Response to Reviewer 2

Manuscript: *BMAT: A footprint-level building facade material dataset for 73 major cities worldwide*

We sincerely thank the reviewer for the careful reading of our manuscript and for the constructive comments. We have carefully considered each point and revised the manuscript accordingly. Below, we respond to the comments individually. The reviewer's comments are shown in blue, and revised manuscript text is quoted in *gray italics*.

General comment

> **Comment General.** *This manuscript presents BMAT, a building facade material dataset that combines large-scale geographic coverage with individual-building-level precision. Overall, the manuscript is well organized and clearly written, and the experiments and analyses are generally rigorous and comprehensive. However, I still have several concerns that I encourage the authors to carefully consider.*

Response:

Thank you for your careful evaluation of our manuscript and for these constructive suggestions. We have carefully addressed all six comments and revised the manuscript accordingly. Detailed responses are provided below.

Specific comments

Comment 1. *Lines 94–96 describe the factors considered in selecting the 73 cities, including population size, administrative status, and economic importance. However, the manuscript does not provide the specific reference criteria or thresholds used for these factors.*

Response:

Thank you for this helpful suggestion. We did not apply fixed quantitative thresholds for city selection. Instead, the 73 cities were purposively selected to achieve broad geographic coverage, with priority given to national or regional capitals, major administrative and economic centers, and large population centers across the included countries and regions. We have clarified this selection rationale in the revised manuscript, and the characteristics of the selected cities are summarized in Table 1. We have also added a detailed description of the city selection rationale in Supplementary Section S2.

Table 1: Cities included in BMAT and their main selection characteristics.

Country	City	Platform	Continent	Selection rationale	Country	City	Platform	Continent	Selection rationale
Argentina	Buenos Aires	GSV	South America	National capital; major metro	China	Beijing	BSV	Asia	National capital
Australia	Sydney	GSV	Oceania	State capital; economic hub	China	Changchun	BSV	Asia	Provincial capital
Brazil	Rio de Janeiro	GSV	South America	State capital; major metro	China	Changsha	BSV	Asia	Provincial capital
Canada	Ottawa	GSV	North America	National capital	China	Chengdu	BSV	Asia	Provincial capital; regional hub
Canada	Toronto	GSV	North America	Provincial capital; economic hub	China	Chongqing	BSV	Asia	Municipality; major metro
Chile	Santiago	GSV	South America	National capital; major metro	China	Dalian	BSV	Asia	Port and economic hub
China	Hong Kong	GSV	Asia	SAR; economic hub	China	Fuzhou	BSV	Asia	Provincial capital
China	Taipei	GSV	Asia	Regional capital; major metro	China	Guangzhou	BSV	Asia	Provincial capital; economic hub
Colombia	Bogota	GSV	South America	National capital; major metro	China	Guiyang	BSV	Asia	Provincial capital
France	Paris	GSV	Europe	National capital; major metro	China	Harbin	BSV	Asia	Provincial capital
Germany	Berlin	GSV	Europe	National capital	China	Haikou	BSV	Asia	Provincial capital
India	Bangalore	GSV	Asia	State capital; technology hub	China	Hangzhou	BSV	Asia	Provincial capital; economic hub
India	Delhi	GSV	Asia	National administrative center	China	Hefei	BSV	Asia	Provincial capital
India	Mumbai	GSV	Asia	State capital; financial hub	China	Hohhot	BSV	Asia	Autonomous-region capital
Indonesia	Jakarta	GSV	Asia	National capital; major metro	China	Jinan	BSV	Asia	Provincial capital
Italy	Rome	GSV	Europe	National capital	China	Lanzhou	BSV	Asia	Provincial capital
Japan	Tokyo	GSV	Asia	National capital; major metro	China	Lhasa	BSV	Asia	Autonomous-region capital
Kenya	Nairobi	GSV	Africa	National capital; economic hub	China	Nanchang	BSV	Asia	Provincial capital
Mexico	Mexico City	GSV	North America	National capital; major metro	China	Nanjing	BSV	Asia	Provincial capital
Netherlands	Amsterdam	GSV	Europe	National capital; economic hub	China	Nanning	BSV	Asia	Autonomous-region capital
Nigeria	Lagos	GSV	Africa	Population and economic hub	China	Ningbo	BSV	Asia	Port and economic hub
Russia	Moscow	GSV	Europe	National capital; major metro	China	Qingdao	BSV	Asia	Port and economic hub
Senegal	Dakar	GSV	Africa	National capital; regional hub	China	Shanghai	BSV	Asia	Municipality; economic hub
Singapore	Singapore	GSV	Asia	National capital; economic hub	China	Shenyang	BSV	Asia	Provincial capital
South Africa	Cape Town	GSV	Africa	Legislative and provincial capital	China	Shenzhen	BSV	Asia	Technology and economic hub
South Korea	Seoul	GSV	Asia	National capital; major metro	China	Shijiazhuang	BSV	Asia	Provincial capital
Spain	Barcelona	GSV	Europe	Regional capital; economic hub	China	Suzhou	BSV	Asia	Manufacturing and economic hub
Thailand	Bangkok	GSV	Asia	National capital; major metro	China	Taiyuan	BSV	Asia	Provincial capital
Turkey	Istanbul	GSV	Europe/Asia	Population and economic hub	China	Tianjin	BSV	Asia	Municipality; port hub
UAE	Dubai	GSV	Asia	Emirate capital; economic hub	China	Urumqi	BSV	Asia	Autonomous-region capital
UK	London	GSV	Europe	National capital; financial hub	China	Wuhan	BSV	Asia	Provincial capital; regional hub
US	Boston	GSV	North America	State capital; major metro	China	Wuxi	BSV	Asia	Manufacturing and economic hub
US	Chicago	GSV	North America	Population and economic hub	China	Xiamen	BSV	Asia	Port and economic hub
US	Los Angeles	GSV	North America	Population and economic hub	China	Xi'an	BSV	Asia	Provincial capital; regional hub
US	New York	GSV	North America	Population and financial hub	China	Xining	BSV	Asia	Provincial capital
US	San Francisco	GSV	North America	Technology and economic hub	China	Yinchuan	BSV	Asia	Autonomous-region capital
					China	Zhengzhou	BSV	Asia	Provincial capital; regional hub

Note. GSV = Google Street View; BSV = Baidu Street View. Selection rationale summarizes the main administrative, demographic, or economic characteristic of each city.

