International Cartographic Association
Association Cartographique Internationale

What do you see? An XAI approach for VLM-generated map descriptions

Güren Tan Dinga [a]*, Jochen Schiewe [a]

[a] *HafenCity University Hamburg, Lab for Geoinformatics and Geovisualization (g2lab) - gueren.dinga@hcu-hamburg.de, jochen.schiewe@hcu-hamburg.de*

* Corresponding Author

Abstract: Over the last decades, significant progress has been made in enabling diverse communities to create and share cartographic maps. However, advancements in map accessibility, for blind and visually impaired users in particular, still lag behind. A critical challenge remains in generating effective and efficient text descriptions that are supported by screen-readers. Vision Language Models (VLMs) offer a promising solution, as they can produce image descriptions quickly. However, their outputs depend heavily on network architecture and prompt engineering. Further, VLMs usually are complex and outputs are difficult to interpret. To address the interpretation of outputs in particular, we propose an Explainable AI (XAI) approach using Shapley Explanations to analyze and understand the contributions of specific map regions to the text outputs generated by a VLM. Our contribution lies in applying XAI techniques to spatial data, providing a workflow to evaluate and improve the interpretability of VLM-generated map descriptions. Data and further information can be found on a corresponding GitHub repository: https://github.com/grndng/CartoXAI

Keywords: CartoAI, Explainable AI (XAI), Shapley Values

1. Introduction

Robinson and Griffin (2024) state that over the past few decades, significant advances have been made in enabling a wide variety of communities and individuals to create and share cartographic maps. It is also noted that the same progress does not apply for ensuring map accessibility, and particularly highlight the challenge of ensuring useful text descriptions to support screen-readers for blind and visually impaired users. This statement can be substantiated by a brief investigation, as many English and German websites do not provide alternative texts for images and maps. A brief correspondence with editorial teams in Germany has reinforced this impression, due to the generation of good and descriptive alternative texts being not only time-consuming but also subjective and dependent on context. An approach to generate map and image descriptions for accessibility purposes could incorporate Vision Language Models (VLMs) which, in essence, are multimodal Large Language Models (LLMs), that can then be interpreted by screen-readers (Xu and Tao (2024), Zhang and Ochiai (2024), Zhao et al. (2024)). While VLMs can generate image descriptions quickly, the quality is heavily dependent on carefully engineered prompts. As highlighted by Xu and Tao (2024), predicting the behaviour of such large models is challenging due to their complexity. The complexity leads to difficulties in understanding and interpreting the output of such networks. Developing a better understanding of VLM output has the potential to lead to improvements in AI-based description generation for e.g. maps. Therefore, we approach the challenge from an XAI perspective: using Shapley Explanations to better understand AI-based description generation for maps. As a case study, we aim to quantify the contribution of map areas to the generation of specific textual output in cartographic maps.

Explainable AI (XAI) can help to interpret why machine learning models make certain estimations by introducing new metrics such as contributions of features. In our research, we use Shapley Additive Explanations (Lundberg and Lee (2017)) to make the output of a relatively modest open-source VLM[1] interpretable. Shapley Values, originating from game theory, measure individual contributions to a collaborative games outcome (Shapley (1952)). In this study, we compute the individual contribution of each map region to the VLM's textual outputs, which can be compared to a collaborative games outcome. By focusing on *why* a VLM generates specific outputs rather than just *what* it generates, we aim to provide insights into the decision-making of networks to deliver another set of metrics to improve the quality of map descriptions.

2. Related Work

Image captioning refers to the process of generating descriptions for images. One of the key challenges in this task is identifying the purpose of the image in order to create a precise description to pull together the overarching elements. While the use of machine learning, particularly deep learning, in image captioning is well-researched (Ghandi et al. (2024), Hossain et al. (2019)), the application of image captioning networks to spatial data, es-

[1] https://moondream.ai/

pecially maps, requires further work. Ongoing research shows that image captioning can address the problem of e.g. cultural heritage inaccessibility: Díaz-Rodríguez and Pisoni (2020) aim to enhance the accessibility of historical images by making them understandable only through words. While this is comparable to the generation of captions for cartographic maps, the term "explainability" is used in a different context: for Díaz-Rodríguez and Pisoni (2020), "explainability" means that a multimodal network generates a description for an image, thereby "explaining" it. Contrary to that, in our research, we aim to utilize XAI to make the decision-making process of such networks less opaque and more understandable. Han et al. (2020) present an XAI approach to image captioning that links detected objects in an image to a particular word or phrase in the generated sentence, while Dewi et al. (2023) generate image captions and employ cosine similarity (Vasiliev (2020)) and term frequency inverse document frequency (Salton and Buckley (1988)) to compare the output of a propriety and open source image captioning network. While they look into Shapley Values for classification purposes, it is not employed for explainability purposes. After careful and thorough literature research to the best of our ability, it became clear that the topic of XAI is not yet addressed in terms of description generation for cartographic maps.

