What Pedestrians Lack Around the World: Auditing Street Infrastructure in 75 Cities with On-Device Vision-Language Models

Md Shadab Alam[1,*], Alexandros Rouchitsas[2] and Pavlo Bazilinskyy[1]

[1] *Eindhoven University of Technology, Eindhoven, The Netherlands*

[2] *KTG Research & Innovation*

m.s.alam@tue.nl, a.rouchitsas@ktg-research.eu, p.bazilinskyy@tue.nl

[*] *Corresponding author: Md Shadab Alam, m.s.alam@tue.nl*

ABSTRACT

Pedestrians account for almost a quarter of road deaths worldwide, and much of their risk follows from missing footways, crossings and lighting. Auditing streets in person is slow, so the infrastructure of most of the world's streets is unknown. We asked whether vision-language models (VLMs), run locally, can audit pedestrian infrastructure from dashcam video and reveal differences between regions. Six open VLMs from four countries (Qwen3.8-27B, Gemma 4 31B, Molmo2-8B, InternVL3.5-8B, Ministral 3 14B and Kanana-1.5-V 3B) coded 150 clips from 75 cities, drawn as a spatially balanced probability sample from a global dashcam dataset, with items from two established instruments: the iRAP road attributes and the MAPS-Mini streetscape audit. Each model also listed the infrastructure needed, pointed to it in the frame and suggested a change. Validity was checked on 120 labelled JAAD clips, robustness with a reworded prompt, one frame, half resolution and resampling (8,280 answers in total), and suggestions were rated by two VLM judges. The crossing item matched human labels for five models (balanced accuracy 0.86–0.92). A crossing where people had reason to cross was the most common need (38% of cities; 57% in Africa), followed by footways (50% of African and 23% of North American cities). Needs rose with national road deaths (Spearman's $\rho = 0.33$) and fell with median age (-0.43) and literacy (-0.44). However, the models agreed slightly less on crossings outside Europe and North America and not at all on footway type, and only the two largest models listed needs that the judges could see. VLMs can screen streets at scale, but the measurement is not yet strongest where the needs are greatest.

Keywords: Vision-language models, Pedestrian infrastructure, Street audit, Road safety, Cross-cultural comparison, Dashcam video

INTRODUCTION

About 1.19 million people died on the world's roads in 2021, and almost a quarter of them were pedestrians (World Health Organization, 2023). Much of this risk is built into the street: people walk in the road where there is no footway, cross between moving vehicles where there is no crossing, and are hard to see where there is no lighting. The burden is uneven. Most road deaths occur in low- and middle-income countries, where streets are often shared by people on foot, motorcycles, carts and market stalls, and where the infrastructure for walking is least documented (World Health Organization, 2023).

Road assessment programmes therefore code such features in standard ways. The International Road Assessment Programme (iRAP) codes road attributes, including footway provision and pedestrian crossing facilities, for 100 m segments and uses them to rate how safe a road is for each road user (iRAP, 2022). Public-health research uses streetscape audits such as MAPS-Mini, a short tool that records the features that support walking, such as street lighting, kerb ramps, trip hazards and transit stops (Sallis et al., 2015). Both kinds of audit are carried out by trained people, on site or from imagery. Virtual audits with Google Street View give results close to on-site audits for many items (Badland et al., 2010; Rundle et al., 2011), but they still need a person to look at every street. Computer vision on street-level imagery scales further and has been used to measure, for example, perceived safety across cities (Naik et al., 2014; Zhang et al., 2018; see Biljecki & Ito, 2021, for a review). Such models, however, are trained for fixed labels and need annotated data for every new item and, often, every new region.

Vision-language models (VLMs) change this. A VLM can follow an audit protocol written in plain language, answer each item in a fixed format, point to what it sees and suggest an improvement, without being trained for the task. They can already take a pedestrian's perspective: the best models rate crossing in front of an automated car much as people do (Alam & Bazilinskyy, 2025). If their answers are valid, VLMs could audit the streets of cities that have never been audited, and communicate what is missing in a form that residents and planners understand. Two questions stand in the way. First, vision models perform worse on images from lower-income and non-Western settings (DeVries et al., 2019; Rojas et al., 2022), and VLMs show a Western bias in image understanding (Ananthram et al., 2025). An audit

that is less reliable in exactly the regions with the greatest needs would mislead. Second, VLMs are increasingly used to judge other models' output, but such judges can favour particular answers or their own generations (Zheng et al., 2023; Panickssery et al., 2024), so their ratings of suggestions need checking too.

Dashcam footage now covers the streets of cities on every continent. The CROWD dataset holds continuous urban driving videos from cities worldwide, with each video mapped to its city and country (Alam et al., 2026). Together with open VLMs that run on one consumer GPU, this makes a worldwide audit of pedestrian infrastructure possible without sending footage to a cloud service, which matters for privacy and for reproducibility.

Aim of the Study

This study examines whether on-device VLMs can audit pedestrian infrastructure from dashcam footage in a way that supports comparison between regions and cultures. Specifically, we ask (1) which pedestrian infrastructure is present and which is missing on ordinary streets, (2) how this differs between continents and whether it follows country indicators such as road deaths, median age, inequality and literacy, (3) whether the models audit streets outside Europe and North America as reliably as inside, (4) how robust the answers are to the wording of the prompt, the number of frames, the image resolution and random sampling, and (5) whether the infrastructure the models report as needed is visible in the footage and their suggested improvements are specific and feasible.