Comment 2. *Given the importance of this dataset to the proposed framework, the authors should briefly describe the annotation process, including image sampling, annotator expertise, and the treatment of ambiguous or mixed-material facades.*

Response:

Thank you for this helpful suggestion. We have added a concise description of the manual annotation process, including image selection, annotator expertise, and the treatment of ambiguous and mixed-material facades. Specifically, 38,873 images were sampled from the street-view imagery collected in this study, and 532 additional images were obtained from previously published annotated resources. To ensure consistent application of the material definitions, images were labeled over one month by a doctoral researcher in urban planning with experience in urban built environments and street-view image analysis. Mixed-material facades were assigned the material occupying the largest visible proportion of the main facade. The annotation records produced in this study are publicly available at https://github.com/yhyJoy/BMAT/blob/main/data/label/image_label.csv, while the labels compiled from previously published resources are provided separately at https://github.com/yhyJoy/BMAT/blob/main/data/label/image_label_otherSource.csv. To the best of our knowledge, this is the largest publicly available manually annotated facade-material image collection based on street-view imagery, comprising 39,405 labeled images. The annotation collection itself therefore provides an additional reusable resource for future facade-material research.

Table 2: Comparison of recent street-view-based facade material resources.

Study	Year	Cities	Labels	Classes	Open	Main output
Raghu et al. (2023)	2023	3	985	7	Image data	Annotated images and material detection model
Wang et al. (2024b)	2024	2	2,567	6	Image data	Facade image dataset for cladding classification
Chen et al. (2025)	2025	8	17,914	9	No	Training data and city-scale material mapping
Sun et al. (2025)	2025	4	11,787	9	No	Roof and facade material identification
Liang et al. (2025)	2025	7	2,871	7	Toolkit	OpenFACADES attribute-enrichment workflow
BMAT (ours)	2026	73	39,405	9	Dataset	Footprint-level material records for 22.09 million buildings

Note: Labels refer to the number of annotated facade or material samples used for material classification or evaluation. Classes refer to material categories. Open indicates the type of publicly available resource reported by each study.

Changes in the manuscript:

Manual annotation of facade images

We assembled 39,405 manually annotated building facade images. The collection comprised 38,873 images sampled from the street-view imagery acquired in this study and 532 images obtained from previously published annotated resources (Supplementary Section S5). To ensure consistent

application of the material definitions, all images were annotated over one month by a doctoral researcher in urban planning with research experience in urban built environments and street-view image analysis. Following previous studies, facade materials were classified into nine categories (Figure 3): brick (5,502 images), concrete (5,490), glass (4,804), metal (2,831), stone (4,773), stucco (7,359), tile (2,067), wood (5,521), and other (1,058). Common materials were represented by larger sample sizes, while all nine categories retained sufficient observations for model training. The other category mainly included buildings under construction or renovation, as well as facades that could not be reliably assigned to the main categories. For mixed-material facades, the label was assigned according to the material occupying the largest visible proportion of the main facade. For example, a building with a concrete ground floor but tile cladding across most of its upper floors was labeled as tile.

Comment 3. *The material-inference prompt refers to the target as the "central building." However, whether the building located at the image center is unique and corresponds to the intended building footprint depends directly on the image FOV, zoom level, and camera-to-building distance. The manuscript currently reports only the heading and pitch parameters, without explaining how the FOV or zoom level was determined.*

Response:

Thank you very much for raising this important point. We agree that, in addition to heading and pitch, the field of view (FOV) is a key parameter in street-view image collection. It determines both the extent of the scene captured and the apparent size of the target building within the image.

We did not use a fixed FOV for all buildings. Instead, we adjusted the FOV according to building height and the distance between the street-view location and the building. We first used these two variables to estimate the angle (α) from the camera to the top of the building. This angle represents the apparent size of the building from the selected viewpoint. Taller or closer buildings have a larger α and therefore require a wider FOV to reduce the risk of cropping the roof, lower facade, or building edges. Shorter or more distant buildings have a smaller α, allowing a narrower FOV to enlarge the building and preserve more facade detail.

We then calculated the FOV as the ratio of α to a target proportion parameter. During preliminary testing, we compared images generated using different parameter values. We found that a value of 0.2 provided sufficient space around the building and more consistently captured the full facade. However, an FOV that is too narrow may capture only part of the building, whereas an FOV that is too wide may include too many neighboring buildings and unrelated background features. We therefore constrained the FOV to a range of 40° to 120°. Values outside this range were set to the nearest limit. When valid building-height information was unavailable, we used a default FOV of 90°.

We acknowledge that the FOV cannot always isolate the exact boundary of the target building. As a result, a small number of images may also contain neighboring buildings. However, the heading was calculated separately for each image using the street-view sampling point and the centroid of the predefined target building. This placed the target building near the image center, even when nearby buildings were also visible. Therefore, our model prompt also explicitly instructed the model to focus on the building at the center of the image.