3. Method and Material

Figure 1 shows a simplified workflow to generate Shapley Values: we generate image captions, tokenize them and compute alignment scores to, in the end, determine which image region makes an impact on the generation of a specific word in the generated map description.

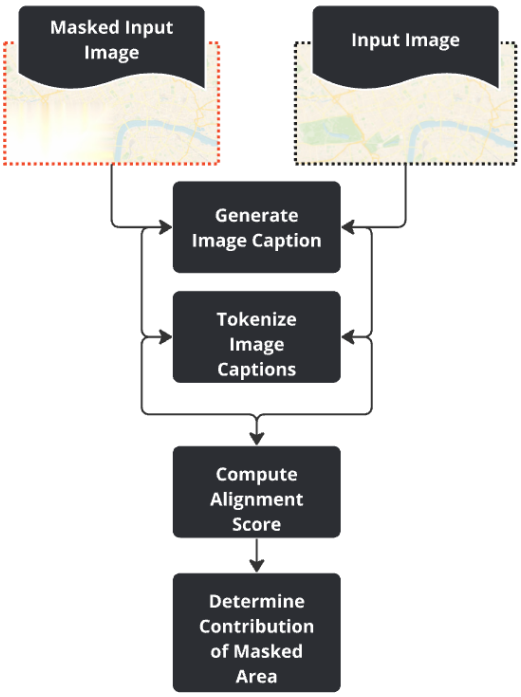


Figure 1. Simplified workflow to generate Shapley Values (Contribution Scores) for specific image regions

The workflow begins with generating a caption for the input image using the VLM "Moondream2". The same process is applied to a masked version of the original image, where certain regions have been obfuscated by masking. Figure 2 shows an exemplary VLM output for a map of London.

Figure 2. Generation of a map description by a VLM, using the prompt "What do you see?" and a map of London

After generating image captions for the original and masked image, we employ a transformer-based language model and tokenizer, "distilbart"[2], to tokenize the image caption. Tokenization is the process of converting the text generated by the VLM into smaller units (tokens) that can be processed by a model to make computations. The map description generated by the VLM of the original image might be "A detailed map of London, showing streets and waterways". This description is being *tokenized* using distilbart into the tokens ["A", "detailed", ...]. Since we eventually have to work with numeric values, we will map each token to a unique numerical ID based on the tokenizers' vocabulary, e.g. [100] for "A", [250] for "detailed". The following process includes the vectorization of each token ID so we receive a suitable representation of each token for processing by the model. The tokenization is crucial for analysing which parts of the input image regions contribute to the generation of specific tokens or words in the VLM output. Figure 3 summarizes the process of creating tokens from a sentence.

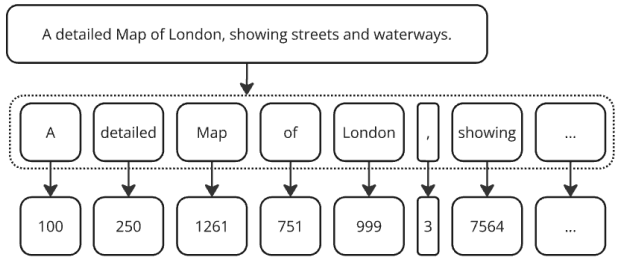
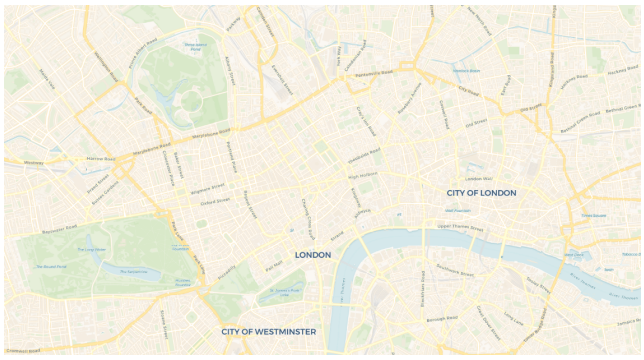