METHODS

Footage and City Sample

Cities were drawn from the CROWD mapping of dashcam videos to cities (Alam et al., 2026). Eligible cities had at least one hour of daytime footage from a car, valid coordinates and a known population. From these, 75 cities in 43 countries were drawn with the local pivotal method, which gives a spatially balanced probability sample (Grafström et al., 2012), with inclusion probability proportional to the square root of population. The sample holds the following cities, by continent and country (state or province where recorded). *Asia* (32): Azerbaijan: Baku; China: Beijing, Chengdu, Huai'an, Jiangmen (Guangdong), Jiujiang, Lhasa (Tibet), Nanning, Tianjin, Zhuhai; India: Chennai (Tamil Nadu); Indonesia: Demak, Denpasar, Jakarta, Pekanbaru; Israel: Jerusalem; Japan: Fukuoka, Kanazawa (Ishikawa), Yokohama (Kanagawa); Kazakhstan: Almaty; Lebanon: Beirut; Pakistan: Karachi, Peshawar; Philippines: Davao (Davao del Sur), Lucena; Russia: Krasnoyarsk; South Korea: Incheon; Taiwan: Taipei; Thailand: Rayong; Türkiye: Istanbul; Vietnam: Ho Chi Minh City; Yemen: Taiz. *Europe* (15): France: Montpellier; Germany: Bamberg; Italy: Florence; Kosovo: Peja; Luxembourg: Luxembourg City; Netherlands: Nijmegen; Poland: Gdańsk; Romania: Bucharest; Russia: Kazan, Saint Petersburg, Smolensk; Ukraine: Kharkiv (Kharkiv Oblast), Yalta (Crimea); United Kingdom: Bath, Edinburgh. *North America* (15): Canada: Montreal (QC); Jamaica: Kingston; Mexico: Chilpancingo, Puebla; United States: Austin (TX), Carlsbad (CA), Cleveland (TN), Golden (CO), Jersey City (NJ), Madison (WI), Napa (CA), Oregon City (OR), Spring Hill (FL), Springfield (MO), The Woodlands (TX). *Africa* (7): Algeria: Annaba; DR Congo: Kinshasa; Morocco: Tétouan; Nigeria: Benin City; Rwanda: Kigali; Senegal: Dakar; South Africa: Pretoria. *South America* (5): Argentina: Buenos Aires; Brazil: Manaus (Amazonas), São Paulo; Colombia: Valledupar; Ecuador: Ambato. *Oceania* (1): Australia: Sydney (NSW).

For each city, the video with the most qualifying footage was cut into 10 s windows. Windows were ranked by how much their keyframes matched descriptions of pedestrian-relevant scenes (footway condition, obstructions, crossings, pedestrians in the road, accessibility and vulnerable road users) rather than routine scenes such as motorways or tunnels, using SigLIP2 image–text embeddings, and one random window was added. The best-ranked and the random window of each city formed the 150 city clips; the random clips keep the sample representative of ordinary street stretches. Each clip was given to the models as five frames, 2 s apart, with the longest side at 1,280 pixels.

To check validity against human labels, 120 clips from the JAAD dataset of pedestrian behaviour (Rasouli et al., 2017) were added, half with a zebra crossing visible and half without, each with JAAD's labels for the same frames: zebra crossing, traffic signal, pedestrians visible and pedestrians in the road.

Audit Instrument

The prompt asks the model to code the clip as a trained road-safety auditor would code a 100 m segment, using iRAP road-attribute definitions and MAPS-Mini items (Table 1). iRAP supplies the crossing facility and its quality, footway provision on each side, pedestrian fencing, number of lanes, median, area type and street lighting; MAPS-Mini adds land use, traffic calming, transit stops, bike facilities, kerb ramps and trip hazards. Footway obstruction was added because it is common in many cities. The prompt tells the model to code only what it can see and to ignore the car's

bonnet, reflections and overlays such as speed or time stamps, and it defines each option, for example that a kerb alone is not a physical barrier and that an informal path is a trodden strip without a sidewalk.

From the item answers, eight needs follow by fixed rules: a footway on a side where there is none or only an informal path; clearing a blocked footway; repairing trip hazards; a crossing where people have reason to cross and there is no facility; a better crossing where its quality is poor; a refuge island on a road with four or more lanes and an at-grade crossing without one; kerb ramps; and street lighting on an urban road without any. The model also lists each need itself, with the evidence it sees, the frame and point where it is clearest, a standard measure and a one-sentence suggestion that changes the street rather than people's behaviour. Answers are constrained to a JSON schema, so every answer contains every item.

Table 1. Audit items, their source and answer options.