All images were generated at a fixed resolution of 640×640 pixels. At this fixed resolution, the FOV also controlled the effective framing and scale of the image. We therefore did not specify a separate zoom level. We have now added the detailed calculation, parameter settings, limits, and default values to the revised main text and Supplementary Methods.

Changes in the manuscript:

"This study aims to acquire time-series street-view imagery at the individual-building scale. We developed a geometric workflow that links building footprints to the road network. For each building, we first identified a theoretical sampling location by projecting its geometric centroid onto the nearest road segment. Because this location may not coincide with an available street-view panorama, the nearest recorded panorama was used as the actual sampling point. Camera parameters, including heading, pitch, and field of view (FOV), were then determined from the spatial relationship between the actual sampling point and the target building (Figure 2(a)). This parameterized workflow enabled automated, large-scale collection of building-centered imagery across entire cities.

The heading was defined as the clockwise angle from true north to the line connecting the actual sampling point and the building centroid, thereby directing the camera toward the target building. The pitch represented the vertical angle toward the midpoint of the building facade, estimated from half of the building height. For buildings without valid height information in Overture Maps, a default pitch of 30° *was used. The FOV was adjusted according to building height and the distance from the panorama to the building (Supplementary Section S3). Taller or closer buildings were assigned a wider FOV to reduce the risk of cropping, whereas shorter or more distant buildings were assigned a narrower FOV to preserve facade detail. All images were generated at a fixed resolution of* 640×640 *pixels."*

Changes in the supplementary materials:

Automated adjustment of the field of view for street-view image acquisition

A fixed field of view (FOV) is not suitable for all buildings because their heights and distances from the street-view camera vary substantially. Taller or closer buildings require a wider FOV to avoid cropping the roof, ground floor, or building edges. Shorter or more distant buildings can be captured with a narrower FOV, which increases the apparent size of the facade and preserves more visible detail. We therefore calculated the FOV separately for each building.

First, we calculated the distance d between the actual street-view sampling point and the building centroid from their WGS84 coordinates using the Haversine formula:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right),$$

$$d = 2R \arctan 2\left(\sqrt{a}, \sqrt{1-a}\right),$$

where ϕ_1 *and* ϕ_2 *are the latitudes of the street-view location and building centroid, respectively;* $\Delta\phi$ *and* $\Delta\lambda$ *are their latitude and longitude differences; and* $R = 6{,}371{,}000$ *m is the mean Earth radius.*

We then estimated the viewing angle from the camera to the top of the building:

$$\alpha = \arctan\left(\frac{H - h_c}{d}\right),$$

where H is the building height and $h_c = 2$ *m is the approximate camera height. The angle* α *was*

converted to degrees and used as an estimate of the apparent vertical extent of the building from the selected viewpoint.

The initial FOV was calculated as:

$$FOV_{\text{raw}} = \frac{\alpha}{r},$$

where r is a target angular proportion parameter. During preliminary testing, we compared images generated using different values of r. We found that $r = 0.2$ more consistently retained the complete building facade while leaving sufficient space around the roof, ground floor, and lateral edges.

A very narrow FOV may capture only part of the building, whereas a very wide FOV may include too many neighboring buildings and unrelated background features. We therefore constrained the final FOV to between $40°$ and $120°$. Values outside this range were set to the nearest boundary. When valid building-height information was unavailable, a default FOV of $90°$ was used.

$$FOV = \min\left[\max\left(FOV_{\text{raw}}, 40°\right), 120°\right].$$

Comment 4. *The temporal matching strategy adopted in Section 4.1 is essentially disjunctive: the more timestamps available for a building, the greater the probability of obtaining a match, which may systematically overestimate the agreement rate. Accordingly, the reported accuracy of 0.80 should be interpreted as an upper bound rather than a best estimate, with the true accuracy likely falling between 0.72 and 0.80. Where feasible, the authors are encouraged to report accuracy stratified by the number of available observation years per building, or at least revise the corresponding statements to explicitly acknowledge this limitation.*

Response:

Thank you for this important suggestion. We agree that buildings with more annual records have more opportunities to obtain a historical match. We have revised the manuscript to present the latest-record agreement of 0.72 as the primary result and the historical-match agreement of 0.79 as an upper-bound sensitivity estimate.

We extracted the "facade_material" attribute from Overture Maps buildings in the cities covered by BMAT and harmonized the material categories between the two sources, yielding 50,823 comparable records. Comparing each Overture Maps record with the latest BMAT material produced an agreement of 0.719 (36,542/50,823). Because Overture Maps does not report the observation date of this attribute, we also applied historical matching, in which a building was considered matched if any annual BMAT record agreed with Overture Maps. This produced an agreement of 0.793 (40,296/50,823; Figure 1).

We further reported the latest-record and historical-match agreements by the number of BMAT observation years per building (Table 3). Among the 50,823 buildings, 39,985 had records from at least two years. The historical-match gain was 0.065 for buildings with two observation years, approximately 0.094-0.097 for those with three to five years, 0.114 for those with six to nine years, and 0.078 for those with ten or more years.

In addition, we examined which material categories contributed to the historical-match gain. We identified 3,754 buildings whose latest BMAT record did not match Overture Maps but whose historical record did. Figure 2 shows the distribution of observation years among these cases, including 941 buildings with ten or more annual records. The material correspondence matrix shows that the gain was concentrated mainly among brick, concrete, and stucco (Figure 3). Differences between concrete and stucco may reflect different category definitions: BMAT labels facades with a visible cement-rendered or plastered surface as stucco, whereas Overture Maps may retain concrete as the underlying wall material. Differences between brick and concrete or stucco may also reflect facade renovation or errors in either source.