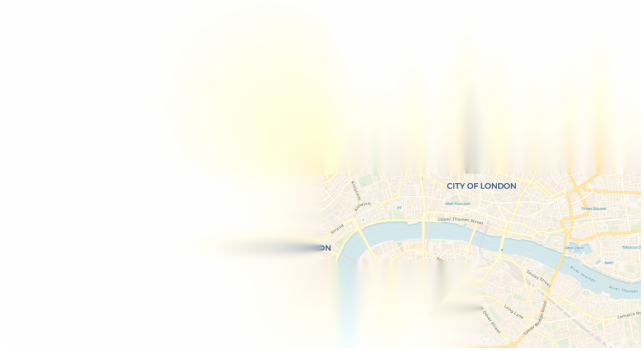
Figure 3. Tokenization for the VLM output "A detailed Map of London, showing streets and waterways."

[2]https://huggingface.co/sshleifer/distilbart-cnn-12-6

To compute Shapley Values, the image is altered systematically by masking regions to observe their impact on the generated map descriptions. The default masking strategy is to divide the image into grid-shaped regions. Each region is masked, and map descriptions are generated for each masked map image. Figure 4 shows the original, unaltered image of one of our study areas and a masked version of the same area. Further masked images and their corresponding VLM outputs can be found in the appendix. Note that the granularity of the final computation of the contributions can be controlled by increasing or decreasing the size of the masked areas.



(a) London, original image containing map labels



(b) One of many masked versions of the original image 4a

Figure 4. Overview of an exemplary original image and one of many masked versions used for the creation of Shapley values

While the original image caption could be "A detailed map of London, showing streets and waterways", an exemplary caption of a masked image could be "A detailed map of a city, showing streets and waterways.". The goal of masking is to obfuscate specific regions to generate new captions to be able to quantify the importance of those regions to specific tokens (such as "London" in this example). For each masked image, the VLM generates a caption. The caption then is tokenized to identify which tokens are affected by masking. Following through with our example, we now have two captions: for the original image the VLM generates "A detailed map of London, showing streets and waterways." and for the masked version of the original image the VLM generates "A map of a city, showing streets and waterways.".

To now compute Shapley Values, we need to quantify how much the masking process has affected the likelihood of specific tokens occurring in the caption. To do so, a process called "Teacher Forcing Logits" is used: both captions are input to the process. The logits (the raw model output) for each token are compared between the original and masked version. To do so, the model is provided with the ground truth tokens (the ones generated from our original image). By measuring the alignment between the ground truth token and the next token that is supposed to be generated, we can determine the impact of masking and thereby the amount of contribution of the masked area to a specific token. Shapley Values quantify the contribution of each image region to the likelihood of specific tokens in the caption with marginal contributions: For each region, the change in probability for each token when that specific region is masked is computed. This process is repeated by masking all possible combinations of regions to ensure that the contribution of each region is evaluated in every possible context. This makes the computation of Shapley Values computationally expensive.

In a final step, the computed Shapley Values are visualized to show the contribution of specific regions to each token: high positive contribution is depicted in red, while negative contribution is depicted in blue. The amount of contribution is controlled by opacity, where a high opacity means that there is a low contribution.

The map segment shown in Figure 4a represents central London, one of the two areas included in our case study. The second area, illustrated in Figure 5, corresponds roughly to the downtown region of Vancouver. For the purpose of this research, we conducted all experiments for map scales of 1:50,000 and 1:25,000 as well as including and excluding map labels.

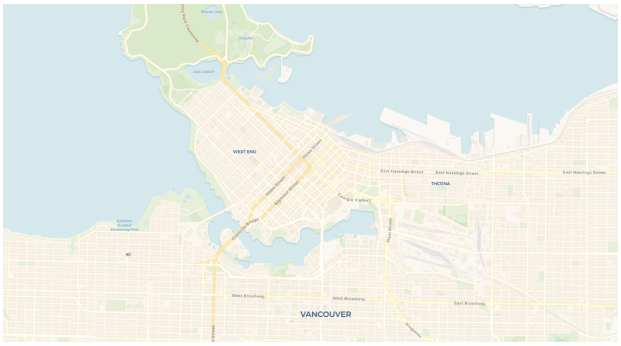


Figure 5. Overview of the area used to generate VLM outputs for Vancouver

4. Results

Preliminary results were obtained for the aforementioned maps of London and Vancouver, both using the "Voyager" styling from CARTO Basemaps[3]. What follows are the map descriptions generated by our VLM for the scale of *1:50,000*:

- **London with labels:** *"The image is a detailed map of London, England, showing the city's streets, landmarks, and surrounding areas."*

[3]https://carto.com/basemaps

Advances in Cartography and GIScience of the International Cartographic Association, 5, 12, 2025.
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada. This contribution underwent double-blind peer review based on the full paper. https://doi.org/10.5194/ica-adv-5-12-2025 | © Author(s) 2025. CC BY 4.0 License

- **London without labels:** *"The image displays a detailed map of London, England, showing the city's streets, landmarks, and waterways."*
- **Vancouver with labels:** *"The image is a detailed map of Vancouver, Canada, showing streets, landmarks, and a body of water."*
- **Vancouver without labels:** *"The image is a detailed map of a city, showing streets, parks, and a body of water."*

While the captions themselves give us a first indication about the performance of the VLM, we want to go further to determine which areas of the map image contribute to the generation of the specific text output, or more precisely, the specific token. To address this, we applied the workflow introduced in Section 3 to visualize the contribution of map regions to specific tokens within the generated text. Figure 6 shows a shortened and simplified plot for the map of London, highlighting the most relevant areas. Some full Shapley plots can be found in the Appendix, as well as on the corresponding GitHub repository.

In Figure 6 we can see red and blue areas in the images that resemble high and low contributions to text outputs. We can see that for the word "London", very specific areas in the map are highlighted. Looking closer, we can see that the highlighted areas have map labels saying "London" and "City of London".

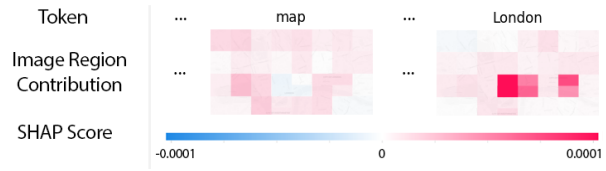


Figure 6. Contributions computed for the map of London, *including map labels*. The red/blue areas implicate high/low contribution to the corresponding text output of the network.

Figure 7 shows the Shapley plots for the same tokens that have been generated by the VLM for the map image without map labels. While the differences in Shapley Values for the token "map" are not significantly different, we clearly see that there are huge differences for the token "London". When no map labels are available, seemingly more areas of the map are contributing to the generation of the token "London".

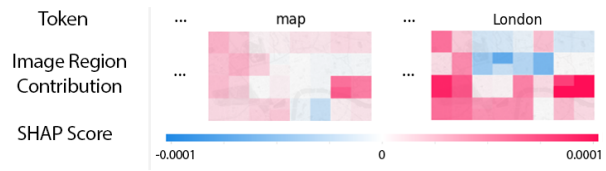


Figure 7. Contributions computed for the map of London, *excluding map labels*. The red/blue areas implicate high/low contribution to the corresponding text output of the network.

As we know by the VLM output for Vancouver including labels, the VLM in use has in fact generated the token "Vancouver". Figure 8 shows the Shapley Values for the particular token.

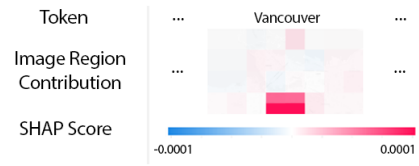


Figure 8. Contributions computed for the map of Vancouver, *including map labels*. The red/blue areas implicate high/low contribution to the corresponding text output of the network.

Upon investigation of the image the VLM was provided with, we see that the highlighted area in Figure 8 contains the map label "Vancouver". "Vancouver" not being mentioned in the VLM output for the map image that does not contain labels, a meaningful comparison is not possible.