Part	Item (source)	Options
Road	Area type, number of lanes, median, street lighting, pedestrian fencing (iRAP)	Urban/rural; 1, 2, 3, 4+; none/painted/physical; none/some/ample; present/not present
Road	Land use, traffic calming, transit stops, bike facility (MAPS-Mini)	Residential/commercial/mixed/other; present/not present; 0, 1, 2+; none/painted line/physical barrier
Each side	Footway provision (iRAP)	Physical barrier; sidewalk ≥ 3 m, 1–3 m, < 1 m; informal path ≥ 1 m, < 1 m; none
Each side	Obstruction (added); trip hazards (MAPS-Mini)	None/partly blocked/blocked; none/some
Crossing	Reason to cross (demand); crossing facility and quality (iRAP)	Grade separated; signalised or marked, each with or without refuge; raised; refuge only; none; adequate/poor
Crossing	Kerb ramps (MAPS-Mini)	Both kerbs/one kerb/none

Models and Runs

Six open VLMs from six developers in four countries (the United States, China, France and South Korea) were used (Table 2), so that a finding shared by several models is not one family's or one culture's bias. All ran on one NVIDIA RTX 5090 GPU (32 GB) with vLLM (Kwon et al., 2023) at temperature 0, with the answer schema enforced during decoding. The two larger models ran as 4-bit (W4A16) checkpoints to fit the GPU, Ministral 3 in its developer's FP8 release with the developer's own tokeniser, and each model was pinned to a fixed checkpoint revision. Five further models were excluded: EXAONE 4.5 33B (LG, South Korea) returned nonsense for every image and Falcon2-11B-VLM (TII, UAE) could not be served, while Aya Vision 8B, North-Micro-Vision (Cohere, Canada) and Chitrarth (Krutrim, India) gave degenerate audits, with every need in the frame's corner or all eight in almost every clip, and found crossings in JAAD at chance (0.50–0.58). No footage left the machine.

Table 2. The six VLMs.

Model	Developer	Run as	Reference
Qwen3.8-27B	Alibaba (Qwen)	4-bit, W4A16	Qwen Team (2026)
Gemma 4 31B	Google	4-bit, quantisation-aware	Google (2026)
Molmo2-8B	Allen Institute for AI	BF16	Allen Institute for AI (2026)
InternVL3.5-8B	Shanghai AI Laboratory	BF16	Wang et al. (2025)
Ministral 3 14B	Mistral AI	FP8 (developer's release)	Liu et al. (2026)
Kanana-1.5-V 3B	Kakao	BF16	Kakao (2025)

The main run answered all 270 clips (150 city and 120 JAAD clips). Four further runs tested robustness: a reworded prompt with the same items and schema, one frame (the middle one) instead of five, half the image resolution (640 pixels), and, on the city clips, two runs at temperature 0.7 with different seeds. This gave 8,280 answers, all of which were valid.

Suggestions Rated by VLM Judges

The two larger models served as judges. For each city answer of the main run that listed at least one need (603 answers, 1,874 needs; Kanana listed none), a judge saw the same frames and the needs with their evidence and suggestions,

and rated each need: is the deficit visible at that place (yes, no, cannot tell), would the suggestion fix it, is the suggestion specific to this street or generic, is it feasible given the space visible (yes, no, unclear), and does it tell people how to behave instead of changing the street. Both judges rated all needs, including those of their own model, so that self-preference could be measured.

Analysis

For each clip and item, the *consensus* answer is the one given by at least two thirds of the models that answered (four of six; at least four answers). Shares of cities with an item present or a need were computed per continent from the consensus, averaging each city's two clips, with 95% intervals from 2,000 bootstrap resamples of cities. Spearman's ρ related each city's shares to national indicators from the CROWD mapping: road traffic deaths per 100,000 people (73 cities with data), median age (75), Gini index (74) and literacy rate (51); ρ was computed only when at least five cities had and five lacked the item. Agreement between the six models was measured with Fleiss' κ (Fleiss, 1971) separately for clips from Europe and North America (60) and from the rest of the world (90), with 95% intervals from 500 bootstrap resamples of clips. Each model's listed needs were compared with the needs its own item answers imply. Robustness was measured with Cohen's κ between each model's main answers and its answers under each changed condition, and between the two temperature-0.7 runs. For the judges, shares rated yes are given with Wilson intervals, judge agreement with Cohen's κ, and self-preference as the difference between the two judges' shares on each auditor's needs. The code and results are public at github.com/Shaadalam9/vlm-street-audit.

RESULTS

Validity on Labelled Clips

On the 120 JAAD clips, the iRAP crossing-facility item found the labelled marked crossings with a balanced accuracy of 0.86 to 0.92 for five of the six models, as high as the direct question about a zebra crossing (0.86–0.92); the smallest, Kanana, reached only 0.57 (zebra crossing 0.67). Visible pedestrians were found with 0.87–0.99 and traffic signals with 0.73–0.98. Pedestrians walking in the road were the hardest scene fact (0.57–0.76).

What the Models Report

Figure 1 shows six needs as the models reported them, one from each continent, with the point each model gave on the frame it named. In five of them, both judges rated the deficit visible and the suggestion as fixing it: a residential street in Spring Hill, Florida, with grass verges instead of footways (a), a crossing sign with no marked crossing in Kazan (b), a cart blocking the only path in Peshawar (c), market stalls and parked motorcycles crowding the path in Valledupar (d), and a man crossing a wide road to a market with no crossing in Kinshasa (e). The sixth shows a typical failure: on the Sydney Harbour Bridge three of the six models reported no footway, and both judges agreed, although a footway runs behind the barrier and fence on the left, separated from the roadway. With three of six, the report falls short of the consensus.

Infrastructure Present and Needed

Street lighting was the most common feature, present in 83% of cities by consensus (Table 3). A footway set back at least 1 m or behind a barrier and a formal footway on both sides were each found in 51% of cities. A crossing facility was in view in 22% of cities and a signalised or grade-separated crossing in 7%; refuge islands and traffic calming were never agreed on. African cities stood out with formal footways on both sides in 14% of cities, against 50–80% elsewhere, crossings in 7% and lighting in 50%.

