Response letter to the Manuscript # Earth System Science Data-2025-710

Response to the referee

Referee #1

Overall comments

> It is a timely and valuable dataset that can be useful for not only building classification but training a multi-modal AI model to a better understanding of the function of urban buildings as well as the underlining motivation of human activities and movements. To the best of my knowledge, it is the first multimodal dataset dedicated to building function classification that offers 26 distinct, fine-grained categories. This is a significant improvement over existing schemes that often mirror coarse land-use classifications. The study challenges the conventional belief that large models cannot handle multimodal spatial data, which is a high-quality contribution. Overall, the paper is well-structured, the methodology is sound, and the dataset indeed fills a clear gap in the Earth System Science community. Some issues in below should be addressed before publication.

Reply:

> We thank referee #1 for recognizing the value of our work. We are glad to address his/her comments below. We highlighted the added text as text and the deleted text as ~~text~~ in both this letter and the revised manuscript.

Referee Comment 1.1

> The dataset is exclusively sampled from New York City rather than from a broader geography. The authors acknowledge that conclusions may be subject to urban bias, as variations in model performance and error distributions might occur in cities with different layouts. A more in-depth discussion of impacts of the geography bias on the inference of models as well as the results of the baselines should be provided.

Reply:

> Thanks for your highly valuable comment. We fully agree that an in-depth discussion on the impact of geographic sample bias on both the inference models and the baselines should be provided. In response to your comment, we have added a subsection titled "Limitations of BuildingSense" in the Discussion section, where we thoroughly discuss the impact of this bias on the baselines in our study, as well as on the performance of models trained on this dataset in future research.

Changes:

6.3 Limitations of BuildingSense

6.3.1 City sample bias

NYC, as an international metropolis, has a diverse ethnic composition and high population density, which contribute to its varied architectural styles, including row houses, modern commercial buildings, grand public structures, and towering skyscrapers. Thus, the building function dataset we constructed from NYC samples exhibits a certain degree of representativeness for the North American context. However, the architectural styles that differ considerably from those in North America, such as East Asia and Europe, may lead to two potential outcomes: (1) the evaluation of the optimal large model based on the baseline results may be overestimated, and (2) the performance of models trained on this dataset may degrade when transferred to regions with substantially different architectural styles.

First, our evaluation of the optimal large model (Gemini-2.5-flash (Thinking)) indicates that it can effectively integrate multimodal spatial data and perform logical building function inference, even though it was not specifically designed for this task. During this process, it demonstrates strong capabilities in image understanding, text processing, and spatial reasoning. As a flagship product of Google, we hypothesize that its training data likely includes extensive Google spatial data (e.g., street view imagery, POI reviews, and remote sensing data), which may account for its superior performance compared to other models. However, in regions such as China, the distribution of information across data modalities (visual, textual, spatial) differs significantly from that in NYC. For instance, Chinese urban villages present a unique challenge: densely built, low-rise residential areas contain informal commercial activities on the ground floor. Visually, a building may appear entirely residential (with laundry hanging from windows, narrow alleyways, and residential-style architecture), while POI data might indicate numerous small businesses (e.g., convenience stores, hair salons, street food vendors) operating within. Such a condition may lead to brittle reasoning chains and performance decline of baselines in four primary aspects: (1) an inability to analyze architectural styles from street view images; (2) a failure to semantically interpret POI names within buildings due to linguistic differences; (3) an incapacity to assess the built environment and architectural styles from remote sensing imagery; and (4) incorrect spatial reasoning resulting from any of the aforementioned errors. This implies that our evaluation of the optimal large model may be overestimated.

Second, despite these limitations, Gemini's strong performance on BuildingSense within its training distribution still offers valuable insights: the technical approach of inferring building functions using large models is feasible, and fine-tuning such models with relevant data could significantly enhance their performance on building function classification tasks. Consequently, subsequent large models fine-tuned on BuildingSense may only achieve performance improvements within the North American context, with relatively limited gains when applied to other regions. Furthermore, for traditional deep learning models, limited

generalizability beyond the training distribution has always been a primary drawback. Models trained on BuildingSense may therefore not be suitable for application outside North America. Nevertheless, BuildingSense remains an important dataset for validating model performance.

