Exploring Veo 3's Capabilities for Generating Urban Traffic Scenes in 76 Cities Worldwide

MD SHADAB ALAM*, Eindhoven University of Technology, The Netherlands
ZI WANG, Augusta University, United States of America
LINGHAN ZHANG, Eindhoven University of Technology, The Netherlands
PAVLO BAZILINSKYY, Eindhoven University of Technology, The Netherlands

This study explores the potential of Google Veo 3, a generative video model, to synthesise 8-second dashcam-style urban traffic scenes solely based on text prompts in 76 cities across six continents. YOLOv11x was used to count facts like the number of road users, traffic lights, and stop signs, revealing variations across cities: Karachi had the most objects detected (79), while Muscat had only four cars. Audio analysis using dBFS showed that Montevideo was the loudest, while Copenhagen was the loudest. Through a qualitative visual analysis, the authors assessed and confirmed the perceived authenticity of most traffic scenes and highlighted AI errors, including the inability to handle non-English languages in these videos. Moreover, we compared 10 synthetic videos of New York City and Kampala, each, and verified that Veo 3 is consistent. To summarise, Veo 3 is capable of synthesising authentic, logical traffic scenes worldwide; nevertheless, it still poses non-negligible errors.

CCS Concepts: • **Computing methodologies** → **Image and video acquisition**; Machine learning; • **Human-centered computing** → **Visualization**; • **General and reference** → *Experimentation*; *Verification.*

Additional Key Words and Phrases: Veo 3, Traffic scenes, Cross-cultural, Generative AI, Dashcam

ACM Reference Format:
Md Shadab Alam, Zi Wang, Linghan Zhang, and Pavlo Bazilinskyy. 2024. Exploring Veo 3's Capabilities for Generating Urban Traffic Scenes in 76 Cities Worldwide. 1, 1 (June 2024), 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The proliferation of Generative Artificial Intelligence (Gen AI) models [5] has affected various practical and research domains, including automated driving, urban planning, and human behaviour analysis, among others [2, 4]. Datasets of traffic scenes, including images and videos, serve as fundamental resources for these fields, providing invaluable information for training and benchmarking machine learning models, as well as for analysing human judgment in complex urban scenarios [10, 16, 22].

Traditionally, well-established datasets such as KITTI [16], Cityscapes [13], NuScenes [10], Caltech Pedestrian Detection Benchmark [14], One Thousand and One Hours [17] and Waymo Open Dataset [22] have been widely used for a variety of purposes, ranging from object detection and semantic segmentation to trajectory prediction and human behaviour modelling. For instance, the Mask2Former model leverages Cityscapes for urban semantic segmentation tasks [12], while RobMOT uses KITTI and Waymo datasets to achieve robust multi-object tracking [20]. Furthermore,

*Corresponding Author

Authors' Contact Information: Md Shadab Alam, m.s.alam@tue.nl, Eindhoven University of Technology, Eindhoven, North Brabant, The Netherlands; Zi Wang, zwang1@augusta.edu, Augusta University, Augusta, Georgia, United States of America; Linghan Zhang, l.zhang1@tue.nl, Eindhoven University of Technology, Eindhoven, North Brabant, The Netherlands; Pavlo Bazilinskyy, p.bazilinskyy@tue.nl, Eindhoven University of Technology, Eindhoven, North Brabant, The Netherlands.

2024. Manuscript submitted to ACM

datasets have been used to explore cultural differences in pedestrian behaviour and risk perception worldwide, using video content gathered from diverse geographical locations [1, 8].

Despite their utility, existing datasets rely predominantly on data collected from limited geographical areas, mainly focused on urban centres in developed countries. For example, KITTI[16] was collected in Karlsruhe (Germany), Waymo Open Dataset [22] contains videos collected from the US cities San Francisco (CA), Los Angeles (CA), Seattle (WA), Detroit (MI), Phoenix (AZ) and Mountain View (CA), and One thousand and One hours [17] from Palo Alto, CA, USA. These geographic limitations and the focus on developed Western nations pose biases on the representativeness and applicability of these datasets to reflect global urban diversity. Consequently, generative AI models such as Veo 3 (https://veo3.ai/), SORA (https://openai.com/sora), LTX Studio (https://ltx.studio), etc. present a compelling alternative, promising the potential to synthetically generate realistic urban traffic scenes from various global locations without the constraints of physical data collection.

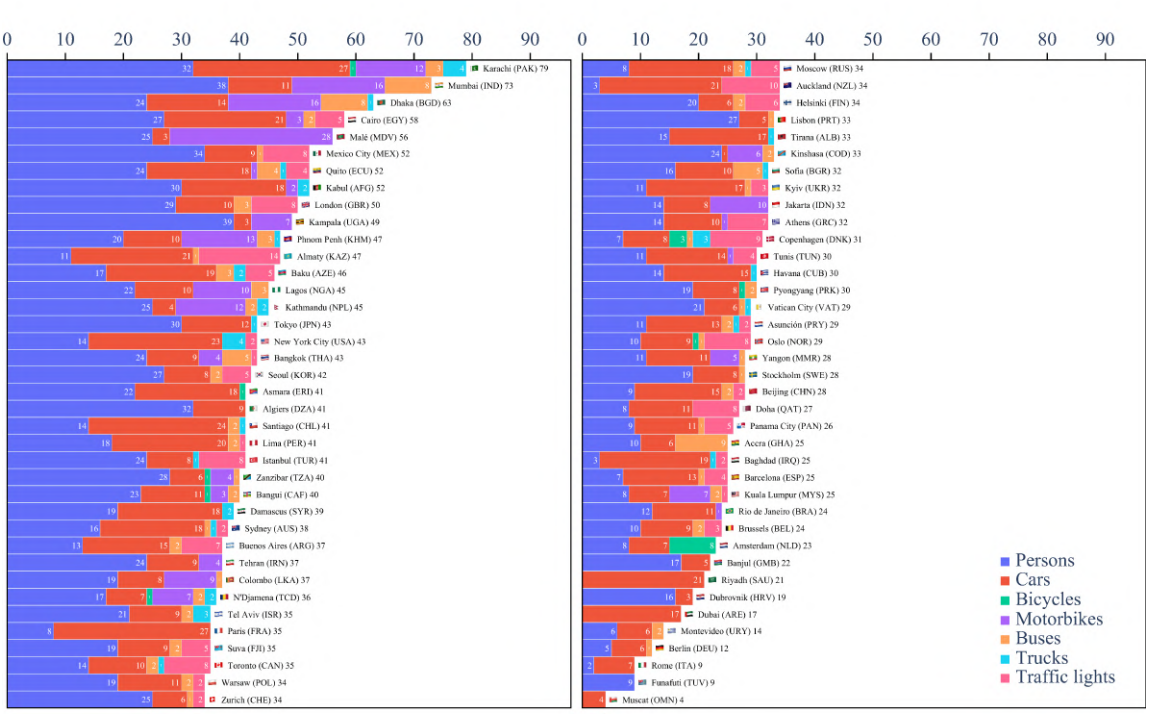


Fig. 1. Counts of objects detected by YOLOv11x. ISO-3 code of the respective country are shown in brackets.

