

Anonymous Authors

Figure 2: Bad examples of geo-tags enhanced text description generated by InstructBLIP.

- **Case 1.** When the Title Multi Objects contain elements outside the LMM's pretraining data (e.g., "tracktype is grade2"), there is a risk of generating nonsensical descriptions.
- **Case 2:** When a term within the Title Multi Objects is polysemous (e.g., "driving range"), the LMM may misinterpret its meaning, resulting in incorrect text generation.
- **Case 3:** If the satellite image itself presents complexities that are difficult for human interpretation, particularly when accompanied by misleading geo-tags information, the resultant descriptions can be significantly degraded.

The causes of the aforementioned issues can be summarized as follows. Firstly, certain details in satellite images pose inherent challenges for human perception. Additionally, there is the issue of "hallucination" within the LMM itself, constrained by the capacity limitations of multimodal foundational models.

3 IMAGE SEGMENTATION ANALYSIS

3.1 Image Segmentation and Text Enhancement

The integration of image segmentation and geo-tags enrichment is essential for generating semantically rich text descriptions that align precisely with visual data. Figure 3 presents a set of illustrative examples demonstrating this intricate interplay. In the top image, we observe a school area, where segmentation and corresponding geo-tags annotations such as "natural: scrub", "building: school", and "leisure: park" provide a detailed textual narrative. This enriching process allows for the precise capture of the surrounding environment, highlighting not just the main amenity but also peripheral elements such as adjacent roads and recreational areas.

Similarly, the bottom image showcases a healthcare facility marked by geo-tags indicating a "healthcare: laboratory" and "leisure: pitch", offering valuable insights into the facility's functionality and layout. These geo-tags are instrumental in guiding the segmentation algorithm to prioritize and detail salient features within the scene, such as parking amenities and service roads, which are pivotal for a comprehensive understanding of the landscape.



Figure 3: Illustrative examples of our image segments and geo-tags enhanced descriptions.

By integrating geo-tags information, we not only refine the image segmentation but also enable the generation of text that captures a more granular and accurate depiction of the scene. This

synergy between visual segmentation and textual enrichment is particularly evident when dealing with complex scenes where multiple objects or features must be precisely identified and described.

However, our methodology extends beyond simple annotation of images with geo-tags. It involves an iterative refinement where image segmentation informs the geo-tag-based text generation, and vice versa, leading to a fine-grained alignment that is greater than the sum of its parts. Such alignment is crucial for applications requiring detailed and context-aware descriptions of geographical spaces, as it allows for a nuanced interpretation of the environment, which is critical for accurate satellite image-text retrieval tasks.

3.2 Fine-tuning Segmentation Parameters

Our methodical calibration of the image segmentation process has conclusively established that six segments ($num_seg = 6$) provide the optimal representation of scenes for our modeling purposes. This specific number of segments was meticulously selected to adeptly capture the inherent complexity of diverse landscapes while avoiding the overburdening of the model with extraneous, non-essential details. The solid rationale behind this particular parameter choice is vividly demonstrated and well-supported by the statistical distributions, which are thoroughly illustrated in Figure 4.

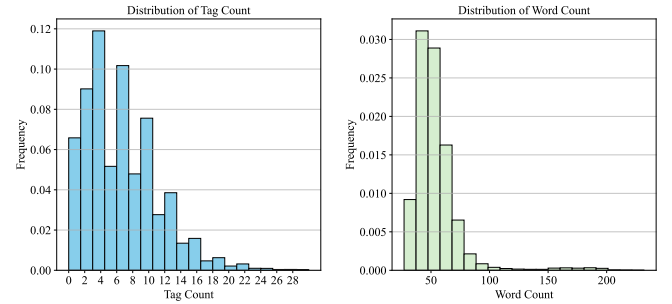


Figure 4: Statistical distribution of tag count and word count of descriptions in our dataset, illustrating the optimal balance of information richness and brevity.

The histogram on the left displays the frequency distribution of tag counts across our dataset. The majority of images have a tag count that clusters between four and ten, with a peak at six. This peak indicates that six tags often provide enough information to describe the essential elements of a scene without introducing noise through over-segmentation. The histogram on the right complements this finding by showing the word count distribution of the corresponding text descriptions. The majority of descriptions average approximately 54 words, achieving a balance between conciseness and detail necessary to comprehensively represent the scene for algorithmic processing and user interpretation.

