**

Jean P. Melendez Villanueva

University of Maryland

Data 665 AI Applications

Professor Jeremy Bolton

**Image-to-Text**

**Introduction**

Image-to-text technologies are rapidly transforming how artificial intelligence interacts with the visual world. These systems, known as vision-language models, are capable of analyzing images and translating that analysis into natural language. This functionality has wide-ranging applications, including content creation, image tagging, and most notably, digital accessibility. By converting visual information into readable text, image-to-text models serve as essential tools for ensuring that digital environments are inclusive, particularly for individuals with visual impairments who rely on screen readers.

A key use of this technology is the generation of alternative text, commonly referred to as alt text. Alt text provides descriptive summaries of images, allowing visually impaired users to access the same visual content as sighted individuals. When done properly, alt text supports legal compliance with web accessibility standards like WCAG and ADA, while also promoting digital equity. However, not all AI-generated alt text is effective. Producing high-quality, context-aware descriptions often requires precise instruction, making prompt engineering, the strategic wording of input commands—an essential skill in the process.

This exercise explores the effectiveness of two vision-language models, OpenAI’s ChatGPT-4o and Google Gemini, in generating alternative text for two different image types: one scenic and one technical. The goal was to evaluate the accuracy, detail, and usefulness of the descriptions produced, while also observing how prompt modifications could enhance the results. This experiment not only highlights the capabilities and limitations of these tools but also reinforces their importance in making digital communication more accessible and inclusive in both academic and professional settings.

**Data Section**

For this experiment, four distinct images were selected to evaluate how vision-language models interpret both expressive and technical visual content. The first scenic image, referred to as Scenic Image 2, depicts a winding mountain road bordered by tall trees beneath a vivid sunset sky. This image was chosen to assess the model’s ability to capture not just objects, but also the mood, depth, and emotional tone of a natural landscape. The second scenic image, Scenic Image 1, presents a peaceful garden setting with a pond, a small bridge, and vibrant flowers, providing a different kind of scenic challenge—one focused more on compositional variety and seasonal cues. Both scenic images tested the models’ capacity for storytelling, sensory detail, and visual nuance.

The experiment also included two technical visualizations. The first, a box plot, displays age distributions segmented by gender and passenger class. As a statistical chart, it required the models to identify and describe features such as medians, interquartile ranges, and outliers across grouped data. The second technical image, a bar graph titled “Age by Transported Status”, represents a frequency distribution of ages with overlapping colored bars to show comparison between transported and non-transported groups. This image tested the models’ ability to recognize axis labels, color coding, overlapping values, and data trends—requiring more than just structural recognition and demanding analytical interpretation.

These four images were intentionally selected to provide a balanced test across aesthetic and analytical domains. Scenic Image 2 and Scenic Image 1 required descriptive, emotive alt text to support accessibility for artistic and nature-focused visuals. In contrast, the box plot and bar graph called for clarity, precision, and a focus on data interpretation. This multi-image approach offered a comprehensive assessment of how vision-language models adapt to different visual contexts. It also reinforced the importance of customizing alt text based on the purpose and audience of the image, ensuring that accessibility solutions are meaningful and informative across disciplines.

**Experiment**

To evaluate the effectiveness of image-to-text technologies, two vision-language models were selected: ChatGPT-4o by OpenAI and Google Gemini. Both models were prompted with the same two images, Scenic Image 2 and the box plot visualization, using tailored prompts to elicit detailed alternative text suitable for visually impaired users. The prompts were designed to guide the models in generating context-rich descriptions, with a focus on clarity, completeness, and relevance to the image type. Each model's initial responses were reviewed, and then refined prompts were applied to assess how prompt engineering impacted the quality of the generated text.

For Scenic Image 2, both models were asked to describe the scene in alt text format. ChatGPT-4o responded with an emotionally rich narrative, highlighting the curvature of the road, the density of the surrounding forest, and the dramatic coloration of the sky. It evoked not only what was present in the image but also the mood it conveyed. Google Gemini, while similarly descriptive, focused more on listing the visual elements such as flowers, trees, and the position of the bridge and water. Its interpretation was more literal, with less emphasis on atmosphere and emotional tone. This contrast revealed that while both models are capable of generating vivid alt text, their narrative depth and interpretation styles differ significantly.

For the box plot, both models initially produced limited descriptions, often identifying the chart type without interpreting the statistical components. Upon prompt revision, specifically asking for technical details such as medians, quartiles, and group comparisons, ChatGPT-4o provided a more analytical summary. It identified the age ranges, explained the differences between genders and classes, and acknowledged the presence of outliers. Google Gemini also improved its output, noting axis labels, color coding, and observed trends in the data, such as the distribution of transported versus non-transported individuals. This demonstrated that prompt engineering is essential when engaging AI for technical image interpretation, as more specific instructions lead to more insightful and relevant responses.

**Concluding Remarks**

This experiment highlighted the capabilities and limitations of two prominent vision-language models in generating alternative text for both scenic and technical images. ChatGPT-4o demonstrated strong performance in crafting expressive, context-aware descriptions, especially when guided by refined prompts. Its responses for the scenic image conveyed not just the visual elements but also the emotional atmosphere of the scene. On the technical side, it was able to accurately interpret and explain statistical components of the box plot once the prompt was clarified. In contrast, Google Gemini provided more structured and factual outputs, excelling at identifying visual components and labeling chart features. While its narrative depth was more limited, Gemini proved reliable in producing clear and accessible content.

