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

Response to the referee

Referee #2

Overall comments

> This paper constructs a multimodal fine-grained dataset for building function classification, covering remote sensing imagery, street-view images, POI data, and building attributes for over 34,000 buildings in New York City. Through a systematic literature review, it identifies current methodological bottlenecks in building function classification and the scarcity of datasets, while revealing both the potential and research gaps of large models in this task. The subsequent benchmark experiments with large models demonstrate the promise of multimodal fusion for understanding building functions, challenging prevailing views in the field. The dataset construction process is solid and includes appropriate quality-control measures. The paper is timely and valuable, and falls within the scope of ESSD. However, there are some issues that need to be addressed before publication.

Reply:

> We thank referee #2 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 2.1

> As the first contact with readers, the abstract should briefly highlight the dataset's scale (e.g., number of buildings, modalities included).

Reply:

> Thank you for your constructive suggestion. We have added the relevant information about the dataset in the Abstract.

Changes:

Building function is a description of building usage. The accessibility of its information is essential for urban research, including urban morphology, urban environment, and human activity patterns. Existing building function classification methodologies face two major bottlenecks: (1) poor model interpretability and (2) inadequate multimodal feature fusion. Although large models with strong interpretability and efficient multimodal data fusion capabilities offer

promising potential for addressing the bottlenecks, they remain limited in processing multimodal spatial datasets. Their performance in building function classification is therefore also unknown. To the best of our knowledge, there is a lack of multimodal building function classification datasets, which results in the challenge of effectively performing their performance evaluation. Meanwhile, prevailing building function categorization schemes remain coarse, which hinders their ability to support finer-grained urban research in the future. To bridge the gap, we constructed a novel multimodal and fine-grained dataset—BuildingSense—for building function classification, comprising over 34,000 buildings, 60,000 annotated images, 71,654 POIs, and 34,00 building description texts in 26 distinct categories. Based on BuildingSense, we evaluated the performance of four state-of-the-art large models from the perspective of classification outcomes and reasoning processes. The results demonstrate that large models can effectively comprehend multimodal spatial data, challenging the conventional concept. Based on that, three directions for future research can be key: (1) build a categorized inference example database, (2) develop cost-effective classification models, and (3) quantify the confidence of model outputs. Our findings not only provide insights into the development of subsequent large model-based classification methods but also contribute to the advancement of multimodal fusion-based classification methods. The dataset and code of this paper can be accessed through https://doi.org/10.6084/m9.figshare.30645776.v2 [1].

Referee Comment 2.2

I noticed that the street-view images in the published dataset do not have the target buildings annotated. This will undoubtedly lead to "Spatial relationship errors" in the current study and similar issues in future models trained on this dataset. Could the authors explain the reasoning behind this?

Reply:

Thank you for your insightful comment. It made us realize that we had not clearly articulated the rationale for the absence of street view annotation in the original manuscript. Therefore, we have added a subsubsection titled "Lack of annotation in street view images" in the Discussion section to elaborate on our reasoning—specifically, to test whether the model could automatically infer the location of the target building in street view images based on the other provided information.

Changes:

6.3.2 Lack of annotation in street view images

In BuildingSense, the street view images are not annotated with the target building. It was designed to assess whether large models can infer the target building's location within a street-view image from other provided information (e.g., remote-sensing imagery, building height). In our prompting templates, we only indicate to the model that the target building is located directly ahead. Strikingly, we found that the model attempts to match and integrate across

the different viewpoints and is capable of self-correction when it detects logical inconsistencies, ultimately producing a revised—and correct—label. An illustrative example is provided in Figure 1.

Figure 1: Example of the Gemini-flash 2.5 (Thinking)'s output

It can be observed that the model's textual description of the street view image did not match the actual characteristics of the target building: the model mistakenly took a taller background tower as the target structure (see the blue text in Figure 1, "Description of building in Street view image"). Despite the condition arising from the unannotated target building in the street view image, this setup enables testing whether the model can align the building's location across the remote sensing and street view modalities using the provided information. In a subsequent inference, the section "Semantic information of the building in the content "in Figure 1 mentions that given the building height that we supplied, it recognized a recognition error in the street view with textual cues present in the street image and revised its street view judgment. The later stages of the inference retained this revised decision consistently (see the blue text in Figure 1, "Analyze each POI name of the building" and "Combined the analysis and judgment").

This annotation gap was intentionally designed to test whether large models could integrate multimodal spatial data and perform coherent reasoning—a finding that ultimately proves our assumptions. However, this approach inevitably increases annotation costs in practical applications, as target buildings in streetview images must be labeled to support specific tasks. Therefore, we plan to include target building annotations in future updates of the dataset.

Referee Comment 2.3

> There is a typo in Equation (1): "arctan" should replace "acrtan"

Reply:

> Thank you for pointing out the spelling errors in our manuscript. We apologize for overlooking these mistakes. In response to your comment, we have not only corrected the spelling error in Equation (1) and (2a).

Changes:

$$Pitch = \sout{acrtan}arctan\left(\frac{h}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}\right) \tag{1}$$

$$\theta = \sout{acrtan}arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \tag{2a}$$

$$Heading = \begin{cases} 90 - \theta, & \textbf{if } x_2 - x_1 > 0 \\ 270 + \theta, & \textbf{if } x_2 - x_1 \leq 0 \end{cases} \tag{2b}$$

Referee Comment 2.4

> In the POI section, adding a field table would improve readability.

Reply:

> Thank you for your constructive comment. We have added a detailed field table in the POI section.