We know from previous experience that changing the scale of the maps has an effect on the output of the VLM. This is why we have decided to repeat the same experiments for the scale of *1:25,000*. The VLM outputs generated for each map image are as follows:

- **London with labels:** *"The image displays a detailed map of a city, featuring streets, parks, and rivers."*
- **London without labels:** *"The image displays a detailed map of a city, featuring a river running through the center, surrounded by numerous streets and buildings. The map is predominantly yellow and green, with some blue and brown accents, providing a clear representation of the city layout."*
- **Vancouver with labels:** *"The image is a detailed map of a city, showing streets, buildings, and a body of water."*
- **Vancouver without labels:** *"The image is a detailed map of a city, featuring a grid-like pattern of streets and roads. The city is surrounded by water, with a coastline visible in the background. The map is predominantly light blue, with some darker blue areas representing water or land features. The streets are clearly marked, and the overall layout appears well-organized and easy to navigate."*

We can observe from the VLM outputs that the scale has a significant impact on the generated descriptions. Unlike the outputs generated for images with a scale of 1:50,000, when labels are excluded from the 1:25,000-scale maps, the VLM is not able to generate any location names. However, the level of detail in the descriptions has increased for 1:25,000-scale maps without labels. In the absence of labels, the VLM seems to emphasize visual features such as colors for water bodies and structural patterns such as street infrastructure.

Advances in Cartography and GIScience of the International Cartographic Association, 5, 12, 2025.
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada. This contribution underwent double-blind peer review based on the full paper. https://doi.org/10.5194/ica-adv-5-12-2025 | © Author(s) 2025. CC BY 4.0 License

5. Discussion

In our first experiment we have generated map descriptions with a VLM for London and Vancouver with a scale of 1:50,000, including and excluding map labels. When highlighting the contributions of image areas to the generated text, we can see that the highest contributions to the token "London" correspond to the regions of the map containing the map labels "London" (Figure 9). Especially the contribution on the right side of the image is interesting, since the contribution of the area containing the map label "London" has a higher contribution value of the area left of it containing the map label "City of". This indicates that, in fact, the VLM is able to recognize the word "London", ranking its contribution higher than the neighbouring area containing the words "City of". While this is what we would expect a VLM to do, we now not only have our intuition to rely on, but have been able to quantify and visualize this behavior using contribution analysis.

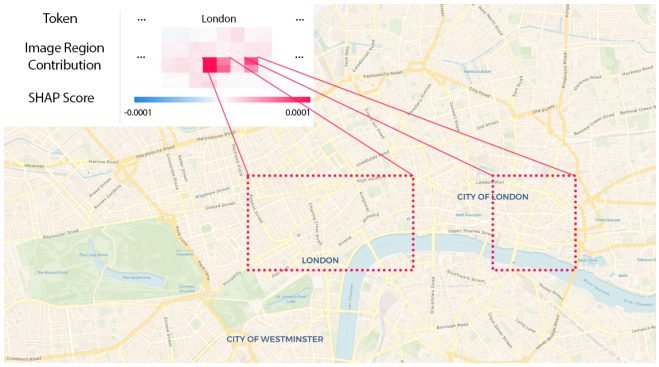


Figure 9. Locations of highest contributing areas to the generation of the token "London". Both areas include the literal word "London". Notably, the contribution of the area including the map label "London" in "City of London" is significantly higher than the contribution of the surrounding areas.

Examining our map of Vancouver, we can observe similar behaviour. In this case, the highest contributions to the token "Vancouver" align closely with the regions of the map that contain the map label "Vancouver" (Figure 10). The contribution values for the area containing the word "Vancouver" exceed those of surrounding regions without the label. This, again, indicates, that the VLM is able to directly identify and interpret specific map labels. Furthermore, the VLM seems to be able to generalize this behaviour across different map contexts.

The findings when providing the VLM with maps including labels confirm our intuition: the VLM is able to read map labels and by recognizing that it is being shown a map, it can connect the dots to indicate that we are dealing with a map of a specific location. At the same time, especially for the example of London including labels, we see that the word "London" is prioritized over "City of" as well as "City of Westminster", although they have the same style in terms of size and weight. This indicates that the used model has further context information it was provided with during its training. However, to make a solid statement in that regard, further research is needed.

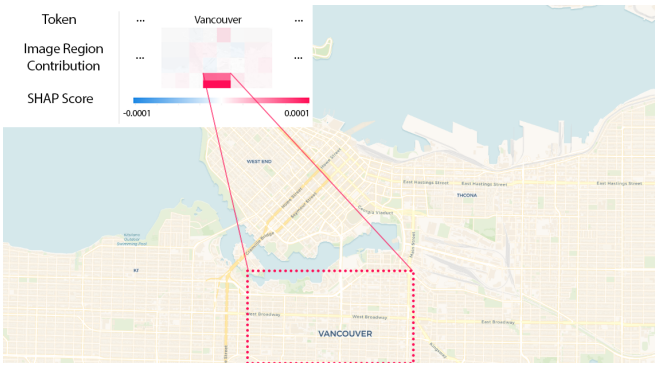


Figure 10. Location of the highest contribution area to the generation of the token "Vancouver". The highlighted area contains the literal word "London". This indicates, that the model in use is able to recognize the letters and make the connection that the underlying map is a map of Vancouver.