Veo 3, developed by Google DeepMind (https://deepmind.google) and released on 20 May 2025, represents an advanced generative model capable of producing realistic video content with synchronised audio based solely on textual prompts. This innovative capability allows Veo 3 to uniquely generate dynamic and realistic traffic scenarios for locations traditionally under-represented in existing traffic scene datasets, such as cities in developing countries or regions with limited resources for comprehensive data collection. Given the capacity for rapid, cost-effective video generation without the logistical complexities associated with traditional methods, generative models such as Veo 3 have tremendous potential to democratise access to diverse urban scenarios for researchers and practitioners alike.

However, a critical aspect of effectively employing AI-generated content is the fidelity of the videos produced [18]. For generative models to truly serve as viable supplements or alternatives to real-world datasets, their generated outputs must accurately reflect the distinctive cultural, infrastructural and architectural characteristics of specific cities worldwide. Ensuring such realism is essential for applying synthesised data such as traffic scenes to applications in urban planning, driver training simulations, and policy-making visualisations.

The objective of this study is *to evaluate the ability of Veo 3 to synthesise realistic urban traffic scenes for various global cities (76 cities in total, 12 in Africa, 27 in Asia, 21 in Europe, 4 in Oceania, 7 in South America, 5 in North America), with a particular focus on visual and aural authenticity and urban specificity*. Specifically, the study aims to investigate whether Veo 3 can accurately depict urban environments, not only in commonly represented locations such as New York City (USA), but also in less frequently documented settings such as Kampala in Uganda. Through prompt-based video generation and qualitative and quantitative visual and auditory analysis, the study aims to assess the degree to which Veo 3 generated traffic scenes authentically capture city-specific visual and audio characteristics.

2 METHOD

The selection of cities prioritised those with significant international recognition and prominence in global urban research, similar to the approaches used in previous cross-cultural urban studies [3, 8, 9, 11]. The selected cities were fed into Veo 3 prompts through Google Gemini (https://gemini.google.com/app). The prompt used was: *"A realistic dashcam video was captured from the dashboard of a car driving through the streets of {city} during the day"*, where {city} serves as a placeholder for city names (e.g. Almaty, Brussels, Doha). Furthermore, nine additional videos of New York City (United States) and Kampala (Uganda) were generated using the same prompt. Each generated video was 8 seconds long (the maximum duration of videos Veo 3 allows). The videos were generated between 8 June 2025 and 27 June 2025.

To analyse the videos quantitatively, we used YOLOv11x [19] with a confidence of 0.7 to detect various objects, including persons, bicycles, cars, motorbikes, buses, trucks, and traffic lights. We developed a systematic procedure to extract and analyse the loudness of a video's audio track. The audio is programmatically isolated, saved as an uncompressed WAV file, and converted to mono if needed. After normalising the waveform to floating-point values in the range [-1, 1], we calculate the root mean square (RMS) amplitude as a proxy of loudness. The RMS is then converted to dB relative to full scale (dBFS), where 0 dBFS is the maximum digital amplitude and all real audio values are negative. This approach follows industry standards and ensures reproducibility in digital audio analysis [21]. Furthermore, each video was visually analysed by the authors to determine whether the generated frames appeared realistic or not. For reference, the authors have used the PYT dataset that includes dashcam videos from all over the world [1].

3 RESULTS

A total of 124.99 USD was spent on generating the videos. On average, each video took 92.94 s to generate (SD = 17.48 s). The generated videos were 1280 pixels wide by 720 pixels tall in the MP4 format, encoded with H.264 video and AAC stereo audio, at a bitrate of about 5500 kbps. The longest generation time was for Pyongyang (North Korea), at 136.13 s, while the fastest was for Banjul (Gambia), at 59.33 s. In addition, 10 videos were generated for both New York City (USA) and Kampala (Uganda); it took Veo 3 86.32 ± 7.69 s to generate the New York videos and 86.58 ± 4.88 to generate the Kampala ones.

The file sizes of the generated videos varied, with Lisbon (Portugal) producing the largest file at 7.5 MB and Almaty (Kazakhstan) the smallest at 2.79 MB; the average file size was 4.59 MB (SD = 1.03 MB). These differences in size likely reflect variations in video quality, despite consistent generation settings. Some videos, such as those for New York City

