essd-2024-140 — Revision

ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models

By Z. Yuan, Z. Xiong, L. Mou, X. X. Zhu

General remarks to all reviewers and editors:

We sincerely thank the editors and anonymous reviewers for their valuable comments and suggestions. Below, we provide point-by-point responses to the reviewers' comments. The reviewers' comments are in black, and our responses follow in blue. The revised parts are marked in red in the manuscript.

Reviewer #2:

1. Throughout the paper, the author mentioned image-text datasets many times. Image-text datasets cover multiple different types of annotations, such as image caption, VQA, and visual grounding. Since this paper focuses on image captioning, the writing should be modified accordingly.

R: We appreciate the reviewer's suggestion on this point. We agree with the reviewer's understanding of the concept of image-text dataset. Image-text datasets indeed contain multiple types of text annotations, including image captioning, visual question answering, and visual grounding. However, our dataset specifically provides long, detailed descriptions of images, particularly focusing on land cover types and their spatial distribution. While these descriptions are closely related to image captioning tasks, they contain richer information that extends beyond typical captioning tasks. It can be readily used for other generative tasks, such as image-to-text and text-to-image synthesis. In addition, the dataset can also be easily extended to visual question answering by leveraging the capabilities of current large language models. This versatility is why we refer to it as an "image-text dataset," a high-level term that captures its potential for a range of tasks.

In light of this, we choose to use the broader term "image-text dataset" to reflect the higher-level concept of images paired with textual descriptions. This is consistent with prior works [1]-[5], which also use "image-text dataset" when focusing primarily on image captioning tasks.

[1] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data," *arXiv preprint arXiv:2001.07966*, 2020.
[2] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[3] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "RedCaps: Web-curated image-text data created by the people, for the people", *arXiv preprint arXiv:2111.11431*, 2021.
[4] Y. Okamoto, H. Toyonaga, Y. Ijiri, and H. Kataoka, "Constructing image-text pair dataset from books," *arXiv preprint arXiv:2310.01936*, 2023.
[5] Q. Yu, Q. Sun, X. Zhang, Y. Cui, F. Zhang, Y. Cao, X. Wang, and J. Liu, "Capsfusion: Rethinking image-text data at scale," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

2. The authors use land cover labels from WoldCover products to formulate prompts. The information carried by image captions mainly covers land cover information, this limits the usage of the proposed dataset. This is a big drawback when compared to previous datasets (e.g., RSICap) that provide more diverse information (such as object counting, position, size, and complex reasoning).

R: Thanks for the insightful comments. RSICap [6] is an excellent dataset that offers diverse and detailed annotations, but it has a relatively smaller volume and relies on manual annotation. In contrast, our dataset leverages automated methods to generate a significantly larger volume of image-text pairs, ensuring broader coverage and scalability. Our dataset consists of satellite images with global coverage and lower resolution. This makes object counting and complex reasoning more challenging due to the granularity of the images.

Although we use land cover labels from WorldCover products to formulate prompts, our dataset also includes detailed descriptions related to position and size. For example: "This image reveals a mix of developed areas and trees, with developed areas showing expansive coverage particularly in the top left, signifying widespread human settlement or infrastructure. Bodies of water are substantially present, especially in the top left, forming large open shapes indicative of lakes or wide rivers. Trees spread significantly across the bottom half, offering a sense of a forested or natural region, while grasslands are present but less dominant. Varying shapes in the pattern of developed areas and the strong presence of water features characterize this image alongside the notable forest coverage."

In summary, we believe that both RSICap and our dataset offer valuable contributions to the community, but with distinct focuses. RSICap emphasizes high-resolution object recognition, counting, and attribute analysis, while ChatEarthNet focuses on land cover types and global coverage.

[6] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "RSGPT: A remote sensing vision language model and benchmark." *arXiv preprint arXiv:2307.15266*, 2023.

3. Information Overlap. To generate image captions, the proposed method divides each image of 256x256 into 5 patches of size 128x128, top-left, top-right, bottom-left, bottom-right, and middle patches. The center patch overlaps with other patches. This causes two issues: 1) duplicated object description; 2) duplicated object counting.

R: We appreciate the reviewer's comment on this point. In real-world scenarios, it is common for land cover types to span across multiple regions, including overlapping areas.

While our method involves dividing the image into patches with some overlap, this does not affect describing each patch individually. Specifically, we divide one image into five patches, including a central one that overlaps with the others, to ensure a comprehensive description of the spatial distribution across the entire image. Without this overlap, the central portion of the image might be overlooked, leading to incomplete coverage of the spatial pattern.

Since our dataset focuses on land cover types, which often lack distinct object boundaries, there is no issue of duplicated object descriptions. We also notice that there are no redundant descriptions in overlapping areas. Additionally, ChatEarthNet does not involve object counting in its captions. Therefore, the overlap does not introduce any issues related to duplicated object counting, ensuring that the dataset remains unaffected in this case.