Finally, we reported agreement by city and material category, together with the corresponding sample sizes (Table 4). Concrete showed relatively low agreement, largely because BMAT and Overture Maps may apply different boundaries between concrete and stucco.

Figure 1: Comparison of model predictions with Overture Maps records for (a) latest and (b) historical material attributes, including (c) visual examples with green for correct and red for error.

Table 3: Overture–BMAT agreement stratified by the number of BMAT observation years.

BMAT observation years	*N*	**Latest agreement**	**Historical agreement**	**Gain**
1	10,838	0.683 (7,397/10,838)	0.683 (7,397/10,838)	0.000
2	4,321	0.648 (2,801/4,321)	0.713 (3,081/4,321)	0.065
3	3,028	0.622 (1,884/3,028)	0.716 (2,168/3,028)	0.094
4	2,804	0.678 (1,901/2,804)	0.774 (2,169/2,804)	0.096
5	2,523	0.664 (1,675/2,523)	0.761 (1,919/2,523)	0.097
6–9	15,193	0.702 (10,660/15,193)	0.816 (12,397/15,193)	0.114
≥10	12,116	0.844 (10,224/12,116)	0.922 (11,165/12,116)	0.078
Overall	**50,823**	**0.719 (36,542/50,823)**	**0.793 (40,296/50,823)**	**0.074**

Note. BMAT observation years denote distinct annual material records per building. Latest agreement uses the latest BMAT label; historical agreement allows matching with any annual label. Gain is their difference.

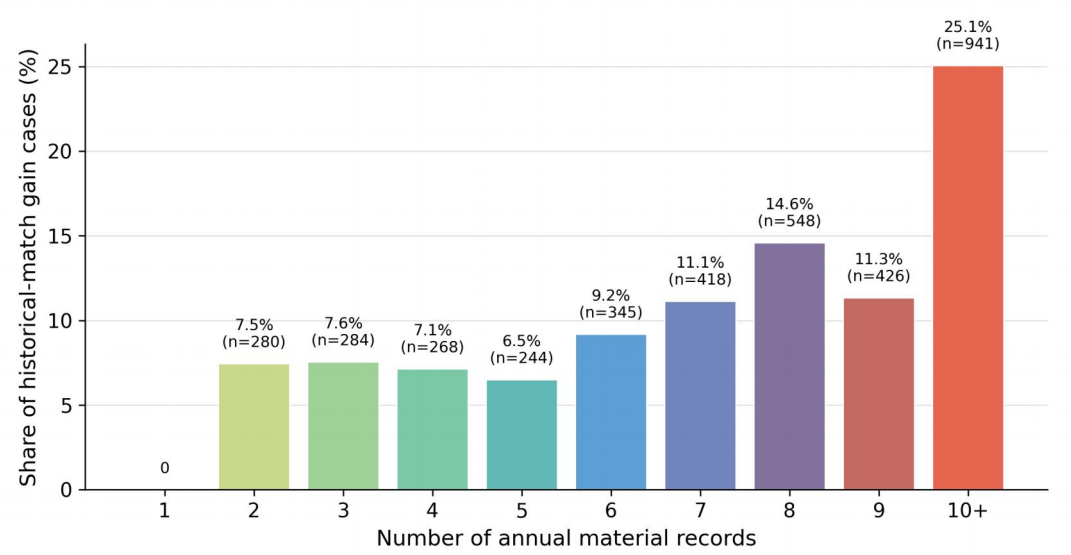

Figure 2: Number of annual material records in historical-match gain cases. Gain cases refer to buildings unmatched by the latest BMAT record but matched by at least one historical BMAT record.

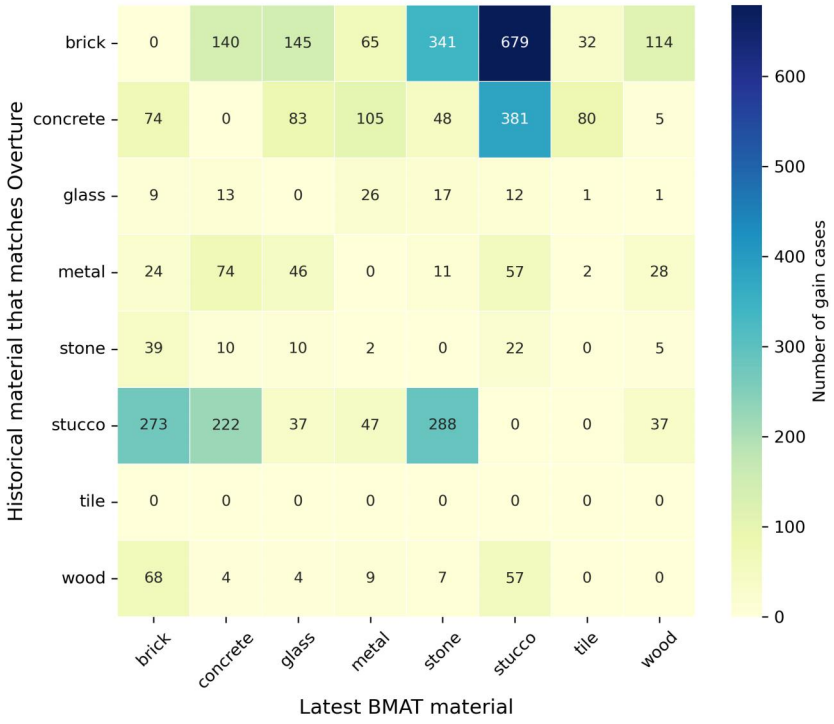

Figure 3: Material correspondence in historical-match gain cases. Rows indicate the historical BMAT material that matches Overture Maps, and columns indicate the latest BMAT material.