Ultimately, this exercise emphasized the importance of prompt engineering in working with multimodal AI systems. The clarity, specificity, and context of a prompt directly influenced the relevance and completeness of the models’ outputs. These findings have practical implications for accessibility-focused design, especially in fields like education, data analytics, and web development. As AI continues to evolve, understanding how to direct these systems effectively will be an essential skill for ensuring content remains inclusive and meaningful across diverse audiences. This experiment not only demonstrated the current potential of image-to-text technologies but also reinforced the human role in shaping AI-driven communication.

**References**

Bureau of Internet Accessibility. (2023, October 26). *Be careful when using A.I. for alternative text*. <https://www.boia.org/blog/be-careful-when-using-ai-for-alternative-text>

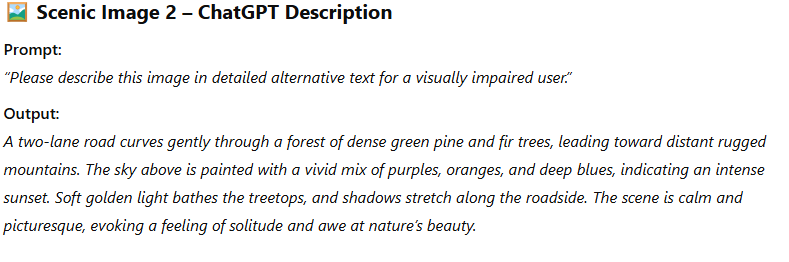
OpenAI. (n.d.). *ChatGPT*. <https://chatgpt.com/>

Google. (n.d.). *Gemini*. <https://gemini.google.com/>

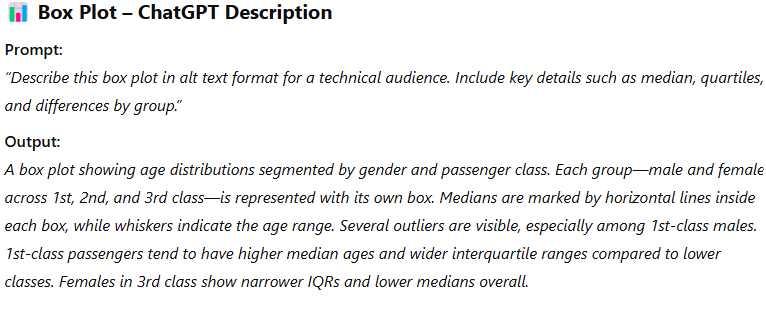
Perkins School for the Blind. (2024, August 15). *How to write alt text and image descriptions for the visually impaired*. <https://www.perkins.org/resource/how-write-alt-text-and-image-descriptions-visually-impaired/>

**Appendix**

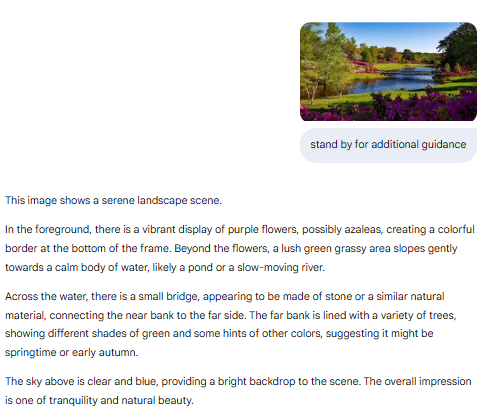
**Appendix A –** Screenshots



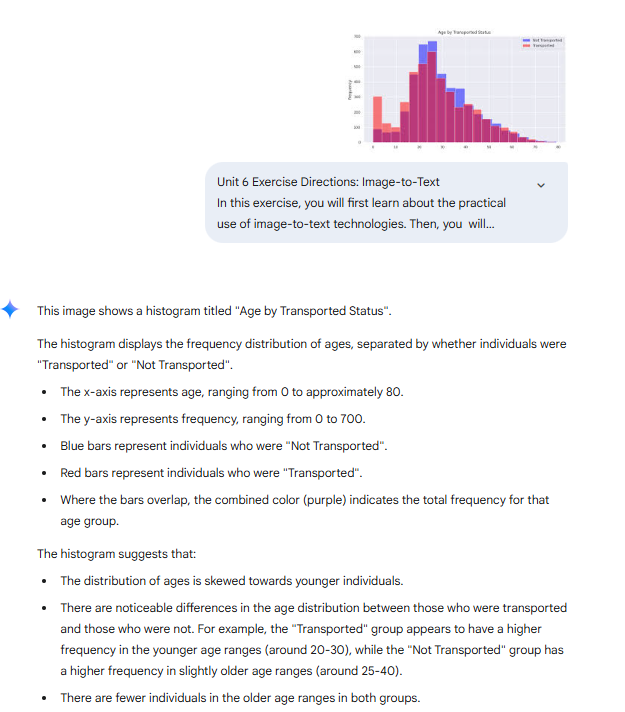
(Figure 1) Scenic Image 2 – ChatGPT-4o Output



(Figure 2) Box Plot – ChatGPT-4o Output



(Figure 3) Scenic Image 1 – Google Gemini Output



(Figure 4) Bar Graph – Google Gemini Output