What cities lack falls into a clear geography (Figure 1). In 39 of the 75 cities, the models agreed that at least one clip needed a crossing where people had reason to cross, in 22 that a footway was needed, and in 12 both; 22 cities had no agreed need. Ten of the 12 cities lacking both a crossing and a footway lie in Africa, South Asia, the Middle East and around the Caribbean, while in North America a footway alone is the most common gap (5 of 15 cities).

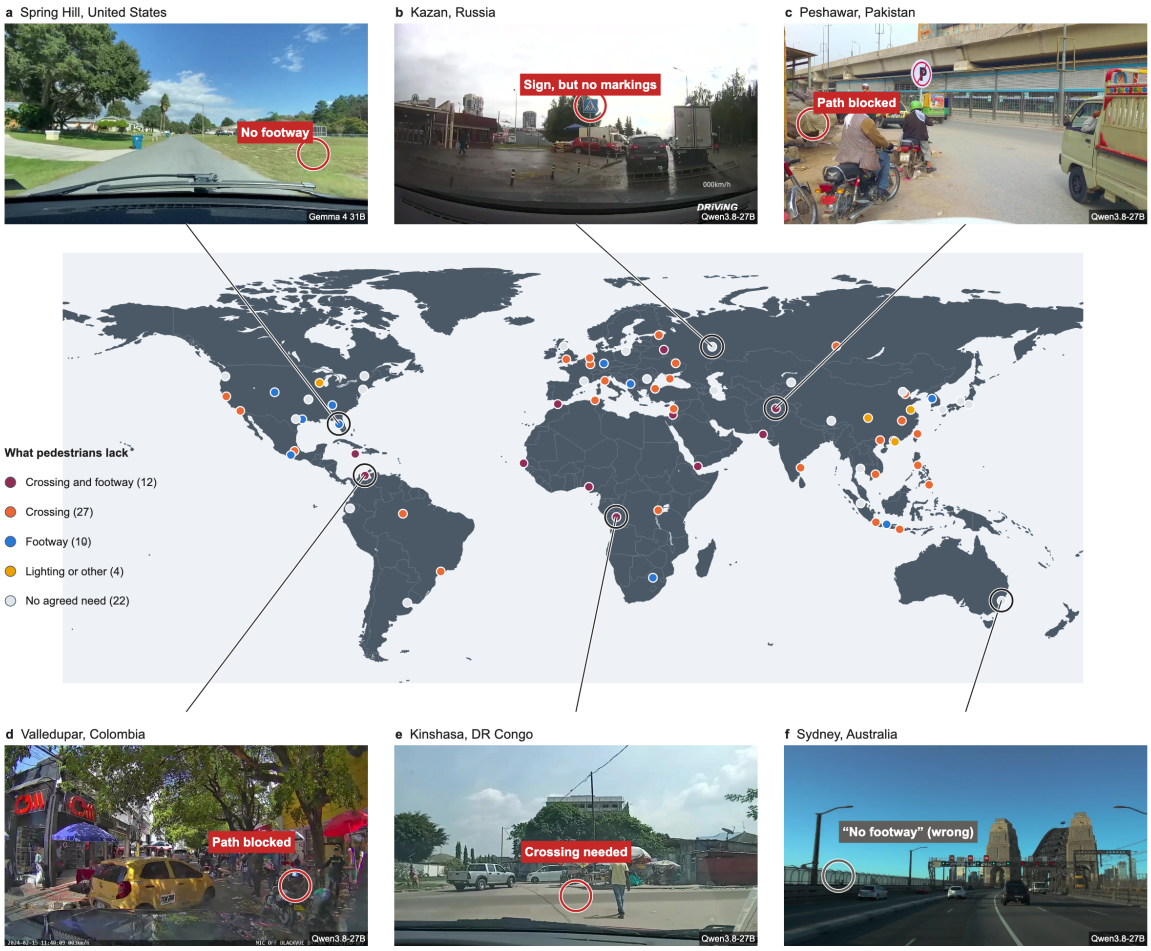


Figure 1. What pedestrians lack in each of the 75 cities (needs on which at least four of the six models agree in at least one of the city's two clips), with six of the streets and the point each model gave (ring) on the frame it named. Suggestions: (a) a continuous sidewalk; (b) a marked crossing with dropped kerbs at the sign; (c) remove the cart and debris to clear the path; (d) regulating the stalls and a no-parking zone for motorcycles; (e) a marked crossing with a central refuge island. (f) A false report: the bridge's footway runs behind the barrier and fence, separated from the roadway.

By continent, a crossing where people had reason to cross was needed in 38% of cities overall: 57% in Africa, 42% in Asia, 40% in South America, 37% in Europe and 23% in North America (Figure 2a). A footway was needed in 17% of cities, but in 50% of African and 23% of North American cities, against 10–12% in Europe and Asia. Street lighting was seldom needed (5%; 10% in North America and 14% in Africa). Clearing blocked footways, better crossings and kerb ramps were never needed by consensus. With 7 cities in Africa and 5 in South America, their intervals are wide (for crossings in Africa, 36–79%).