To address this limitation, our future work will focus on three directions: (1) extending BuildingSense to include multiple cities representing diverse urban typologies (e.g., a European compact city, an Asian high-density city, and a Latin American spontaneous city); (2) developing domain adaptation techniques to transfer knowledge from NYC to data-scarce regions; and (3) conducting systematic cross-city evaluations to quantify the generalizability of both traditional and large models.

Referee Comment 1.2

One important contribution of the dataset is the rich function Table 3 lists existing building datasets without indicating the number of building function categories. I would suggest adding # of categories in the table.

Reply:

Thanks for your constructive feedback, which made us realize that the "N" column we originally added in Table 3 was not intuitively clear in terms of its abbreviation. Therefore, we have replaced it with the "CN" column, which represents the number of building-function categories for each dataset.

Changes:

Table 1: Building-related dataset. NB refers to the number of buildings; Balance refers to whether the dataset considers each category sample of the dataset is balanced or not; POI refers to point of interest; RSI refers to remote sensing image; SVI refers to street view image; RN refers to road network; H refers to building height; Y refers to building year; roof refers to building roof; F refers to building function; ~~N~~CN refers to the number of function category.

Dataset	NB	Balance	POI	RSI	SVI	RN	H	Y	Roof	F	~~N~~CN	Task
KITTI [1]	-				✓							S
Cityscapes [2]	-				✓							S
EuroCity [3]	-				✓							O
WildPASS [4]	-				✓							S
PASS [5]	-				✓							S
HoliCity [6]	-				✓							S
SkyScapes [7]	-			✓								S
SpaceNet [8]	-			✓								S
Li et al.[9]	-		✓									S
TorontoCity [10]	400000			✓	✓		✓					S
Wojna [11]	9674			✓	✓				✓	✓	6	C
OmniCity [12]	-			✓	✓		✓			✓	7	S
CMAB [13]	31000000						✓	✓	✓	✓	6	P
He et al.[14]	500000			✓							10	C
Ours	**34458**	✓	✓	✓	✓	✓	✓	✓	✓	✓	**26**	**C**

Referee Comment 1.3

> The method section describes a workflow. I suggest adding a workflow chart at the beginning of the method section to make the readers easily understanding the process.

Reply:

> Thank you for pointing out the insufficient clarification of the connections between the methods. Following your suggestion, we have added a workflow chart and corresponding textual descriptions in the Method section, which clearly illustrate the relationships between our methods.

Changes:

As illustrated in the Figure 1, the methodological workflow of our dataset consists of three components: (1) building-related data and annotation collection, (2) building annotation and data cleaning, and (3) evaluation on large models. The first component describes how building footprints were sampled and how the relevant data and annotations were collected based on these footprints (Section 3.1). The second component details the cleaning process for the collected data and annotations, ultimately yielding a high-quality multimodal dataset (Section 3.2). The third component outlines the setup for baseline comparisons to evaluate the performance of different large models and the evaluation method for the optimal model (Section 3.3).