Table 4: Agreement between Overture Maps and the latest BMAT material labels.

City	*N*	Overall	Brick	Concrete	Glass	Metal	Stone	Stucco	Wood
Overall	**50,823**	**0.72 (36542/50823)**	**0.88 (26381/30031)**	**0.20 (1379/6929)**	**0.79 (1271/1612)**	**0.49 (1034/2100)**	**0.61 (700/1147)**	**0.67 (5316/7965)**	**0.44 (461/1039)**
London	16,510	0.90 (14924/16510)	0.93 (13462/14404)	0.23 (30/131)	0.80 (76/95)	0.41 (17/41)	0.78 (335/431)	0.73 (990/1365)	0.33 (14/43)
Moscow	12,675	0.61 (7775/12675)	0.75 (3423/4557)	0.33 (537/1641)	0.63 (112/179)	0.51 (892/1753)	0.24 (10/41)	0.62 (2679/4329)	0.70 (122/175)
Toronto	7,165	0.83 (5958/7165)	0.88 (5316/6071)	0.27 (41/150)	0.81 (127/156)	0.48 (13/27)	0.75 (203/269)	0.54 (242/451)	0.39 (16/41)
Tokyo	2,379	0.08 (186/2379)	0.09 (4/45)	0.05 (107/2215)	0.69 (43/62)	0.62 (23/37)	0.22 (2/9)	0.67 (2/3)	0.62 (5/8)
Amsterdam	1,979	0.66 (1302/1979)	0.98 (1028/1050)	0.34 (169/497)	0.55 (6/11)	0.07 (1/15)	–	–	0.24 (98/406)
Berlin	1,894	0.80 (1517/1894)	0.87 (274/314)	0.84 (256/305)	0.59 (27/46)	0.56 (10/18)	0.68 (15/22)	0.79 (929/1172)	0.35 (6/17)
New York	1,296	0.84 (1087/1296)	0.86 (937/1090)	0.32 (6/19)	0.90 (113/126)	0.56 (9/16)	0.83 (15/18)	0.24 (6/25)	0.50 (1/2)
Sydney	1,147	0.67 (774/1147)	0.75 (550/730)	0.18 (12/68)	0.58 (14/24)	0.22 (2/9)	0.80 (12/15)	0.66 (170/257)	0.32 (14/44)
Singapore	820	0.11 (91/820)	0.04 (4/94)	0.02 (13/582)	0.67 (33/49)	0.44 (24/55)	0.20 (5/25)	0.92 (12/13)	0.00 (0/2)
Istanbul	760	0.15 (115/760)	0.00 (0/2)	0.13 (90/711)	0.85 (11/13)	0.67 (2/3)	0.44 (4/9)	0.44 (4/9)	0.31 (4/13)
Boston	577	0.86 (495/577)	0.93 (444/479)	0.53 (17/32)	0.53 (10/19)	0.50 (3/6)	0.51 (19/37)	0.00 (0/1)	0.67 (2/3)
Bogota	504	0.84 (422/504)	0.92 (334/362)	0.25 (6/24)	0.81 (72/89)	0.24 (4/17)	1.00 (1/1)	0.57 (4/7)	0.25 (1/4)
Paris	346	0.70 (243/346)	0.88 (168/192)	0.42 (8/19)	0.50 (31/62)	0.50 (5/10)	0.48 (26/54)	–	0.56 (5/9)
Shenzhen	314	0.81 (253/314)	0.00 (0/1)	0.04 (1/25)	0.93 (249/267)	0.17 (3/18)	0.00 (0/1)	0.00 (0/2)	–
Bangalore	307	0.75 (229/307)	0.50 (9/18)	0.08 (2/26)	0.76 (26/34)	0.00 (0/3)	0.30 (7/23)	0.93 (185/200)	0.00 (0/3)
Ottawa	296	0.86 (254/296)	0.89 (173/194)	0.38 (5/13)	0.86 (6/7)	0.80 (4/5)	0.91 (21/23)	0.29 (2/7)	0.91 (43/47)
San Francisco	227	0.63 (142/227)	0.75 (12/16)	0.25 (1/4)	0.75 (6/8)	0.00 (0/1)	–	0.50 (1/2)	0.62 (122/196)
Mumbai	211	0.16 (34/211)	–	0.15 (28/193)	–	0.17 (2/12)	–	0.80 (4/5)	0.00 (0/1)
Buenos Aires	188	0.74 (140/188)	0.84 (59/70)	0.33 (7/21)	0.84 (73/87)	0.10 (1/10)	–	–	–
Rome	152	0.27 (41/152)	0.56 (5/9)	0.11 (2/19)	0.67 (4/6)	0.61 (11/18)	0.13 (11/87)	0.67 (8/12)	0.00 (0/1)
Chicago	150	0.70 (105/150)	0.86 (85/99)	0.11 (2/19)	0.73 (11/15)	0.33 (1/3)	0.20 (1/5)	0.00 (0/1)	0.62 (5/8)
Santiago	124	0.57 (71/124)	0.46 (33/71)	0.33 (6/18)	0.94 (30/32)	–	1.00 (1/1)	1.00 (1/1)	0.00 (0/1)
Hong Kong	95	0.77 (73/95)	1.00 (1/1)	0.19 (4/21)	0.94 (67/71)	1.00 (1/1)	0.00 (0/1)	–	–
Delhi	93	0.25 (23/93)	0.24 (12/50)	0.11 (3/27)	–	0.33 (2/6)	1.00 (4/4)	0.40 (2/5)	0.00 (0/1)
Guangzhou	80	0.23 (18/80)	0.25 (1/4)	0.10 (6/59)	0.85 (11/13)	0.00 (0/4)	–	–	–
Shanghai	69	0.84 (58/69)	0.00 (0/1)	0.00 (0/1)	0.75 (9/12)	0.00 (0/1)	1.00 (1/1)	0.91 (48/53)	–
Beijing	68	0.46 (31/68)	0.70 (19/27)	0.00 (0/3)	–	0.00 (0/1)	0.20 (3/15)	0.50 (9/18)	0.00 (0/4)
Cape Town	58	0.09 (5/58)	0.30 (3/10)	0.00 (0/8)	0.00 (0/2)	–	0.00 (0/35)	1.00 (2/2)	0.00 (0/1)
Seoul	55	0.56 (31/55)	0.67 (8/12)	0.18 (3/17)	0.86 (19/22)	–	0.00 (0/3)	–	1.00 (1/1)
Mexico City	55	0.55 (30/55)	0.25 (1/4)	0.48 (10/21)	0.79 (15/19)	0.40 (2/5)	0.40 (2/5)	–	0.00 (0/1)
Jakarta	53	0.38 (20/53)	0.00 (0/28)	0.00 (0/4)	0.95 (20/21)	–	–	–	–
Los Angeles	50	0.58 (29/50)	0.62 (5/8)	0.20 (1/5)	0.64 (9/14)	0.50 (1/2)	0.00 (0/1)	0.72 (13/18)	0.00 (0/2)
Barcelona	49	0.59 (29/49)	0.69 (11/16)	0.00 (0/3)	0.93 (13/14)	1.00 (1/1)	0.14 (1/7)	0.50 (3/6)	0.00 (0/2)
Ningbo	24	0.58 (14/24)	–	0.10 (1/10)	0.93 (13/14)	–	–	–	–
Nairobi	10	0.40 (4/10)	–	0.00 (0/3)	1.00 (2/2)	–	0.00 (0/3)	–	1.00 (2/2)
Taipei	9	0.33 (3/9)	–	–	0.38 (3/8)	–	–	–	0.00 (0/1)
Dubai	7	0.71 (5/7)	–	0.00 (0/1)	1.00 (5/5)	0.00 (0/1)	–	–	–
Dakar	6	0.50 (3/6)	–	0.50 (3/6)	–	–	–	–	–
Chengdu	3	1.00 (3/3)	–	–	1.00 (3/3)	–	–	–	–
Hohhot	3	0.00 (0/3)	0.00 (0/1)	–	0.00 (0/2)	–	–	–	–
Lanzhou	3	0.67 (2/3)	–	–	0.67 (2/3)	–	–	–	–
Lagos	2	0.00 (0/2)	–	0.00 (0/2)	–	–	–	–	–
Rio de Janeiro	2	0.50 (1/2)	–	0.50 (1/2)	–	–	–	–	–
Bangkok	2	0.00 (0/2)	0.00 (0/1)	0.00 (0/1)	–	–	–	–	–
Nanjing	2	0.50 (1/2)	–	1.00 (1/1)	–	0.00 (0/1)	–	–	–
Harbin	1	1.00 (1/1)	–	–	–	–	1.00 (1/1)	–	–
Taiyuan	1	0.00 (0/1)	–	0.00 (0/1)	–	–	–	–	–
Tianjin	1	0.00 (0/1)	–	0.00 (0/1)	–	–	–	–	–
Changsha	1	0.00 (0/1)	–	–	–	–	–	0.00 (0/1)	–