More interestingly, the model at hand was also able to recognize London when map labels were absent. In Figure 11 we observe that the areas of high contribution are more evenly distributed. This suggests, that the model identified "London" by recognizing enough spatial and contextual concepts resembling the city.

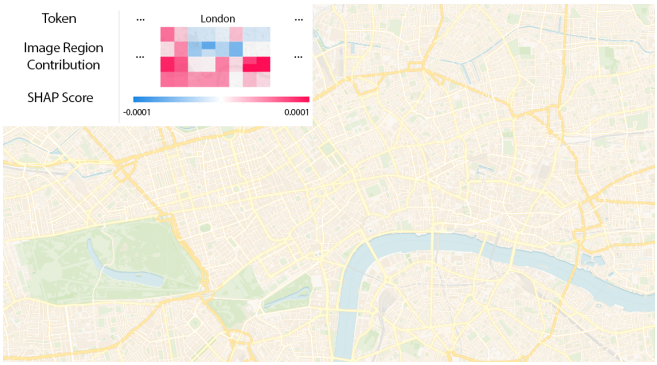


Figure 11. The contributing areas to generating the token "London" are more evenly distributed across the map than before. This indicates that the model identified "London" not through optical character recognition, but by recognizing enough spatial and contextual concepts of a map of London.

While this suggests that map labels can be excluded when using VLMs to generate map descriptions, excluding labels from the 1:50,000-scale map of Vancouver resulted in the VLM not being able to recognize the location. This could have many reasons - while it can be assumed that the model in itself is able to pick up enough context of a map, Vancouver might not have been included in the training dataset in the same extent as London. That being said, an overall limitation of our approach is that our findings are only valid for this specific network and model. We therefore expect that the results may vary with other models. Another general limitation when working with most models that are derived from foundation models is that we can not guarantee that the model never has seen any of the data used for our experiments.

Our second set of experiments was conducted on roughly the same areas with a scale of 1:25,000. Irrespective of labels, the VLM generates a more universal output, mentioning "map of a city" instead of a specific place or location name. When using labels, the generated captions were more concise and brief. In cases where labels were omitted, seemingly, the VLM compensates the absence of labels by adding more detail to its generated output. The exact reasoning behind the VLM focusing on design elements of the map is, however, still open to interpretation and requires further research.

6. Conclusion and Future Work

Our study aims to underscore the importance of understanding how neural networks interpret spatial data and derive meaning from map features and relationships. By utilizing Explainable AI (XAI), we can gain insights into the decision-making process of these models and contribute the development of more robust, interpretable and trustworthy algorithms. Our findings suggest that the scale and the presence of map labels can have significant impacts on VLM outputs. In particular, the inclusion of map labels led to mostly concise and accurate descriptions of the input images while maps without labels led to longer and more universal outputs. We could also find that with this specific stack, the model did not solely rely on map labels but seemingly understood the concept of "London" when conducting the experiment without map labels and the 1:50,000-scaled map.

Future work will focus on altering the masking techniques: The current Shapley-based approach relies on grid-like masking which comes with the risk that relationships between map elements are not captured. An approach such as selectively masking specific features and graphical elements such as the street networks, water bodies and altering colors could provide deeper insights into how individual map components influence model outputs. Future studies should expand the scope of experiments to include diverse map styles, scales and especially geographic regions to make statements about the generalizability of VLMs across different contexts. Since we believe that prompt engineering is not sustainable and does not align with reproducibble scientific work, we do not intend to pursue this approach further. All in all, with this work, we aim to contribute to pave the way for the development of more robust and interpretable VLMs to enhance their trustworthiness for critical tasks.

References

Dewi, C., Chen, R. C., Yu, H. and Jiang, X., 2023. XAI for Image Captioning using SHAP. *Journal of Information Science and Engineering* 39(4), pp. 711–724.

Díaz-Rodríguez, N. and Pisoni, G., 2020. Accessible cultural heritage through explainable artificial intelligence. In: *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, Association for Computing Machinery, New York, NY, USA, p. 317–324.