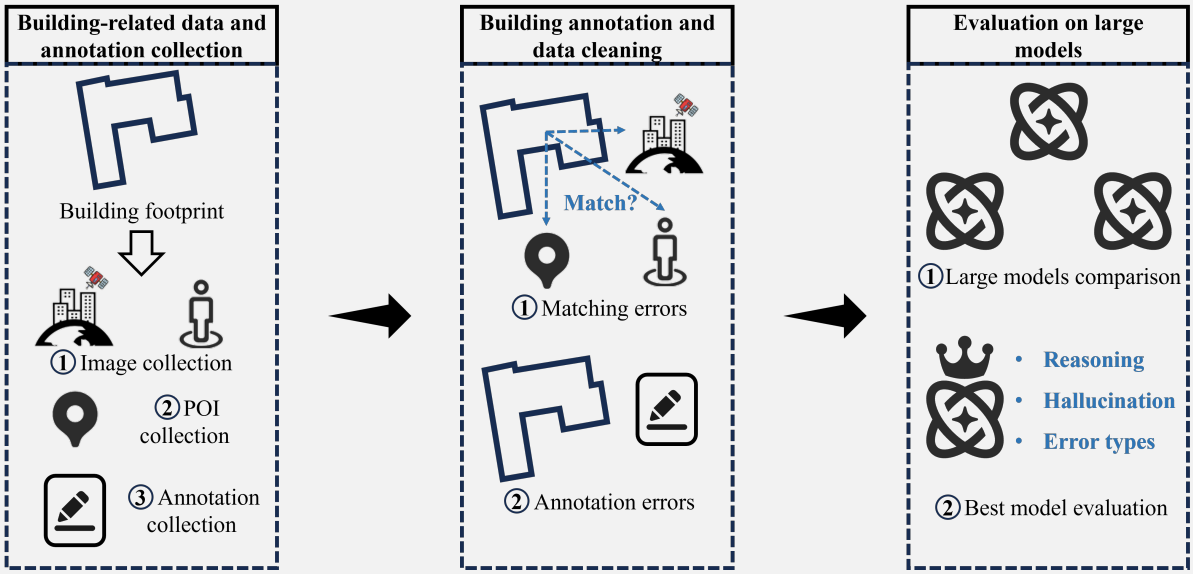

Figure 1: The workflow chart

Referee Comment 1.4

It is good to have an evaluation for the performance of the large models handling the multimodal dataset. However, you should give a logical and clear justification of the reasons of doing the evaluation. Why did you only involve those large models? Any specific reasons of not testing popular machine learning and deep learning methods?

Reply:

Thank you for highlighting the lack of clarity in the rationale for excluding popular machine learning and deep learning methods, which Referee #1 also mentions. In response to your comment, we have added detailed illustrations in the Baselines subsection.

Here is the logic: the exclusion of traditional deep learning and machine learning methods from the baselines is motivated by two points: (1) two specific bottlenecks identified in the existing literature on building function classification (Section 2.1) and (2) the fundamental differences in their evaluation paradigms and spatial scalability compared to large models.

Given these limitations, large models have emerged as a promising approach. Its extensive world knowledge and human-like reasoning abilities (spatial scalability) enable it to jointly interpret visual and textual cues (sufficient feature extraction and fusion), thereby explicitly articulating its inference processes (interpretability). Therefore, testing the hypothesis that large models can overcome traditional limitations in multimodal building-function classification is extremely valuable.

Changes:

The exclusion of traditional deep learning and machine learning methods from the baselines is motivated by two points: (1) two specific bottlenecks identified in the existing literature on building function classification (Section 2.1) and (2) the fundamental differences in their evaluation paradigms and spatial scalability compared to large models.

First, two specific bottlenecks are as follows: (1) insufficient feature extraction and fusion [15, 16, 17, 18, 19, 20, 21, 22, 14, 23, 24] and (2) poor interpretability of model outputs [21, 22, 14, 23, 24]. Second, the differences in their evaluation paradigms and spatial scalability are as follows: (1) from the perspective of evaluation paradigms, we assess the out-of-the-box zero/few-shot reasoning capabilities of large models on the full dataset. In contrast, evaluating traditional machine learning/deep learning methods requires partitioning the majority of the data into a training set. Such an approach allows the models to learn data characteristics in advance, leading to better-fitting results. Therefore, it is unfair to directly compare large models with deep learning models trained for building function classification. (2) From the perspective of spatial scalability, traditional deep learning models inherently overfit on the training dataset, leading to performance decreasing when transferred to another region or when the data format changes. Such a limitation renders them unscalable for large-scale urban applications.