Note. Entries report agreement as ratio (matched/comparable). Dashes indicate no comparable records. Tile is omitted because no comparable Overture reference records were available.

Changes in the abstract:

> *"Dataset reliability was further assessed through comparison with 50,823 Overture Maps facade-material records, yielding agreements of 0.72 for the latest BMAT records and 0.79 under a historical matching strategy."*

Changes in the results:

"Because Overture Maps material records do not contain acquisition dates, two comparison strategies were used. First, matching the latest BMAT material record with the Overture Maps record yielded an agreement of 0.72 (Figure 5(a), Supplementary Table S4). This comparison is transparent and conservative, but it may underestimate agreement if the undated Overture record corresponds to an earlier facade state rather than the latest BMAT observation. Second, we used a historical matching strategy in which a building was considered matched if any available annual BMAT material record agreed with the Overture Maps record, yielding an agreement of 0.79 (Figure 5(b)). This value should be interpreted as an upper-bound sensitivity estimate rather than a conventional accuracy estimate, because buildings with more annual observations have more opportunities to produce a match by chance."

Comment 5. *In Section 4.1, Qwen2.5-VL-7B appears to have been selected as the sole model for fine-tuning based only on comparisons among models in their non-fine-tuned states. However, zero-shot or untuned performance does not necessarily predict the performance ceiling after supervised fine-tuning. I therefore recommend fine-tuning at least two or three representative vision-language models using the same training data, test set, training protocol, and evaluation metrics. This would provide a more convincing basis for model selection.*

Response:

Thanks for your valuable suggestion. We agree that zero-shot or unfine-tuned performance alone may not reliably indicate the performance ceiling after supervised fine-tuning. To address this concern, we additionally fine-tuned four representative vision-language models, including InternVL2.5-8B, MiniCPM-V-2.6, Qwen2-VL-7B-Instruct, and mLLaMA3.2-11B, using the same training set (27,578 images), test set (11,827 images), training protocol, and evaluation metrics. Specifically, all models were fine-tuned under a consistent supervised fine-tuning setting with the same input prompt, label space, number of training epochs, learning rate schedule, LoRA-based adaptation strategy, and evaluation procedure.