This dual analysis of tag and word count distributions allows us to tailor our segmentation to align closely with the amount of detail that can be effectively described in texts, optimizing both the accuracy of image descriptions and the efficiency of our model. This fine-tuning ensures that each segment adds meaningful information to the generated description, facilitating a high level of detail in the semantic understanding of the scene, which is crucial for tasks such as image-text retrieval and scene comprehension.

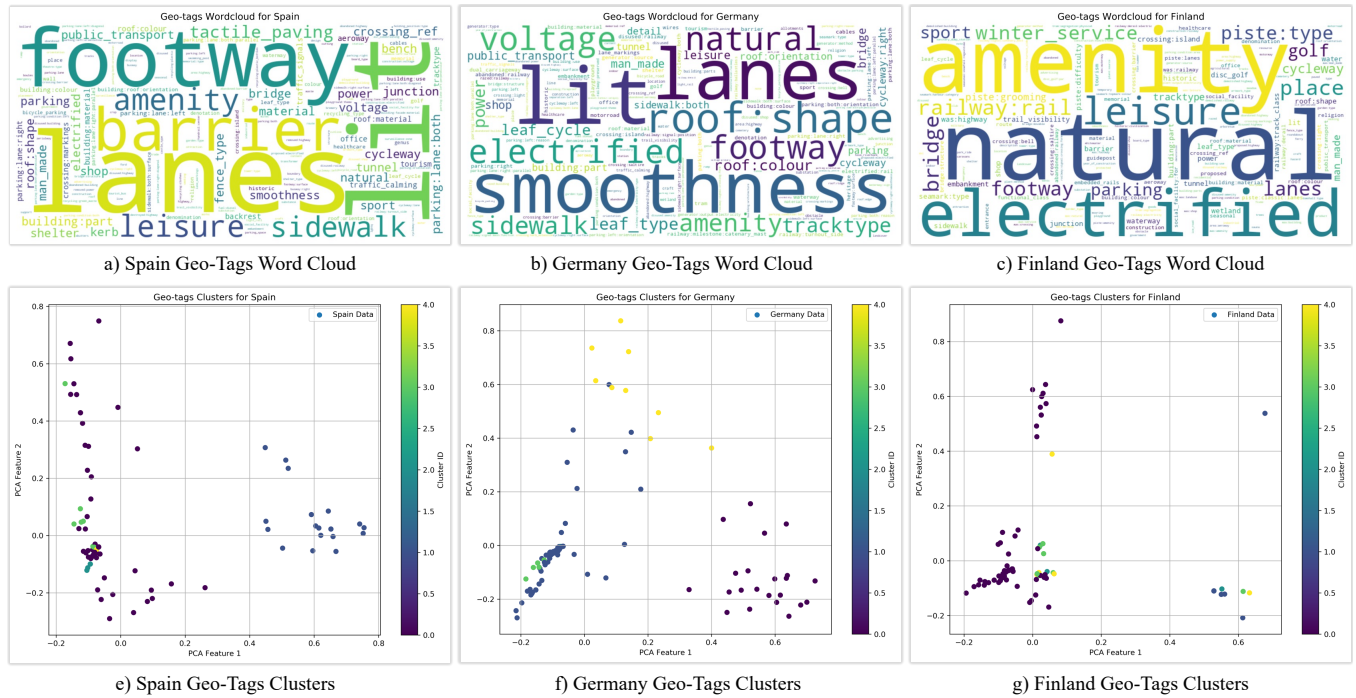


Figure 5: Geo-Tags Word Cloud: Visualizing the Most Frequent Tags for Spain (a), Germany (b), and Finland (c); and Geo-Tags Clusters: Spatial Distribution Based on PCA for Spain (e), Germany (f), and Finland (g) across three countries.

3.3 Future Segmentation Challenges

While current segmentation approaches yield substantial improvements in text-image alignment, several challenges remain. These include poor segmentation performance for small objects, a lack of quantified metrics for assessing text quality, and the absence of an automatic refinement mechanism for the segmentation process. Addressing these issues will be crucial for future advancements. In subsequent iterations, tags could play a more pivotal role in guiding the joint training of image-text pairs [8]. Moreover, the adoption of a Chain-of-Thought mechanism [4], which leverages feedback based on segmentation ratios, could enhance text quality further. These strategies offer promising avenues for improving the functionality and efficiency of our segmentation techniques.