Table 3. Share of cities in which at least four of the six models agree that a feature is present, by continent (number of cities in brackets; 95% bootstrap intervals for all cities). Oceania (one city) is included in the total only.

Feature	Africa (7)	Asia (32)	S. America (5)	Europe (15)	N. America (15)	All (75)
Formal footway on both sides	14%	52%	80%	60%	50%	51% (41–61%)
Footway set back ≥ 1 m or behind a barrier	29%	53%	60%	60%	47%	51% (42–60%)
Crossing facility in view	7%	19%	50%	20%	27%	22% (15–29%)
Signalised or grade-separated crossing	0%	6%	20%	7%	7%	7% (3–12%)
Street lighting	50%	95%	90%	83%	70%	83% (77–89%)

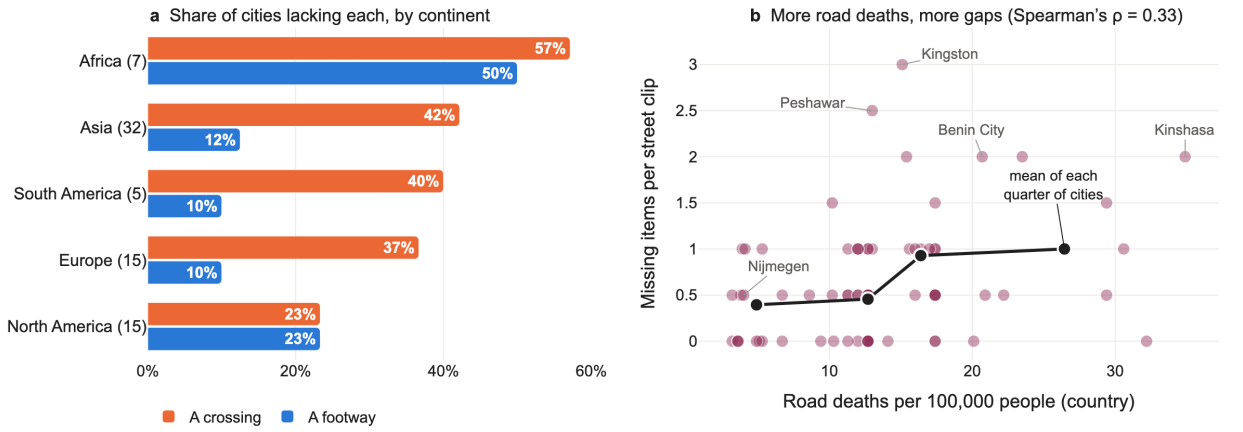


Figure 2. (a) Share of cities needing a crossing where people have reason to cross, and a footway, by continent (number of cities in brackets; Oceania, one city, is left out). (b) Missing items per street clip (consensus of the six models, averaged over each city's two clips) against the country's road deaths per 100,000 people, for 73 cities; the line joins the means of each quarter of the cities.

Country Indicators

Cities in countries with more road deaths had more missing infrastructure (Figure 2b). The number of needs per clip rose with road deaths per 100,000 ($\rho = 0.33$, 95% CI 0.11 to 0.52): Kingston had three missing items per clip, Peshawar two and a half, and Benin City, Dakar, Kinshasa and Valledupar two, while Bamberg and Nijmegen had half an item. Needs also fell with median age (−0.43, −0.61 to −0.21) and literacy (−0.44, −0.66 to −0.17). Formal footways on both sides were more common where median age and literacy were higher (0.43 each). Inequality showed few associations, except that footways were needed more often where the Gini index was higher (0.27). Kingston and Peshawar lie well above the trend: by consensus, their streets lacked footways, crossings and street lighting.

Model Agreement by Region

The six models agreed moderately on area type ($\kappa = 0.46$), crossing facility (0.42) and crossing quality (0.39), but not at all on footway type (0.00), trip hazards (0.01), footway obstruction (0.05) and kerb ramps (0.00) (Figure 4). Figure 3 shows what this means on one street: on a market street in Kingston, all six models saw people walking in the road, a reason to cross and no crossing, but gave five different answers for each footway. Kanana lowered agreement on most items; it coded a third of the city clips as rural (area type 0.74 among the other five). Agreement on the crossing items was somewhat lower for clips outside Europe and North America: reason to cross 0.23 against 0.35, crossing quality 0.35 against 0.45 and crossing type 0.40 against 0.46, while street lighting was the same (0.20). The intervals overlap for each item, and the gap was smaller with Ministral 3 than among the four models without it. For other items, such as bike facilities, lanes and median, agreement was higher outside Europe and North America.