The new results show that supervised fine-tuning substantially improved all tested VLMs. Nevertheless, the fine-tuned Qwen2.5-VL-7B-Instruct still achieved the best overall balance between classification performance and inference efficiency, with an F1 score of 0.91 and an inference time of 237.73 ms. These results support our selection of Qwen2.5-VL-7B-Instruct as the final backbone. We have revised Section 4.1 accordingly and updated the model comparison results in Table 1 and Supplementary Table S2.

Table 5: Comprehensive performance comparison of different models.

Model	**Fine-tuned**	**Inference time (ms)**	**Accuracy**	**Precision**	**Recall**	**F1 score**
Traditional vision backbones						
VGG16	✓	1.26	0.83	0.82	0.80	0.81
ViT-B/16	✓	4.40	0.82	0.81	0.79	0.80
ResNet-50	✓	5.57	0.82	0.80	0.80	0.80
ResNet-101	✓	10.52	0.82	0.81	0.80	0.80
ShuffleNetV2	✓	5.57	0.79	0.78	0.76	0.77
EfficientNet-B0	✓	6.81	0.84	0.82	0.82	0.82
DenseNet-161	✓	18.37	0.84	0.83	0.82	0.82
MobileNetV3-Large	✓	5.54	0.82	0.81	0.80	0.80
Vision-language models (VLMs)						
InternVL2.5-8B	×	185.55	0.69	0.62	0.53	0.54
MiniCPM-V-2.6	×	631.59	0.74	0.55	0.48	0.48
Qwen2-VL-7B-Instruct	×	276.06	0.70	0.51	0.52	0.51
mLLaMA3.2-11B	×	829.64	0.66	0.38	0.36	0.36
Qwen2.5-VL-7B-Instruct	×	221.48	0.73	0.76	0.65	0.66
Fine-tuned InternVL2.5-8B	✓	163.16	0.89	0.88	0.89	0.89
Fine-tuned MiniCPM-V-2.6	✓	636.26	0.90	0.89	0.90	0.89
Fine-tuned Qwen2-VL-7B-Instruct	✓	436.29	0.91	0.90	0.90	0.90
Fine-tuned mLLaMA3.2-11B	✓	545.18	0.91	0.90	0.90	0.90
Fine-tuned Qwen2.5-VL-7B-Instruct (Ours)	✓	237.73	**0.91**	**0.91**	**0.91**	**0.91**

Table 6: Class-wise comparison of model performance

Category	Metric	VGG16	ViT	Res50	Shuff	Res101	EffB0	Dense	MbNet	InternVL	MiniCPM	Qwen2	mLLaMA	Ours
Overall	Accuracy	0.83	0.82	0.82	0.79	0.82	0.84	0.84	0.82	0.89	0.90	**0.91**	**0.91**	**0.91**
	Precision	0.82	0.81	0.80	0.78	0.81	0.82	0.83	0.81	0.88	0.89	0.90	0.90	**0.91**
	Recall	0.80	0.79	0.80	0.76	0.80	0.82	0.82	0.80	0.89	0.90	0.90	0.90	**0.91**
	F1 score	0.81	0.80	0.80	0.77	0.80	0.82	0.82	0.80	0.89	0.89	0.90	0.90	**0.91**
Brick	Precision	0.87	0.82	0.84	0.80	0.84	0.86	0.85	0.85	0.93	0.92	**0.94**	0.93	0.92
	Recall	0.84	0.82	0.84	0.79	0.82	0.83	0.85	0.83	0.91	0.93	**0.94**	**0.94**	**0.94**
	F1 score	0.86	0.82	0.84	0.79	0.83	0.85	0.85	0.84	0.92	0.93	**0.94**	0.93	0.93
Concrete	Precision	0.77	0.76	0.78	0.71	0.75	0.77	0.81	0.77	0.83	0.85	0.85	**0.86**	**0.86**
	Recall	0.75	0.73	0.71	0.70	0.74	0.78	0.74	0.72	0.80	0.82	0.83	0.83	**0.85**
	F1 score	0.76	0.75	0.75	0.70	0.75	0.77	0.78	0.74	0.82	0.84	0.84	0.85	**0.86**
Glass	Precision	0.85	0.88	0.86	0.87	0.88	0.88	0.86	0.86	0.93	0.93	0.94	0.94	**0.95**
	Recall	0.93	0.95	0.94	0.92	0.93	0.95	0.95	0.94	0.95	**0.96**	0.95	0.95	**0.96**
	F1 score	0.89	0.91	0.90	0.89	0.90	0.91	0.91	0.90	0.94	**0.95**	**0.95**	0.94	**0.95**
Metal	Precision	0.78	0.77	0.74	0.73	0.79	0.78	0.79	0.76	0.84	0.86	0.87	0.85	**0.88**
	Recall	0.72	0.73	0.72	0.69	0.72	0.73	0.74	0.72	0.83	0.84	0.84	0.85	**0.88**
	F1 score	0.75	0.75	0.73	0.71	0.76	0.76	0.77	0.74	0.83	0.85	0.85	0.85	**0.88**
Other	Precision	0.81	0.83	0.77	0.76	0.76	0.74	0.78	0.79	0.85	0.87	**0.89**	**0.89**	**0.89**
	Recall	0.67	0.66	0.72	0.60	0.66	0.74	0.69	0.66	0.87	0.88	0.88	0.88	**0.91**
	F1 score	0.73	0.73	0.75	0.67	0.70	0.74	0.74	0.72	0.86	0.87	0.88	0.88	**0.90**
Stone	Precision	0.89	0.88	0.87	0.87	0.89	0.89	0.92	0.89	0.93	0.94	0.94	0.94	**0.95**
	Recall	0.87	0.87	0.88	0.84	0.87	0.89	0.88	0.88	**0.94**	0.92	**0.94**	**0.94**	**0.94**
	F1 score	0.88	0.87	0.88	0.86	0.88	0.89	0.90	0.88	0.94	0.93	0.94	0.94	**0.95**
Stucco	Precision	0.82	0.81	0.80	0.77	0.82	0.85	0.83	0.81	0.87	0.87	0.89	**0.90**	0.89
	Recall	0.84	0.82	0.81	0.81	0.83	0.83	0.85	0.83	0.87	**0.89**	**0.89**	**0.89**	**0.89**
	F1 score	0.83	0.81	0.80	0.79	0.82	0.84	0.84	0.82	0.87	0.88	**0.89**	**0.89**	**0.89**
Tile	Precision	0.73	0.73	0.68	0.66	0.69	0.76	0.74	0.73	0.81	0.83	0.83	0.84	**0.88**
	Recall	0.71	0.68	0.71	0.67	0.71	0.73	0.72	0.72	0.86	0.86	0.87	**0.88**	0.83
	F1 score	0.72	0.70	0.70	0.66	0.70	0.74	0.73	0.72	0.83	0.84	0.85	**0.86**	0.85
Wood	Precision	0.87	0.85	0.86	0.84	0.85	0.87	0.86	0.85	0.95	**0.96**	**0.96**	0.95	**0.96**
	Recall	0.90	0.88	0.86	0.83	0.89	0.91	0.91	0.90	0.95	0.96	**0.97**	**0.97**	**0.97**
	F1 score	0.88	0.87	0.86	0.84	0.87	0.89	0.88	0.88	0.95	0.96	**0.97**	0.96	**0.97**