4. "Moreover, considering the API request limit of ChatGPT-4V, we put four images into one request to generate descriptions more efficiently". By putting four images into one request, do you mean concate the images into one? Merging multiple images will cause undesired interactions between image features caused by self-attention in transformer architecture. As far as I know, GPT-4V allows 10,000 requests per day, it's therefore not necessary to put four images into one request.

R: Thank you for the insightful comment. To clarify, we do not concatenate four images into one. Instead, we send four separate images in a single API request to ChatGPT-4V. Therefore, there are no interactions between image features at the model level. However, in a few cases, the returned descriptions include comparisons between different images, which is not our intention. To address this, we manually review and correct such descriptions to ensure quality.

In addition, we would like to clarify three points regarding our decision to put four images into a single request when generating captions with ChatGPT-4V.

1) **Our work began in 2023, and the first version of the manuscript was submitted in February 2024. At that time, for usage tier 1, the limit was set at 500 requests per day, not the 10,000 requests per day that are available now. Given the resource constraints we faced at that time, we chose to put four images into one request to generate descriptions more efficiently.**
2) **To ensure that the images are described independently, our prompts specifically request: "Generate the four descriptions separately; do not add connections between them."**
3) **Despite the prompt requesting no interactions between images, some descriptions still contain comparisons among the four images. To ensure quality, we manually check all captions generated by ChatGPT-4V and refine comparison-related captions.**

5. Missing experimental verification. By the current version, it's unclear how this dataset can be used to boost the development of LVLMs in remote sensing. As a benchmark dataset, it's better to show the image captioning performance of existing well-known methods on the proposed dataset.

R: We thank the reviewer for the valuable comment. We agree that demonstrating ChatEarthNet's utility for evaluating vision-language geo-foundation models is crucial. To address the reviewer's concerns, we have conducted additional benchmarking experiments using widely established multimodal large language models (MLLMs). Specifically, we evaluated several MLLMs, including LLaVA-v1.5 [7], MiniGPT-v2 [8], MiniGPT-4 [9], and GeoChat [10]. These evaluations further support our conclusion that ChatEarthNet is a valuable resource for training and evaluating vision-language geo-foundation models for remote sensing. Please kindly check out the revised version as follows.

"To demonstrate the effectiveness of ChatEarthNet in evaluating multimodal large language models, we conduct benchmarking experiments using a range of existing models. Given that ChatEarthNet includes long and detailed descriptions, it is not well-suited for evaluating CLIP-based vision-language models like RemoteCLIP [11] and RS-CLIP [12]. Therefore, we focus on evaluating widely used multimodal large language models, including LLaVA-v1.5 [7], MiniGPT-v2 [8], MiniGPT-4 [9], and GeoChat [10]. All experiments are performed using the ChatGPT-4V version of our dataset, which allows us to conduct extensive evaluations across multiple models while significantly reducing computational resource requirements.

Table 2 summarizes the results of these evaluations, detailing the models' performance across several widely used metrics: BLEU, CIDEr, METEOR, ROUGE-L, and SPICE. We evaluate these models in two experimental settings. The first is a zero-shot transfer setting, where pre-trained models are used to generate captions without any additional training or fine-tuning on the ChatEarthNet dataset. The first four rows in Table 2 present the results of this zero-shot transfer setting. The performance is suboptimal due to the domain gap between the models' original training datasets and our test dataset. In addition to zero-shot testing, we fine-tune some of these models on the ChatEarthNet dataset (ChatGPT-4V version) and report their performance. The results clearly show that fine-tuning on our proposed dataset significantly improves image captioning performance in the context of remote sensing data. These findings strongly suggest that ChatEarthNet is a valuable resource for both training and evaluating vision-language geo-foundation models in the remote sensing domain."

Table 2. Performance comparison of different models on the ChatEarthNet (ChatGPT-4V Version) test set.

Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	METEOR	ROUGE_L	SPICE
LLaVA-v1.5	0.285	0.116	0.040	0.014	0.012	0.104	0.186	0.093
MiniGPT-v2	0.279	0.116	0.041	0.015	0.009	0.104	0.180	0.091
MiniGPT-4	0.175	0.072	0.023	0.008	0.000	0.116	0.180	0.079
GeoChat	0.199	0.088	0.034	0.011	0.005	0.067	0.126	0.083
MiniGPT-4 (ChatEarthNet)	0.310	0.184	0.113	0.071	0.001	**0.209**	0.254	0.186
GeoChat (ChatEarthNet)	**0.445**	**0.269**	**0.170**	**0.109**	**0.094**	0.208	**0.286**	**0.211**

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems 36*, 2023.

[8] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
[9] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
[10] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
[11] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "RemoteCLIP: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[12] X. Li, C. Wen, Y. Hu, and N. Zhou, "RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision," *International Journal of Applied Earth Observation and Geoinformation*, 2023.

Minors:

1. ChatGPT-3.5 is not a widely used term. Instead, ChatGPT and gpt-3.5-turbo are more frequently used.

R: Thanks for the valuable comment. In this manuscript, we use "ChatGPT-3.5" to refer to the model technically known as "gpt-3.5-turbo." Similarly, "ChatGPT-4V" refers to "gpt-4-vision-preview." These terms are intended to provide a more intuitive understanding of the models' positions within the ChatGPT series.