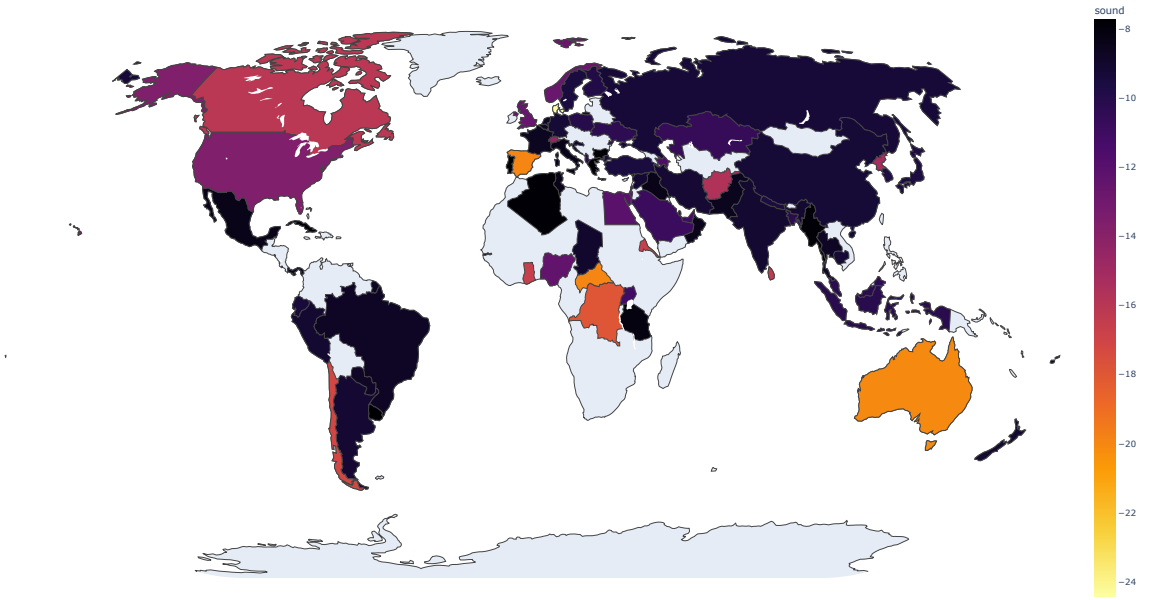


Fig. 2. Mean sound intensity levels (in dBFS) detected in the videos, illustrating the variation in ambient traffic noise captured by the model. The higher the value, the "louder" the video is, and vice versa.

(United States), were noticeably more detailed, while others appeared blurred, with unreadable signs or licence plates. As shown in Fig. 2, analysis of audio tracks revealed considerable variation in maximum dBFS values across the videos. The highest maximum dBFS was observed in Montevideo (Uruguay) at -7.74 dBFS, while the lowest was recorded in Copenhagen (Denmark) at -24.42 dBFS. In general, engine and ambient noise levels tended to be lower in videos of larger cities such as Sydney compared to those from more remote locations.

Across the 76 cities surveyed, the results of object detection with YOLOv11x reveal a wide variation in the counts for each category; see Fig. 1. The person category dominates the detections, with a maximum of 39 in Kampala (Uganda), followed by Mumbai (India, 38), Mexico City (Mexico, 34), and Algiers (Algeria). In contrast, cities such as Dubai (United Arab Emirates), Riyadh (Saudi Arabia), and Muscat (Oman) recorded zero persons. Car detections peak at 27 in both Paris (France) and Karachi (Pakistan), with other high values including Santiago (Chile, 24), Auckland (New Zealand, 21) and Cairo (Egypt, 21). Motorbike detections reach their highest levels in Malé (Maldives, 28), Mumbai (India, 16), Dhaka (Bangladesh, 16), Kathmandu (Nepal, 12), and Karachi (Pakistan, 12), while many cities do not record motorbikes at all, such as Muscat (Oman), Rome (Italy), and Tirana (Albania). The highest bus count is observed in Accra (Ghana, 9), followed by Mumbai (India, 8) and Dhaka (Bangladesh, 8), while the maximum number of trucks is 4 in New York City (USA) and Karachi (Pakistan) each. Traffic light detections are highest in Almaty (Kazakhstan, 14), Auckland (New Zealand, 10), and several cities, including Toronto (Canada), Doha (Qatar), Mexico City (Mexico), London (UK) and Oslo (Norway), register 8 each. The bicycle detections are low overall, with Amsterdam showing the highest count at 8, and Helsinki at 3.

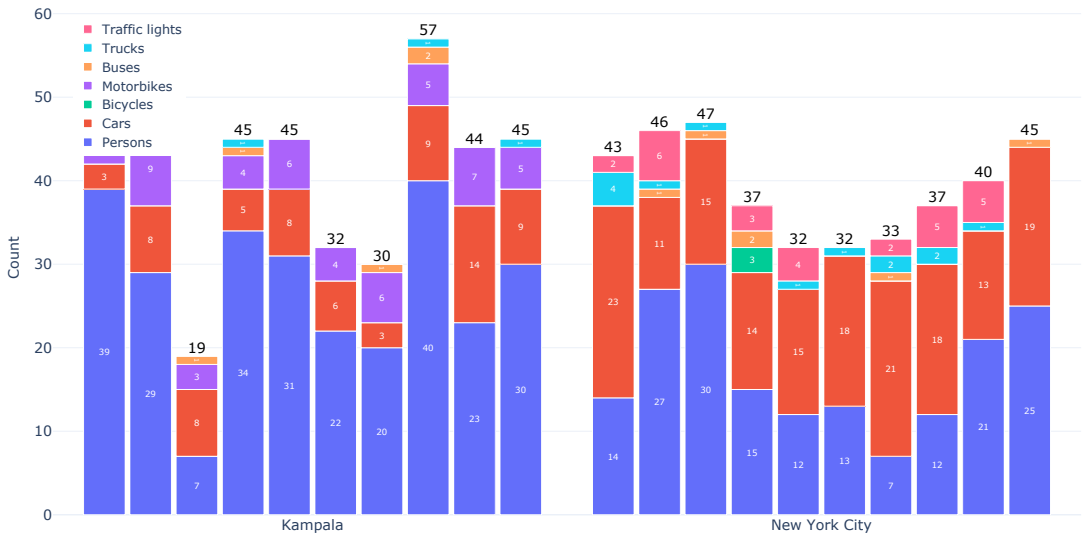


Fig. 3. Counts of objects detected by YOLOv11x for the 10 videos generated for each of New York City (USA) and Kampala (Uganda). Bars are ordered chronologically (oldest to newest), illustrating that Veo 3 can produce highly similar outputs for the same prompt across different generations

Fig. 3 illustrates the objects detected in the 20 generated videos, 10 each for New York City (United States) and Kampala (Uganda) and highlights the consistency in the representation of both cities. Despite some variability, all the videos for New York City showed a typical setting that could be linked to Manhattan. Three videos showed a skyscraper that resembles One World Trade Center. All New York City videos featured yellow taxis, one of the city's iconic symbols. All Kampala videos featured a similar urban setting, characterised by mid-rise buildings, relatively heavy traffic, and a lack of pedestrian infrastructure. Four of the videos generated for Kampala had background music, whereas this was not observed for the videos created for New York City.