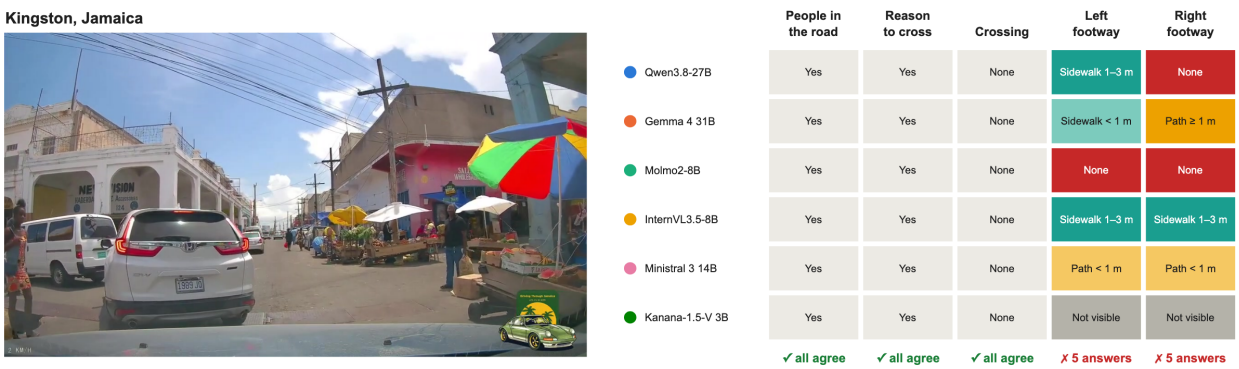


Figure 3. One street in Kingston, Jamaica, as coded by the six models: they agree on what is easy to see and give five different answers for each footway.

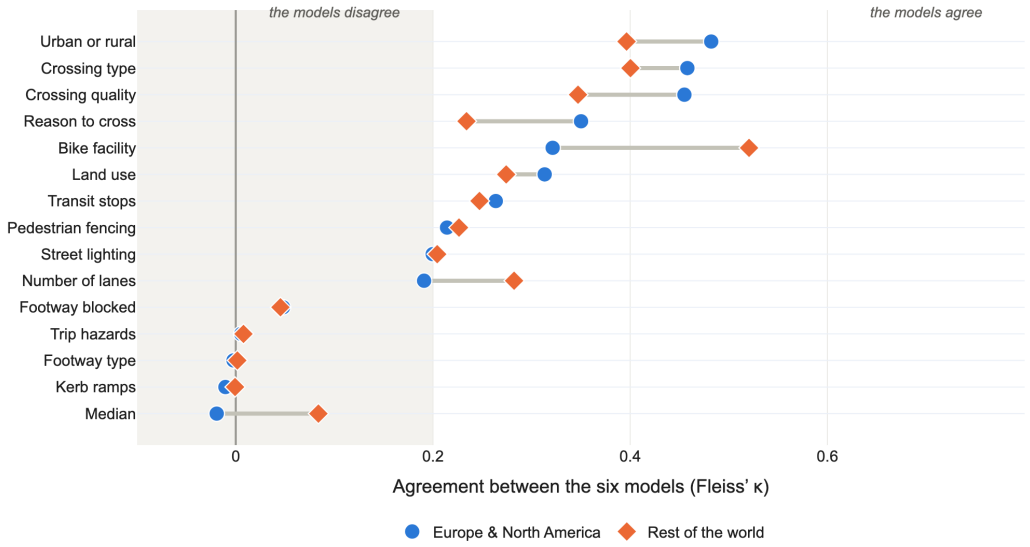


Figure 4. Agreement between the six models on each item (Fleiss' κ) for clips from Europe and North America (60) and from the rest of the world (90). The shaded band marks agreement below 0.2.

The models' own lists of needs did not always follow their item answers. Gemma 4 listed only needs backed by an item and was fully consistent in 90% of answers in both regions. Qwen3.8 was fully consistent in 57% of answers from Europe and North America but in 33% elsewhere. Ministral 3 backed 57–61% of its listed needs with an item answer, Molmo2 and InternVL3.5 only 24–37%, and Kanana listed no needs at all, although its item answers implied 0.6 per clip.

Robustness

Rewording the prompt or halving the image resolution changed few answers (mean κ with the main answers 0.68 and 0.73; Table 4). Giving one frame instead of five changed more (0.56), and two samples at temperature 0.7 agreed with each other at only 0.50. Footway type was the least stable item (0.28 between samples), while crossing facility and quality were among the most stable (0.57–0.76). Robustness was similar in both regions.

Table 4. Robustness: mean Cohen's κ over items and models between the main answers and each changed condition (city clips). Resampling compares two temperature-0.7 runs with each other.

Change	All clips	Europe & N. America	Rest of the world	Footway item
Reworded prompt	0.68	0.70	0.67	0.48
One frame instead of five	0.56	0.58	0.54	0.46
Half resolution (640 px)	0.73	0.73	0.73	0.69
Resampled (temperature 0.7)	0.50	0.53	0.49	0.28

Needs and Suggestions Rated by the Judges

The two larger models were the sound auditors (Figure 5). On average over the two judges, 81% of the needs listed by Qwen3.8 and Gemma 4 were visible in the frames, against 28–31% for Molmo2, InternVL3.5 and Ministral 3; and 77% and 100% of the larger models' listed needs followed from their own item answers, against 27–59% for the smaller ones. Ministral 3 listed the most needs (621, about four per clip), so listing more did not mean finding more. Almost every suggestion was rated as fixing its stated deficit (67–99%), and the judges agreed on this (κ = 0.70). Suggestions were seldom specific to the street (0–61%), and none told people to behave differently.