Note: Abbreviations: VGG16: VGG16 (Simonyan and Zisserman, 2015); ViT: Vision Transformer (ViT-B/16) (Dosovitskiy et al., 2021); Res50: ResNet-50 (He et al., 2016); Shuff: ShuffleNetV2 (Ma et al., 2018); Res101: ResNet-101 (He et al., 2016); EffB0: EfficientNet-B0 (Tan and Le, 2019); Dense: DenseNet-161 (Huang et al., 2017); MbNet: MobileNetV3-Large (Howard et al., 2019); InternVL: fine-tuned InternVL2.5-8B (Chen et al., 2024); MiniCPM: fine-tuned MiniCPM-V-2.6 (Yao et al., 2024); Qwen2: fine-tuned Qwen2-VL-7B-Instruct (Wang et al., 2024a); mLLaMA: fine-tuned mLLaMA3.2-11B (Meta AI, 2024); **Ours**: our fine-tuned Qwen2.5-VL-7B-Instruct (Bai et al., 2025). Bold values indicate the best performance for each metric.

Changes in the manuscript:

"We evaluated several vision-language models on the manually labeled test set (11,827 images) to select the most suitable backbone (Table 1). Among the models tested without fine-tuning, Qwen2.5-VL-7B-Instruct achieved the highest F1 score while maintaining moderate inference time. To further validate this selection, we fine-tuned representative VLMs using the same training set (27,578 images), test set, and evaluation metrics. The fine-tuned Qwen2.5-VL-7B-Instruct achieved an F1 score of 0.91 with a favorable accuracy-efficiency trade-off, substantially outperforming the non-fine-tuned VLMs and traditional vision backbones (e.g., ResNet, ViT) trained on the same data (Figure 4, Supplementary Table S2)."

Comment 6. *There is a capitalization error in line 270. The phrase "In addition, The year..." should be revised to "In addition, the year...".*

Response:

Thank you for pointing out this error. We have corrected "The" to "the" in the revised manuscript.

Changes in the manuscript:

> *"In addition, the year-specific material records can provide exploratory evidence of temporal material changes associated with urban renewal processes."*

We sincerely thank the reviewer again for these thoughtful and constructive comments, which have helped us improve the clarity, methodological transparency, and reliability of the manuscript.

References

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. (2025). Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Chen, X., Ding, X., and Ye, Y. (2025). Mapping sense of place as a measurable urban identity: Using street view images and machine learning to identify building façade materials. *Environment and Planning B: Urban Analytics and City Science*, 52(4):965–984.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., and Wang, W. (2024). Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Liang, X., Xie, J., Zhao, T., Stouffs, R., and Biljecki, F. (2025). Openfacades: an open framework for architectural caption and attribute data enrichment via street view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:918–942.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131.

Meta AI (2024). Llama 3.2 Vision Model Card.

Raghu, D., Bucher, M. J. J., and De Wolf, C. (2023). Towards a 'resource cadastre' for a circular economy–urban-scale building material detection using street view imagery and computer vision. *Resources, Conservation and Recycling*, 198:107140.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Sun, K., Li, Q., Liu, Q., Song, J., Dai, M., Qian, X., Gummidi, S. R. B., Yu, B., Creutzig, F., and Liu, G. (2025). Urban fabric decoded: High-precision building material identification via deep learning and remote sensing. *Environmental Science and Ecotechnology*, 24:100538.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. (2024a). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, S., Park, S., Park, S., and Kim, J. (2024b). Building façade datasets for analyzing building characteristics using deep learning. *Data in Brief*, 57:110885.

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., and Sun, M. (2024). MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.