The [redacted] authors independently reviewed half of the 76 videos ([redacted] coder was assigned 38 cities) and rated their perceived authenticity of synthetic traffic scenes using two questions. Fig. 4 shows the comparison of their ratings on Q1 ("Does the video show a kind of traffic scene that I would expect for this particular city"), Q3 ("Is the traffic density in the video in line with what I would expect for this particular city"), which focus on individuals' perceived authenticity of the traffic scenes. As we can observe, for Q1, there are conflicting opinions on the expected traffic for cities (2/12 in Africa, 3/21 in Europe, 2/27 in Asia, 1/4 cities in Oceania, and 1/7 cities in South America) on all continents except North America. For Q3, we disagreed on the expected traffic densities of cities in Africa (3/12), Europe (2/21), and Asia (6/27). We believe that this, on the one hand, may be caused by our uneven familiarity and misperception of the overall traffic scenes and traffic density of cities across various continents but, on the other hand, could reflect the less authentic nature of the synthetic traffic scenes in those cities. For example, none of the authors had visited Africa and all claimed to be uncertain about their votes. To justify the negative votes, the authors commonly commented "I expected to see a more developed city", "I expected to see different landscapes", or "I expected that the drivers would follow traffic laws in

this city". For Q3, the authors commented: "I had been there, there should be more pedestrians as it is a popular place to visit now, especially in this weather", "I expected to see less pedestrians as it is hot outside in this weather", "I expected the road to have more lanes, and the traffic density is higher as this is the capital of the country", etc.

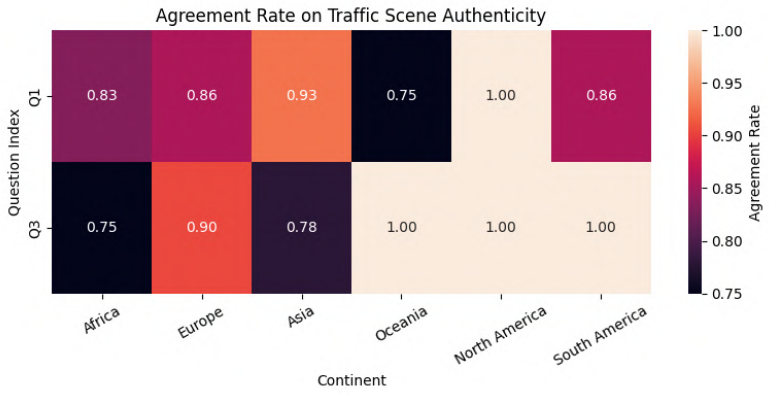


Fig. 4. A comparison between four independent raters for the perceived authenticity of the synthetic traffic scenes.

The authors also voted on "Does the video show a *pedestrian friendly* traffic scene?" (Q4). After a discussion, they all agreed that as long as synthetic videos show sidewalks on both sides and at least one zebra crossing line, the city is considered "pedestrian-friendly". Eventually, they found that 2/21 cities in Europe, 12/12 cities in Africa, 20/27 cities in Asia, and 1/4 cities in Oceania are not pedestrian friendly. All 5 cities in North America and 7 cities in South America are pedestrian friendly. Moreover, we noticed that the weather in almost all synthetic videos is sunny with a blue sky in daytime, even though the weather was not pre-defined in the prompts. The weather is overcast with a grey sky in the videos of 1/5 city in North America, 8/21 cities in Europe, 9/27 cities in Asia, 1/12 cities in Africa, and 1/7 cities in South America. Furthermore, all authors reported AI mistakes they observed in the videos. As presented in Table 1, we categorised AI mistakes in these traffic videos into 5 classes, highlighting the limitations of Veo 3 in handling non-English languages, traffic lines, signs and lights, two-way traffic, and vehicle generation and motion control.

4 DISCUSSION

Differences in object counts detected across cities and countries can be attributed to a combination of socioeconomic, cultural, and infrastructure factors, as seen in Figure 1. Cities with a higher number of pedestrians and vehicles, such as Kampala (Uganda), Mumbai (India), and Karachi (Pakistan), are typically characterised by high population densities and intense urban activity. In these locations, public spaces and roads are often heavily used, leading to greater visibility and frequency of people and vehicles in street imagery. In addition, the prevalence of certain transport modes reflects regional preferences, affordability, and urban planning. For example, motorbikes are much more common in cities such as Malé (Maldives) and Dhaka (Bangladesh), where two-wheelers provide an efficient and economical means of navigating congested urban environments, while cities such as Amsterdam (The Netherlands) stand out for bicycle usage, reflecting robust cycling infrastructure and cultural emphasis on sustainable transport.

In contrast, the low or absent counts of certain categories of objects in other cities highlight differences in urban form and lifestyle. Cities such as Dubai (UAE) and Muscat (Oman) record few or no pedestrian detections, probably

Theme	Description	Example Quote
Wrong language	llm disables to show any languages other than English correctly	"The facade shops' signs were supposed to be written in Arabic, but they are not readable", "all the letters are too blurry to read", "llm 'created' some characters but they are not correct at all"
Phantom road users	A road user appears and disappears suddenly, with aberrant movements	"ghostly persons", "Pedestrians disappeared", "a bicycle pops out of nowhere and the cyclist was riding in the middle of the road", "the car in the front of the ego vehicle appears and disappears", "car appeared out of another car"
Unrealistic motions	Road users' movements were unusual and violate the law of Physics	"cars' trajectories around the roundabout are wrong", "at the end of the street, the car looked like they are messed up", "car in wrong direction", "Horse movement was not ideal", "car ran through pedestrian"
Missing or Wrong traffic lines, signs and lights	Traffic lines and signs can be missing or wrongly painted	"The line in the middle should be double-white lines", "no traffic lines", "traffic lights are either blacked out or flickering", "the red and the green lights were blinking simultaneously", "there were multiple traffic lights for a single road, and all were blinking with different colors"
Only one-way street	almost all the videos show one-way streets, a few videos show two-way streets, but often with mistakes	"it looks like a one-way street but a car driving from the opposite direction just appeared from nowhere", "the opposite lane and cars appear all of a sudden, and the lane ends shortly", "three lanes in the same direction and only one lane in the opposite direction", "strangely slow traffic on the left. Not clear if it is moving or standing still."