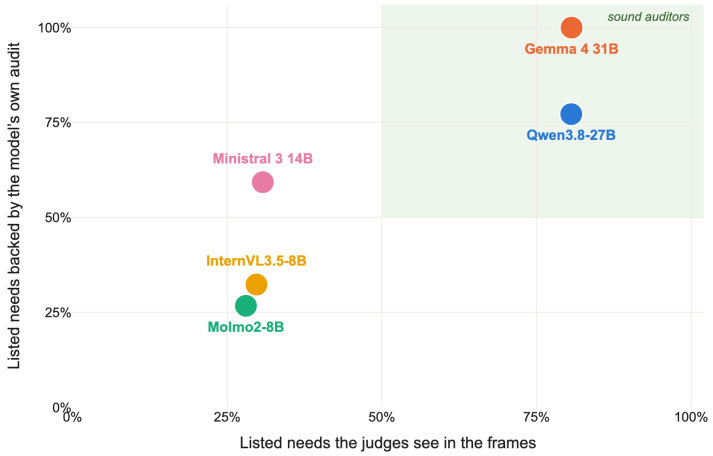


Figure 5. The five models that listed needs, as auditors: share of their listed needs that the judges rated visible in the frames (mean of two judges) against the share backed by the model's own item answers.

The judges disagreed on feasibility ($\kappa = 0.06$): the Gemma judge rated 76% of the larger models' suggestions feasible, the Qwen judge 10–13%. The Gemma judge was more lenient than the Qwen judge by a similar margin on Qwen3.8's suggestions as on its own (deficit visible +22 and +20 points, feasible +62 and +67 points), so there was no clear sign that a judge favoured its own model. Judged quality was about the same for both regions (deficit visible 44% and 47%).

DISCUSSION

Six open VLMs from four countries, run on one GPU, audited pedestrian infrastructure in 75 cities with items from iRAP and MAPS-Mini. Where human labels exist, the crossing item was as accurate as a direct question about a zebra crossing. Across cities, the results describe a plausible geography of missing infrastructure. The most common gap everywhere is a crossing where people have a reason to cross. It is most frequent in Africa and Asia, where pedestrians make up a large share of road deaths (World Health Organization, 2023). Missing footways show two different patterns: in African cities, basic provision is often absent, while in North American cities many roads are built without footways at all, as in the residential street of Figure 1a.

The associations with country indicators point the same way. Cities in countries with more road deaths, younger populations and lower literacy show more needs and fewer formal footways. These indicators all track national income, and the footage covers the streets that video makers drove, so the associations describe a pattern rather than a cause. Still, an audit derived from video recovers the gradient one would expect, which supports its use for comparison. The cities far above the trend are also informative: busy streets where vending, parking and walking compete for the same space produce several needs at once, as in Figure 1c and 1d, which suggests that street management matters as much as construction.

The main concern is that the measurement is not strongest where the needs are greatest. The models agree somewhat less on crossings outside Europe and North America, and Qwen3.8's needs follow its own item answers less often there. Adding a model from a European developer narrowed the regional gap in agreement, so the gap reflects particular models as much as the footage. This matches reports of lower performance and Western bias of vision models on images from other regions (DeVries et al., 2019; Ananthram et al., 2025). In practice, a VLM audit of streets in Africa or South Asia needs more validation than one of European streets, not less. Footway type, the item most central to pedestrians, is not reliable in any region: the models disagree on whether a strip beside the road is a sidewalk, an informal path or nothing. Using the consensus of several models from different developers reduces the influence of any one model's errors, but it cannot replace validation.

The choice of model matters. The two larger models listed needs that both judges mostly found visible, and Gemma 4 kept its needs consistent with its items. The three smaller models (8 to 14 billion parameters) listed needs that were mostly not visible and often not backed by their own answers. The smallest, Kanana (3 billion), listed none at all. They should not be used as auditors for this task on their own. The answers of the three to simple scene questions on JAAD were, however, nearly as accurate, which shows that accuracy on scene facts does not guarantee a sound audit.

VLM judges were useful for some questions and not for others. They agreed on whether a suggestion fixes the deficit and broadly on whether a deficit is visible, which allowed us to separate the auditors. They did not agree on feasibility, which depends on space, cost and local practice that the frames show only in part. Unlike reports of self-preference in text tasks (Panickssery et al., 2024), the more lenient judge was lenient towards both large models alike. Judges should therefore come from different developers, and their absolute ratings should not be taken at face value. The Sydney example in Figure 1f shows a further limit: when the infrastructure is hard to see from the car, auditors and judges can agree on a deficit that is not there; here only the consensus of six models kept it out.

The marked frames themselves may be the most useful output for people. A point on the frame, a plain-language description of what is missing and a concrete suggestion communicate a street's problem more directly than an item code, and they let residents, planners and researchers check each claim at a glance. This supports human-centred uses in which VLMs prepare candidate observations and people confirm, correct or reject them.

In conclusion, on-device VLMs can screen pedestrian infrastructure across many cities at low cost and without sharing footage. They reveal clear regional differences: crossings are the most widespread gap, footways are missing in Africa and on North American roads, and needs grow with national road deaths. The method is suited to deciding where to look first and to communicating what is missing, not to replacing audits by trained people, because the models disagree on footways and are somewhat less consistent outside Europe and North America, where most of the needs are.

LIMITATIONS AND FUTURE STUDIES

First, the city clips have no human ground truth. Agreement between models is not correctness, and the only check against human labels used JAAD, which comes from North America and Europe. Second, each city is represented by two 10 s clips from the video with the most footage, filmed from a car on roads that video makers chose to drive; the results describe these streets, not each city as a whole, and footways and crossings on side streets, out of the camera's view or hard to see are under-represented, as the Sydney example shows. Third, Africa, South America and Oceania contribute 13 cities together, so their estimates are uncertain. Fourth, the country indicators are national and literacy is missing for 24 cities. Fifth, the models saw five daytime frames at temperature 0; footage at night, in rain or on video input may give different answers, the two larger models ran in 4-bit form, and five planned models could not be used, so no developer from Canada, India or the UAE is represented. Finally, the suggestions were rated by VLM judges, not by street designers.