We have added the following sentence in the revised manuscript to clarify this as follows.

"In this manuscript, ChatGPT-3.5 refers to the model gpt-3.5-turbo and ChatGPT-4V refers to the model gpt-4-vision-preview."

2. In line 67, referring image segmentation belongs to visual grounding and therefore should be merged.

R: Thank you for your comment. We believe you may be referring to line 37. While both visual grounding and referring image segmentation are vision-language tasks, they produce different types of outputs. Visual grounding generates a bounding box around the referred object, while referring image segmentation produces a pixel-level mask for the object based on the query. Given this fundamental difference in output, we choose to keep them as separate tasks in the manuscript.

3. In line 44, when mentioning large vision-language foundation models, the authors fail to cover popular models, such as MiniGPT-4, and QWen-VL.

R: Thank you for pointing out this oversight. We have revised the manuscript to include more models as follows.

"For large vision-language foundation models, CLIP [13], LLaVA [7], MiniGPT-4 [9], MiniGPT-v2 [8], and Qwen-VL [14] have revolutionized the computer vision community."

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems 36*, 2023.
[8] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
[9] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *Proceedings of the 38th International Conference on Machine Learning*, 2021.
[14] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," *arXiv preprint arXiv:2308.12966*, 2023.

4. In Table I, it's unclear whether the 10,000 images used with GPT-4V are included in those 163,488 images used with GPT-3.5. If included, the second column can be removed.

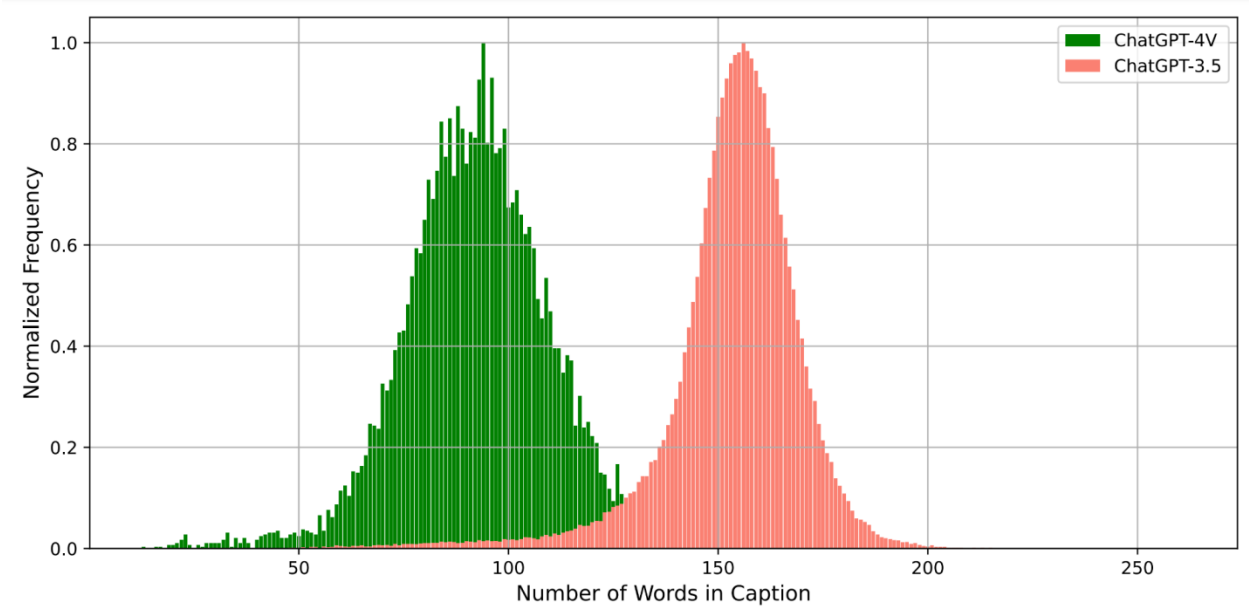
R: Thanks for the insightful comment. The 10,000 images used with ChatGPT-4V are included in those 163,488 images used with GPT-3.5. Following the reviewer's suggestion, we have removed the second column in Table I in the revised manuscript.

Table 1. The number of Sentinel-2 images used for generating captions, along with the corresponding numbers of captions generated by ChatGPT-3.5 and ChatGPT-4V.

Subsets	Number of ChatGPT-3.5 Captions	Number of ChatGPT4-V Captions
Train	98,092	6000
Val	16,348	1000
Test	49,048	3000
Sum	163,488	10,000

5. In Fig. 15, it's better to show the y-axis with probability distribution instead of No. images for a fair comparison between GPT-3.5 and GPT-4.

R: Thanks for the valuable comment. We have revised the Fig. 15 to normalize the frequency for a better visual comparison. Please kindly check it out as follows.



6. Section 3.3 can be compressed.

R: Thank you for the insightful suggestion. We have revised Section 3.3 to make it more concise. Please kindly check out the revised manuscript.