Table 1. Thematic analysis on mistakes in generated videos.

due to low population density, climate conditions that discourage walking, or highly car-centric urban designs. The scarcity of buses, trucks, and bicycles in cities like Mumbai (India) and Malé (Maldives) may also indicate limited public transportation networks or greater dependence on private cars, shaped by local governance, economic conditions, and public policy. Furthermore, variations in detected traffic lights and related infrastructure reflect differences in road regulations and urban management.

The analysis of ten different videos each from New York City and Kampala demonstrates that Veo 3 is capable of generating realistic street scenes from a wide variety of global locations, regardless of a city's prominence or online visibility. Although the number of detected objects varied between different videos from the same city, this inconsistency may reflect natural fluctuations in street activity, such as changes in time of day, or may simply be due to the short duration of each video (just 8 seconds), which limits the range of traffic and urban features captured.

Interestingly, pedestrian infrastructure patterns may reflect genuine urban planning differences, with developing regions often prioritising vehicle infrastructure over pedestrian amenities, which Veo 3 appears to capture accurately. These findings underscore how object detection results not only mirror technical factors but also offer insight into the broader socioeconomic and cultural landscape of each city.

Despite some technical limitations noted in Table 1, the model achieved a high degree of authenticity, capturing distinctive features such as yellow taxis in New York and typical traffic densities. Audio analysis further revealed meaningful variation, with larger cities like Copenhagen (Denmark, -24.42 dBFS) displaying more complex soundscapes, suggesting Veo 3's ability to realistically represent urban traffic characteristics.

4.1 Limitations and Future Work

The current study utilises YOLOv11x for object detection, a deep learning model that can also yield inaccurate results. The authors also conducted a visual analysis to determine whether the videos developed were representative of the respective locations. Future studies may employ other video-generative models. As time progresses, new AI models will emerge with more sophisticated computational models, and more extended video generation capabilities will be utilised to produce videos. The evaluation process highlighted the difficulty of judging urban scenes from unfamiliar cities, as some authors were uncertain about traffic patterns in regions they had not visited. Future analysis can involve more human participants with diverse geographic backgrounds recruited via crowdsourcing platforms such as Appen (https://www.appen.com) or Amazon Mechanical Turk (https://www.mturk.com), as previously done by Bazilinskyy et al. [6, 7]. Future research may also encompass a more representative selection of locations, including the entire world. LLMs are also subject to potential training bias, which is a fundamental limitation of the approach.

REFERENCES

[1] Md Shadab Alam, Marieke H. Martens, Olena Bazilinska, and Pavlo Bazilinskyy. 2025. Understanding global pedestrian behaviour in 565 cities with dashcam videos on YouTube. (2025). Under Review.

[2] Md Shadab Alam, Marieke H Martens, and Pavlo Bazilinskyy. 2025. Generating Realistic Traffic Scenarios: A Deep Learning Approach Using Generative Adversarial Networks (GANs). In *13th International Conference on Human Interaction & Emerging Technologies: Artificial Intelligence & Future Applications, IHIET-AI 2025*. AHFE International, 349–358. doi:10.54941/ahfe1005927

[3] Md Shadab Alam, Marieke H. Martens, and Pavlo Bazilinskyy. 2025. Pedestrian Planet: What YouTube Driving from 231 Countries and Territories Teaches Us About the World.

[4] Md Shadab Alam, Sagar Hitendra Parmar, Marieke H. Martens, and Pavlo Bazilinskyy. 2025. Deep learning approach for realistic traffic video changes across lighting and weather conditions. In *2025 8th International Conference on Information and Computer Technologies (ICICT)*. Hilo, USA.

[5] Joseph Babcock and Raghav Bali. 2021. *Generative AI with Python and TensorFlow 2: Create images, text, and music with VAEs, GANs, LSTMs, Transformer models*. Packt Publishing Ltd.

[6] Pavlo Bazilinskyy, Dimitra Dodou, and J. C. F. De Winter. 2021. Visual attention of pedestrians in traffic scenes: A crowdsourcing experiment. In *Proceedings of International Conference on Applied Human Factors and Ergonomics (AHFE)*. Springer, New York, USA, 147–154. doi:10.1007/978-3-030-80012-3_18

[7] Pavlo Bazilinskyy, Dimitra Dodou, and J. C. F. De Winter. 2022. Crowdsourced assessment of 227 text-based eHMIs for a crossing scenario. In *Proceedings of International Conference on Applied Human Factors and Ergonomics (AHFE)*. New York, USA. doi:10.54941/ahfe1002444

[8] Pavlo Bazilinskyy, Y. B. Eisma, Dimitra Dodou, and J. C. F. De Winter. 2020. Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention* 21, 6 (2020), 347–353. doi:10.1080/15389588.2020.1762871

[9] Joseph R. Burger, Jordan G. Okie, Ian Hatton, Vanessa P. Weinberger, Munik Shrestha, Kyra J. Liedtke, Tam Be, Austin R. Cruz, Xiao Feng, Cesar Hinojo-Hinojo, Abu S. M. G. Kibria, Kacey C. Ernst, and Brian J. Enquist. 2022. Global city densities: re-examining urban scaling theory. arXiv:2210.08067 [physics.soc-ph] https://arxiv.org/abs/2210.08067

[10] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11621–11631. doi:10.48550/arXiv.1903.11027

[11] Wenpu Cao, Lei Dong, Lun Wu, and Yu Liu. 2020. Quantifying urban areas with multi-source data based on percolation theory. *Remote Sensing of Environment* 241 (May 2020), 111730. doi:10.1016/j.rse.2020.111730

[12] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1280–1289. doi:10.1109/CVPR52688.2022.00135

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3223. doi:10.48550/arXiv.1604.01685

[14] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2011. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence* 34, 4 (2011), 743–761. doi:10.1109/TPAMI.2011.155

[15] Patrick Ebel, Pavlo Bazilinskyy, Mark Colley, Courtney Michael Goodridge, Philipp Hock, Christian P. Janssen, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, and Philipp Wintersberger. 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* (Stanford, CA, USA) *(AutomotiveUI '24)*. Association for Computing Machinery, New York, NY, USA, 94–105. doi:10.1145/3640792.3675730

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 3354–3361. doi:10.1109/CVPR.2012.6248074

[17] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. 2020. One Thousand and One Hours: Self-driving Motion Prediction Dataset. arXiv:2006.14480 [cs.CV] https://arxiv.org/abs/2006.14480