Changes:

Table 1: POI sample data

Field	Value
Index	1
Place id	ChIJ2SZR_QdZwokRT5cHeU0dKUo
Name	SBFI Group
Latitude	40.745241
Longitude	-73.9824725
Types	[furniture_store, home_improvement_store, home_goods_store, store]
Primary type	furniture_store
Current opening hours	9:00–17:00

Referee Comment 2.5

There are spelling inconsistencies for "BuildingSense" throughout the manuscript (e.g., line 165).

Reply:

Thank you for pointing out the spelling errors in our manuscript. We apologize for overlooking these mistakes. In response to your comment, we have not only corrected the spelling error at line 165 but also carefully reviewed and revised other similar errors throughout the text.

Changes:

The building footprints collected in ~~BuildingSence~~ BuildingSense are sampled from the official building footprint data published by NYC Open Data. To avoid the geographical bias and imbalance in the categories of samples, the sampling process adhered strictly to two principles: (1) the sampled buildings should be spatially evenly distributed, and (2) the distribution of building functional categories after sampling should be uniform. Based on principle (1), we utilize NYC taxi zones (spatial distribution shown in Supplementary Fig.S5) as the spatial sampling units and assign each building to a corresponding zone number via spatial computations. Regarding principle (2), we reclassify NYC's building functional categories (details on category mappings and definitions are provided in Supplementary Table S1).

The POI data of buildings in ~~BuildingSence~~ BuildingSense are collected from Google Maps. Given the constraints imposed by API return limits, we employed a comprehensive deep-search approach to extract POI information within the boundary of the building footprint. The detailed procedure for this algorithm is shown in the Supplementary Algorithm S1. The collected POI data comprises several attributes: place id, name, latitude, longitude, types, primary type, and current opening hours. Specifically, 'types' and 'primary type' represent the diverse and primary functions of the POIs, respectively, and 'current opening hours' facilitates the future studies of dynamic building functions classification.

Referee Comment 2.6

While large models have surpassed traditional methods in reasoning on complex tasks and can significantly enhance the interpretability of results, it is still necessary to explain why traditional deep learning models were not evaluated on this building function dataset.

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. A detailed illustration is shown below.

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 [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] and (2) poor interpretability of model outputs [8, 9, 10, 11, 12]. 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) [13, 14]. 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 2.7

Please add crucial information about "residential category statistics scaled by a factor of 10" in the title of Figure 7 to prevent misinterpretation.

Reply:

Thank you for pointing out the insufficient illustration in Figure 7. In response to your comment, we have added relevant illustration in the caption of Figure 7.

Changes:

Third, Figure 2 presents a data completeness analysis of BuildingSense ~~(residential category statistics scaled by a factor of 10)~~. It can be seen that most of the buildings include remote sensing imagery, street views, and attribute annotations (Figure 2). However, three functional categories, Entertainment, Public service, and Fundamental infrastructure, exhibit relatively poorer data completeness, primarily due to missing street view imagery. This phenomenon stems from the inaccessible areas of these buildings located beyond the street view vehicles (e.g., secured facilities or locations far from roads).

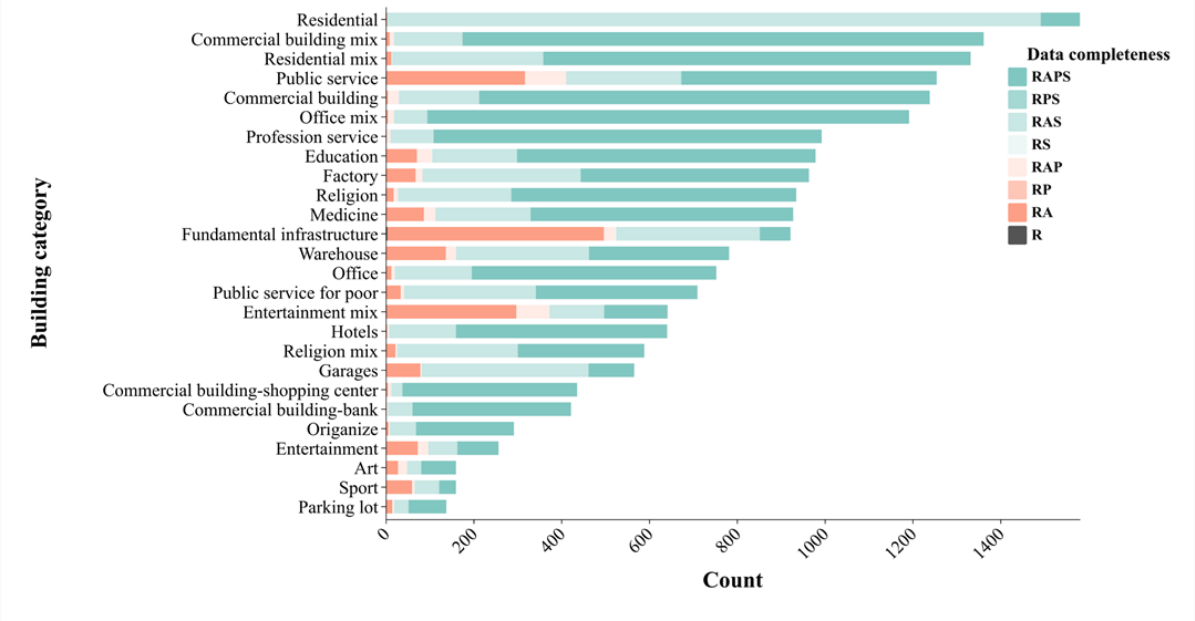

Figure 2: Data completeness of each category. R refers to remote sensing imagery, A refers to annotated building attributes (building height, location, and constructed year), S refers to street view imagery, and P refers to POI data. Residential category statistics scaled by a factor of 10.

References

[1] Pengxiang Su, Runfei Chen, Heng Xu, Wei Huang, Xinling Deng, Wanglin Yan, et al. BuildingSense-A multimodal building function classification dataset, 2025.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.