Future studies should first collect human audits for a stratified subset of the city clips, with at least half from outside Europe and North America. This would turn agreement into accuracy and test the regional difference in reliability directly. More clips per city, including random clips from side streets, would allow city-level estimates. The footway item could be improved with reference images of each footway category or by giving the model the video instead of frames. Finally, street designers and residents from different regions should rate the marked frames and suggestions, since feasibility and priorities depend on local practice.

SUPPLEMENTARY MATERIAL

In line with current open science practices and recommendations for transparency in automotive user research (Ebel et al., 2024), the authors openly provide these research artefacts to support reproducibility, collaboration and further advancements in the field. The materials used in the study, including the video clips, all answers of the models and the judges, and a snapshot of the analysis code, are available at https://www.dropbox.com/scl/fo/cl1e4cgdjlj43i5gy4gsy/ALxGI5IgWYm9RRxrfey5AYA?rlkey=cu2udsf5ooey7cp2ff6efu4ut&st=zjngk7m6&dl=0. A maintained version of the code is available at https://github.com/Shaadalam9/vlm-street-audit.

REFERENCES

Alam, M. S., & Bazilinskyy, P. (2025). Cross or nah? LLMs get in the mindset of a pedestrian in front of automated car with an eHMI. In *Adjunct Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 19–34). ACM. https://doi.org/10.1145/3744335.3758477

Alam, M. S., Bazilinska, O., & Bazilinskyy, P. (2026). CROWD: A global dataset of continuous urban dashcam driving. *arXiv preprint* arXiv:2604.01044. https://doi.org/10.48550/arXiv.2604.01044

Allen Institute for AI. (2026). *Molmo2-8B* [Model card]. Hugging Face. https://huggingface.co/allenai/Molmo2-8B

Ananthram, A., Stengel-Eskin, E., Vondrick, C., Bansal, M., & McKeown, K. (2025). See it from my perspective: How language affects cultural bias in image understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2406.11665

Badland, H. M., Opit, S., Witten, K., Kearns, R. A., & Mavoa, S. (2010). Can virtual streetscape audits reliably replace physical streetscape audits? *Journal of Urban Health, 87*(6), 1007–1016. https://doi.org/10.1007/s11524-010-9505-x

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

DeVries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 52–59). https://doi.org/10.48550/arXiv.1906.02659

Ebel, P., Bazilinskyy, P., Colley, M., Goodridge, C. M., Hock, P., Janssen, C. P., Sandhaus, H., Srinivasan, A. R., & Wintersberger, P. (2024). Changing lanes toward open science: Openness and transparency in automotive user research. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 94–105). https://doi.org/10.1145/3640792.3675730

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Google. (2026). *Gemma 4 31B* [Model card]. Hugging Face. https://huggingface.co/google/gemma-4-31B-it

Grafström, A., Lundström, N. L. P., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics, 68*(2), 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x

International Road Assessment Programme (iRAP). (2022). *iRAP coding manual: Drive on the left edition* (Version 5.3). International Road Assessment Programme. https://irap.org/specifications/

Kakao. (2025). *Kanana-1.5-V-3B-Instruct* [Model card]. Hugging Face. https://huggingface.co/kakaocorp/kanana-1.5-v-3b-instruct

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles* (pp. 611–626). https://doi.org/10.1145/3600006.3613165

Liu, A. H., Khandelwal, K., Subramanian, S., et al. (2026). Ministral 3. *arXiv preprint* arXiv:2601.08584. https://doi.org/10.48550/arXiv.2601.08584

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore: Predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 779–785). https://doi.org/10.1109/CVPRW.2014.121

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems, 37*. https://doi.org/10.52202/079017-2197

Qwen Team. (2026). *Qwen3.8-27B* [Model card]. Hugging Face. https://huggingface.co/Qwen/Qwen3.8-27B

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 206–213). https://doi.org/10.1109/ICCVW.2017.33

Rojas, W. A. G., Diamos, S., Kini, K. R., Kanter, D., Reddi, V. J., & Coleman, C. (2022). The Dollar Street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems, 35* (Datasets and Benchmarks Track). https://doi.org/10.52202/068431-0943

Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to audit neighborhood environments. *American Journal of Preventive Medicine, 40*(1), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034

Sallis, J. F., Cain, K. L., Conway, T. L., Gavand, K. A., Millstein, R. A., Geremia, C. M., Frank, L. D., Saelens, B. E., Glanz, K., & King, A. C. (2015). Is your neighborhood designed to support physical activity? A brief streetscape audit tool. *Preventing Chronic Disease, 12*, 150098. https://doi.org/10.5888/pcd12.150098

Wang, W., et al. (2025). InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint* arXiv:2508.18265. https://doi.org/10.48550/arXiv.2508.18265

World Health Organization. (2023). *Global status report on road safety 2023*. World Health Organization. https://www.who.int/publications/i/item/9789240086517

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning, 180*, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, 36* (Datasets and Benchmarks Track). https://doi.org/10.52202/075280-2020