[18] M Izani, A Assad, A Kaleel, L Wong, D Abdulla, and A Hamdan. 2024. Evaluating AI-Generated Video Quality: A Novel Assessment Model. In *International Symposium on Intelligent Computing Systems*. Springer, 54–66. doi:10.1007/978-3-031-82931-4_4

[19] Glenn Jocher and Jing Qiu. 2024. *Ultralytics YOLO11*. https://github.com/ultralytics/ultralytics

[20] Mohamed Nagy, Naoufel Werghi, Bilal Hassan, Jorge Dias, and Majid Khonji. 2025. RobMOT: Robust 3D Multi-Object Tracking by Observational Noise and State Estimation Drift Mitigation on LiDAR PointCloud. arXiv:2405.11536 [cs.CV] https://arxiv.org/abs/2405.11536

[21] Ken C. Pohlmann. 2010. *Principles of Digital Audio* (6th ed.). McGraw-Hill/Tab Electronics.

[22] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2446–2454. doi:10.48550/arXiv.1912.04838

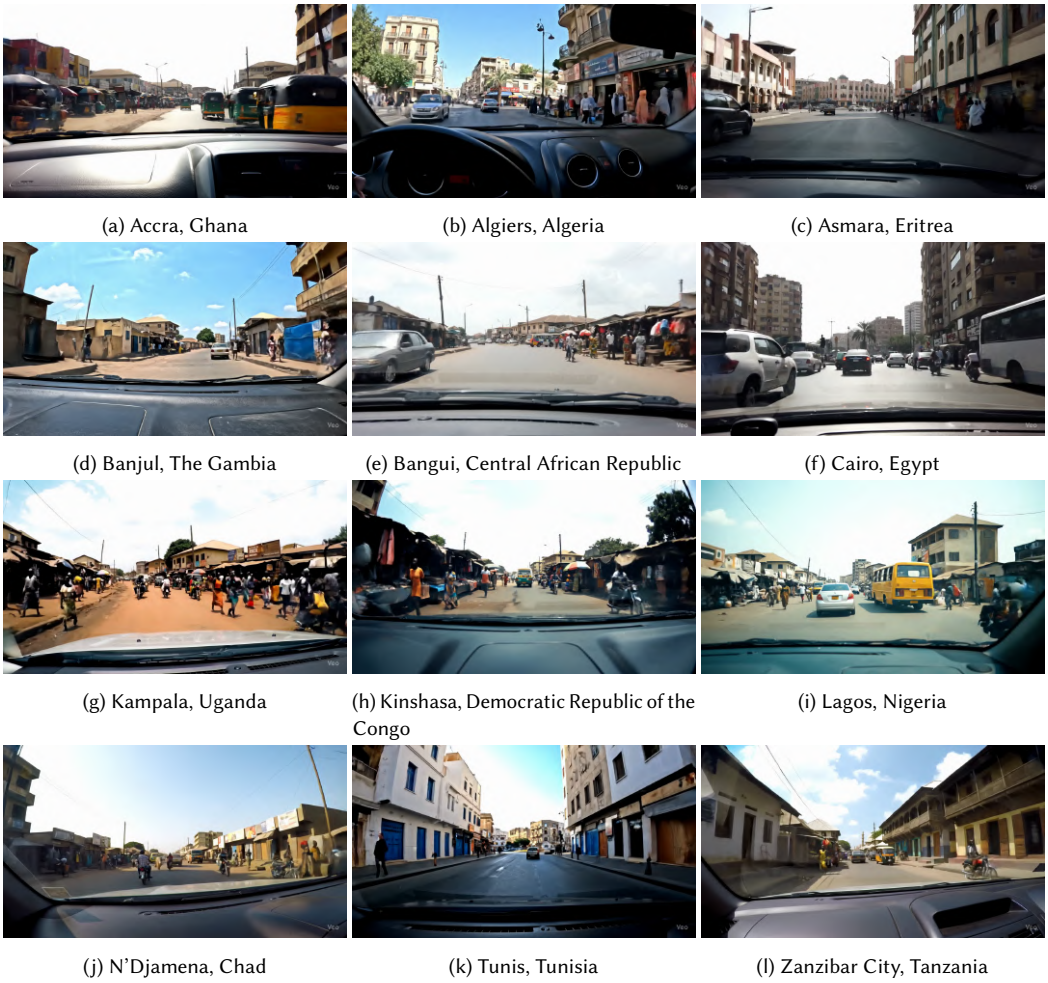
APPENDIX

A SUPPLEMENTARY MATERIAL

In line with current open science practices and recommendations for transparency in automotive user research [15], the authors openly provide these research artefacts to support reproducibility, collaboration and further advancements in the field. The generated videos, their description, and analysis code are available at https://www.dropbox.com/scl/fo/5cwpb5pxdel7415j4g1ol/AFJRjiNqJXtw-ZWKpcmKQDU?rlkey=8ij3ue3agbuk0wq7ltude1p5j. A maintained version of the code is available at https://github.com/Shaadalam9/llm-traffic-scene.

B FRAMES FROM THE GENERATED VIDEO

For each of the 76 cities included in this study, the first frame of the corresponding 8-second dashcam-style video generated by Veo 3 is shown below. The frames are grouped by continent (Africa, Asia I, Asia II, Europe I, Europe II, North America, Oceania, South America) and are arranged in a grid of subfigures. Each caption of the subfigure gives the city name and country, allowing a visual comparison of the model's ability to synthesise distinctive urban features in diverse global locations.