Ghandi, T., Pourreza, H. and Mahyar, H., 2024. Deep Learning Approaches on Image Captioning: A Review. *ACM Computing Surveys* 56(3), pp. 1–39. arXiv:2201.12944 [cs].

Han, S.-H., Kwon, M.-S. and Choi, H.-J., 2020. EXplainable AI (XAI) approach to image captioning. *The Journal of Engineering* 2020(13), pp. 589–594.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F. and Laga, H., 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys* 51(6), pp. 1–36.

Lundberg, S. M. and Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc., pp. 1–10.

Robinson, A. C. and Griffin, A. L., 2024. Using AI to Generate Accessibility Descriptions for Maps. *Abstracts of the ICA* 7, pp. 1–2.

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5), pp. 513–523.

Shapley, L. S., 1952. *A Value for N-Person Games*. RAND Corporation.

Vasiliev, Y., 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press.

Xu, J. and Tao, R., 2024. Map Reading and Analysis with GPT-4V(ision). *ISPRS International Journal of Geo-Information* 13(4), pp. 127.

Zhang, Z.-X. and Ochiai, Y., 2024. A Design of Interface for Visual-Impaired People to Access Visual Information from Images Featuring Large Language Models and Visual Language Models. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, Association for Computing Machinery, New York, NY, USA, pp. 1–4.

Zhao, Y., Zhang, Y., Xiang, R., Li, J. and Li, H., 2024. VIALM: A Survey and Benchmark of Visually Impaired Assistance with Large Models.

Appendix

The following images show, from left to right, further exemplary masked versions of the research area in 1:50,000 with and without labels, in 1:25,000 with and without labels and their corresponding VLM outputs.

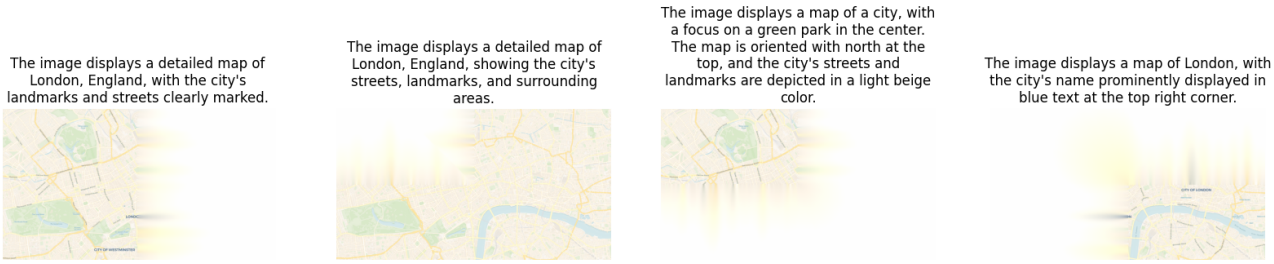


Figure 12. VLM outputs for London in 1:50,000 and 1:25,000 with and without labels.

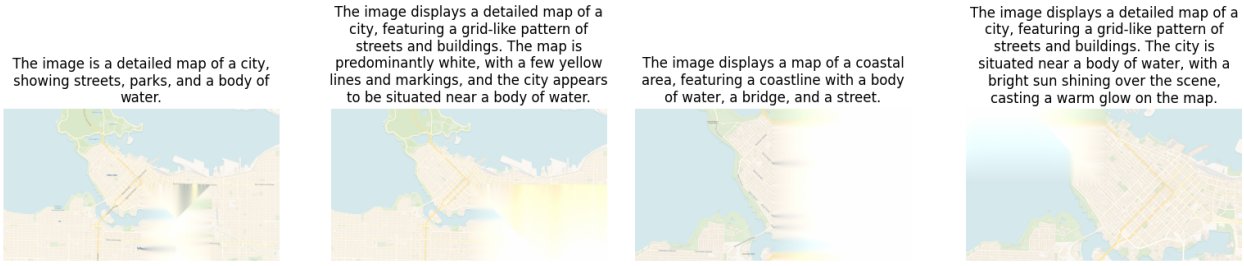


Figure 13. VLM outputs for Vancouver in 1:50,000 and 1:25,000 with and without labels.

The following images show the full Shapley plots for London and Vancouver, with labels with a scale of 1:50,000.

Figure 14. Shapley Plot showing contributions for the map of London with a scale of 1:50,000 including labels

Figure 15. Shapley Plot showing contributions for the map of Vancouver with a scale of 1:50,000 including labels