Given these limitations, large models have emerged as a promising approach. Its extensive world knowledge and human-like reasoning abilities (spatial scalability) enable it to jointly interpret visual and textual cues (sufficient feature extraction and fusion), thereby explicitly articulating its inference processes (interpretability) [25, 26]. These characteristics directly target the limitationsmentioned above, which remain underexplored. Therefore, we select four state-of-the-art large models to test the hypothesis that large models can overcome traditional limitations in multimodal building-function classification.

Referee Comment 1.5

The authors claim that the sampling processing of the building footprints significantly reduces geospatial bias. In the BuildingSense section, a map showing the geographical distribution of the samples should be provided. It can help readers better understand the data quality in terms of geographical coverage.

Reply:

Thank you for pointing out the insufficient illustration of our data coverage. In response, we have added a map showing the geographic distribution of building samples in the BuildingSense section to better visualize our data coverage (Figure 2).

Changes:

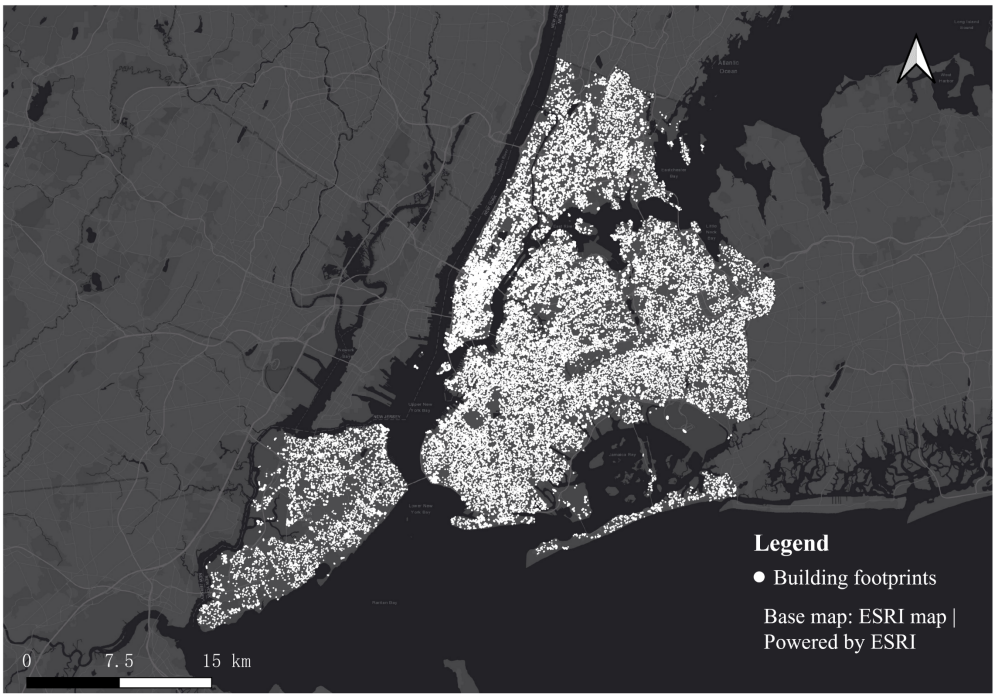

Figure 2: Distribution of building footprints

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[3] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1844–1861, 2019.

[4] K. Yang, X. Hu, and R. Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30:1866–1881, 2021.

[5] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4171–4185, 2020.

[6] Yichao Zhou Ma, Jingwei Huang, Xili Dai, Shichen Liu, Linjie Luo, Zhili Chen, and Yi. Holicity: A city-scale data platform for learning holistic 3d structures. *arXiv*, 2008.03286, 2021.

[7] Seyed Majid Azimi, Corentin Henry, Lars Wilko Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7392–7402, 2019.