(a) Accra, Ghana (b) Algiers, Algeria (c) Asmara, Eritrea

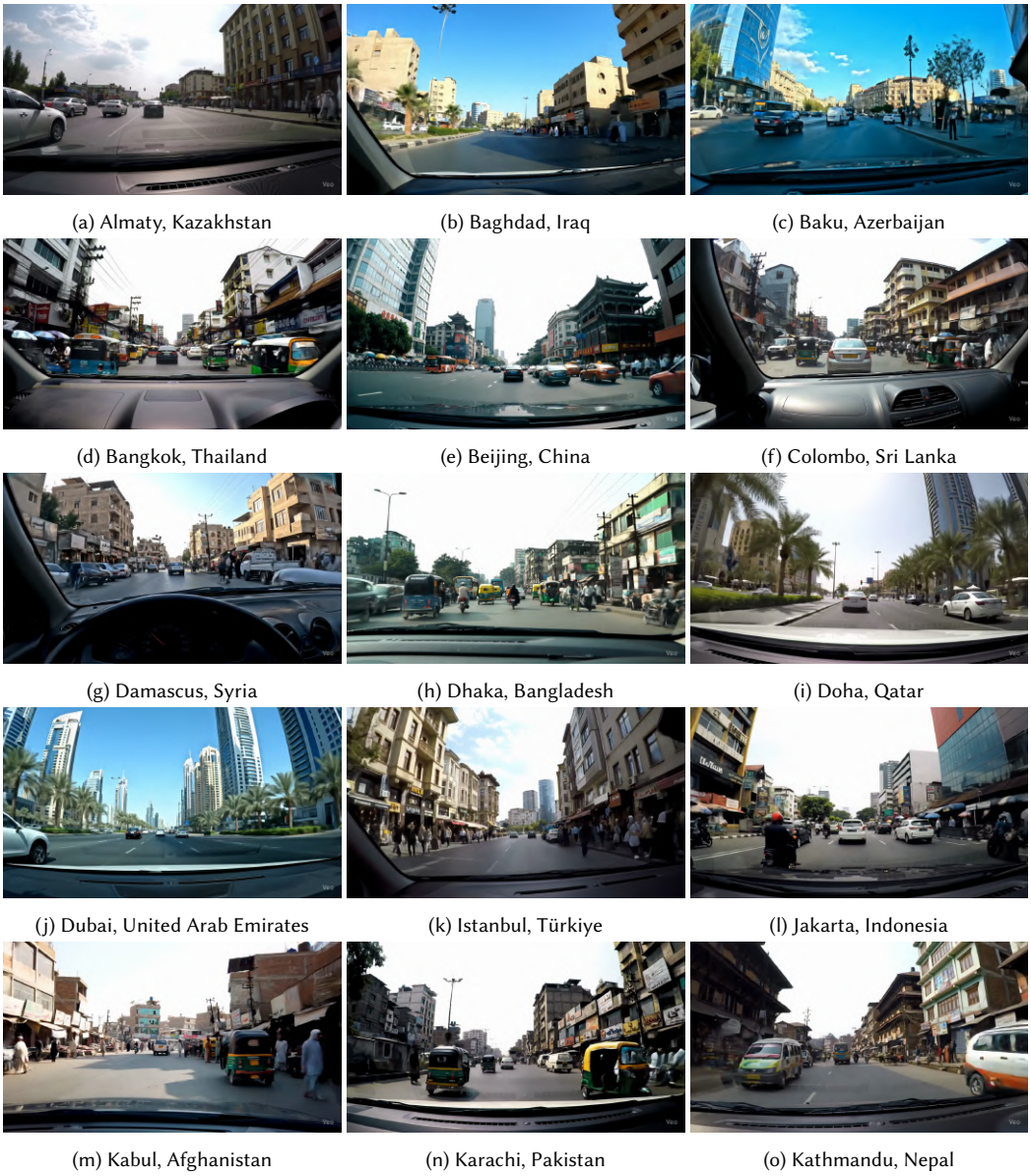
(d) Banjul, The Gambia (e) Bangui, Central African Republic (f) Cairo, Egypt

(g) Kampala, Uganda (h) Kinshasa, Democratic Republic of the Congo (i) Lagos, Nigeria

(j) N'Djamena, Chad (k) Tunis, Tunisia (l) Zanzibar City, Tanzania

Fig. 5. Frames from Veo 3-generated dashcam videos showing daytime urban traffic scenes in major African cities.

Received 19 June 2025; revised 19 July 2025; accepted 19 August 2025



(a) Almaty, Kazakhstan (b) Baghdad, Iraq (c) Baku, Azerbaijan

(d) Bangkok, Thailand (e) Beijing, China (f) Colombo, Sri Lanka

(g) Damascus, Syria (h) Dhaka, Bangladesh (i) Doha, Qatar

(j) Dubai, United Arab Emirates (k) Istanbul, Türkiye (l) Jakarta, Indonesia

(m) Kabul, Afghanistan (n) Karachi, Pakistan (o) Kathmandu, Nepal

Fig. 6. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in major Asian cities (l).

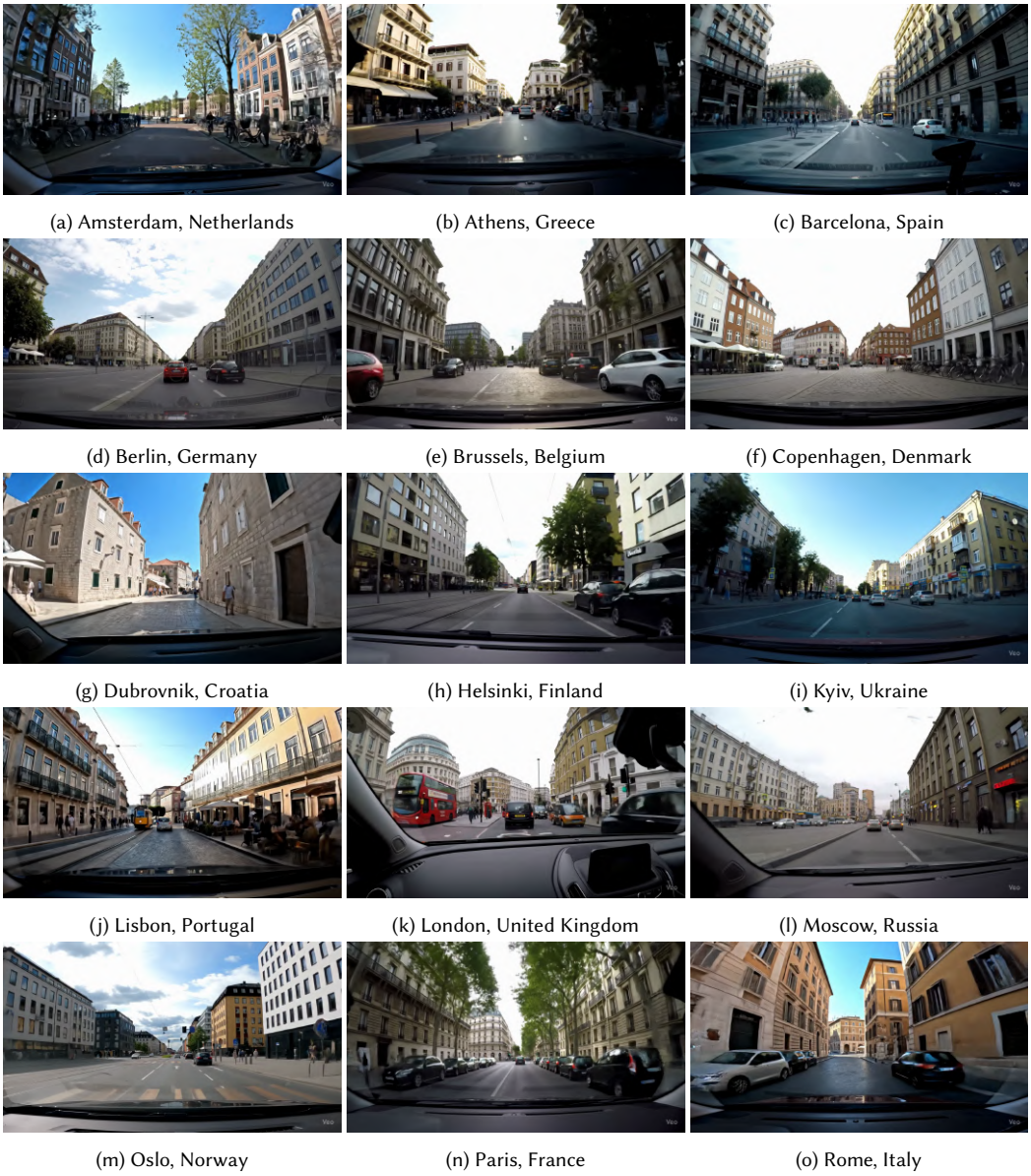
(a) Kuala Lumpur, Malaysia (b) Malé, Maldives (c) Mumbai, India