4 GEO-TAGS ANALYSIS

4.1 Geo-Tags Distribution and Cluster Insights

In exploring geospatial metadata, geo-tags serve as critical tools for extracting geographical semantics embedded within satellite imagery. This analysis employs frequency analysis of geo-tags sourced from satellite images across three distinct European regions: Finland, Germany, and Spain. Utilizing word clouds and PCA clustering, this study visually represents the distribution and frequency of these tags, thereby highlighting the most prominent geographical features captured within the imagery.

Word clouds and PCA clustering diagrams provide dual modalities for in-depth visual analysis. Word clouds are particularly effective for quickly identifying the most frequent and prominent geo-tags, with larger font sizes indicating higher frequencies. This

visualization technique is augmented by PCA clustering, which groups geo-tags based on the similarity of their occurrences across various images, revealing underlying patterns that may not be immediately evident from the word clouds alone.

For instance, word clouds for Finland, Germany, and Spain showcase a distinct predominance of tags such as "natural", "building", and "waterway", amongst others. However, the PCA clustering diagrams further elucidate the relationships between these tags. In the case of Finland, the PCA clustering diagram shows that geo-tags associated with "piste:type" and "winter_service" cluster together separately from common tags, highlighting the distinctive winter sports environment in the area.

The clustering technique used in this analysis involves a multi-step process. Initially, the frequency of geo-tags is computed from the dataset, and a high-dimensional vector space is created. This space is then simplified to two principal components through PCA, capturing the most significant variances among the geo-tags. Subsequently, the clustering algorithm subsequently divides these tags into coherent clusters, each marked by a unique color.

4.2 Semantic Insights of Geo-Tags

The insights from the word cloud and PCA clustering analyses significantly enhance our understanding of image-text alignment in cross-modal retrieval tasks. Specifically, PCA clustering illustrates the semantic proximity of geo-tags within their clusters, indicating potential shared functionalities or relationships. These insights not only emphasize the diverse geographical semantics inherent to the studied countries but also have profound implications for cross-modal retrieval tasks.

By leveraging clustering insights, searches for "recreational areas" could, for instance, pull images tagged with "leisure", "park", and "natural", knowing these tags share semantic relationships.

In summary, the frequency and cluster analysis of geo-tags not only lay a foundational layer for semantic comprehension but also illustrate distinct geographical semantics across different countries, which is essential for enhancing cross-modal retrieval tasks. The refined analytical methods in this study pave the way for extracting more detailed and context-rich insights from geo-tagged data, enriching the narratives that satellite imagery can offer.

5 CONCLUSION

The analyses conducted in this supplementary material underscore the critical role of detailed component examination in advancing the UrbanCross framework. Through rigorous investigation into text generation, image segmentation, and geo-tags clustering, we have identified several key areas for enhancement and potential pitfalls that may impede performance. The insights derived from these studies not only refine our current understanding but also pave the way for future research, particularly in enhancing cross-modal retrieval accuracy and the semantic alignment of geo-spatial metadata. Moving forward, these findings will guide the further development of the UrbanCross framework, ensuring its adaptability and efficacy in handling diverse and complex geo-spatial datasets. This document not only serves as a testament to the depth of our analysis but also as a blueprint for ongoing and future innovations in the field of satellite imagery retrieval.

REFERENCES

- [1] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478* (2023).
- [2] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793* (2023).
- [3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv:2305.06500* [cs.CV]
- [4] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems* 36 (2024).
- [5] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).
- [6] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2024. UrbanVLP: A Multi-Granularity Vision-Language Pre-Trained Foundation Model for Urban Indicator Prediction. *arXiv preprint arXiv:2403.16831* (2024).
- [7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7514–7528. <https://doi.org/10.18653/v1/2021.emnlp-main.595>
- [8] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2024. Tag2Text: Guiding Vision-Language Model via Image Tagging. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=x6u2BQ7xcq>
- [9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

- et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [11] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).