[8] N. Weir, D. Lindenbaum, A. Bastidas, A. Etten, V. Kumar, S. Mcpherson, J. Shermeyer, and H. Tang. Spacenet mvoi: A multi-view overhead imagery dataset. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 992–1001, 2019.

[9] W. Li, L. Meng, J. Wang, C. He, G. S. Xia, and D. Lin. 3d building reconstruction from monocular remote sensing images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12528–12537, 2021.

[10] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun. Torontocity: Seeing the world with a million eyes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3028–3036, 2017.

[11] Zbigniew Wojna, K. Maziarz, L. Jocz, R. Paluba, R. Kozikowski, and I. Kokkinos. *Holistic Multi-View Building Analysis in the Wild with Projection Pooling.* 2021.

[12] W. Li, Y. Lai, L. Xu, Y. Xiangli, J. Yu, C. He, G. S. Xia, and D. Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17397–17407, 2023.

[13] Yecheng Zhang, Huimin Zhao, and Ying Long. Cmab: A multi-attribute building dataset of china. *Scientific Data*, 12(1):430, 2025.

[14] Da He, Xiaoping Liu, Qian Shi, and Yue Zheng. Visual-language reasoning segmentation (larse) of function-level building footprint across yangtze river economic belt of china. *Sustainable Cities and Society*, 127, 2025.

[15] Yue Zheng, Xucai Zhang, Jinpei Ou, and Xiaoping Liu. Identifying building function using multisource data: A case study of china's three major urban agglomerations. *Sustainable Cities and Society*, 108, 2024.

[16] Dongfeng Ren, Xin Qiu, and Zehua An. A multi-source data-driven analysis of building functional classification and its relationship with population distribution. *Remote Sensing*, 16(23), 2024.

[17] Abdulkadir Memduhoglu, Nir Fulman, and Alexander Zipf. Enriching building function classification using large language model embeddings of openstreetmap tags. *Earth Science Informatics*, 17(6):5403–5418, 2024.

[18] Shouhang Du, Meiyun Zheng, Liyuan Guo, Yuhui Wu, Zijuan Li, and Peiyi Liu. Urban building function classification based on multisource geospatial data: a two-stage method combining unsupervised and supervised algorithms. *Earth Science Informatics*, 17(2):1179–1201, 2024.

[19] Wei Chen, Yuyu Zhou, Eleanor C. Stokes, and Xuesong Zhang. Large-scale urban building function mapping by integrating multi-source web-based geospatial data. *Geo-Spatial Information Science*, 27(6):1785–1799, 2024.

[20] Yingbin Deng, Renrong Chen, Ji Yang, Yong Li, Hao Jiang, Wenyue Liao, and Meiwei Sun. Identify urban building functions with multisource data: a case study in guangzhou, china. *International Journal of Geographical Information Science*, 36(10):2060–2085, 2022.

[21] Weijia Li, Jinhua Yu, Dairong Chen, Yi Lin, Runmin Dong, Xiang Zhang, Conghui He, and Haohuan Fu. Fine-grained building function recognition with street-view images and gis map data via geometry-aware semi-supervised learning. *International Journal of Applied Earth Observation and Geoinformation*, 137, 2025.

[22] Zhiyi He, Wei Yao, Jie Shao, and Puzuo Wang. Ub-finenet: Urban building fine-grained classification network for open-access satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 217:76–90, 2024.

[23] Eike Jens Hoffmann, Karam Abdulahhad, and Xiao Xiang Zhu. Using social media images for building function classification. *Cities*, 133, 2023.

[24] Jiaxin Zhang, Tomohiro Fukuda, and Nobuyoshi Yabuki. Development of a city-scale approach for facade color measurement with building functional classification using deep learning and street view images. *ISPRS International Journal of Geo-Information*, 10(8), 2021.

[25] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,

Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[26] Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng-Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yi-Chen Guo, Hui Xiong, Qun Liu, and Zhenguo Li. A survey of reasoning with foundation models. *ArXiv*, abs/2312.11562, 2023.