(d) Muscat, Oman (e) Phnom Penh, Cambodia (f) Pyongyang, North Korea

(g) Riyadh, Saudi Arabia (h) Seoul, South Korea (i) Tel Aviv, Israel

(j) Tehran, Iran (k) Tokyo, Japan (l) Yangon, Myanmar

Fig. 7. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in major Asian cities (II).



(a) Amsterdam, Netherlands (b) Athens, Greece (c) Barcelona, Spain

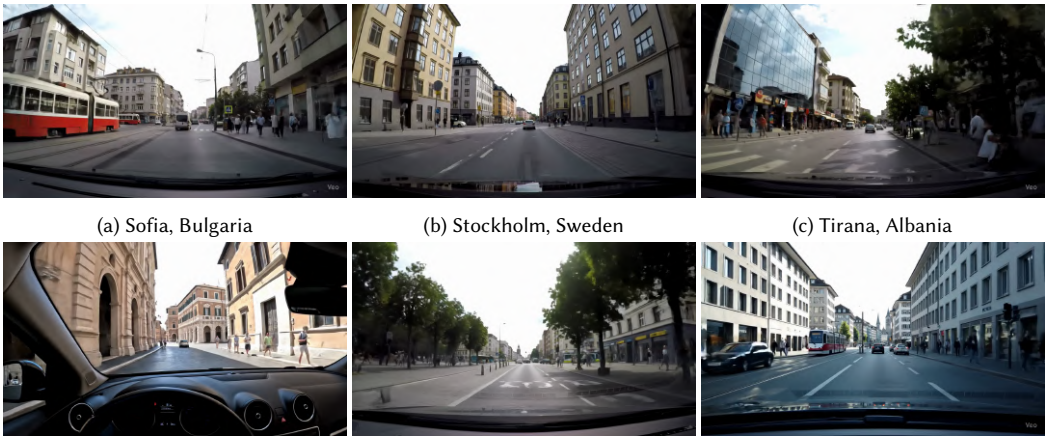
(d) Berlin, Germany (e) Brussels, Belgium (f) Copenhagen, Denmark

(g) Dubrovnik, Croatia (h) Helsinki, Finland (i) Kyiv, Ukraine

(j) Lisbon, Portugal (k) London, United Kingdom (l) Moscow, Russia

(m) Oslo, Norway (n) Paris, France (o) Rome, Italy

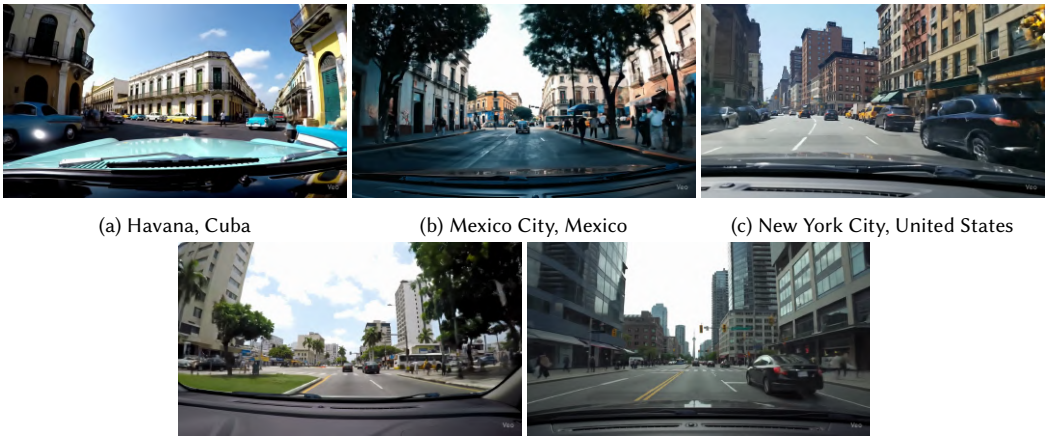
Fig. 8. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in major European cities.



(a) Sofia, Bulgaria (b) Stockholm, Sweden (c) Tirana, Albania

(d) Vatican City, Vatican (e) Warsaw, Poland (f) Zurich, Switzerland

Fig. 9. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in various European cities.



(a) Havana, Cuba (b) Mexico City, Mexico (c) New York City, United States

(d) Panama City, Panama (e) Toronto, Canada

Fig. 10. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in major North American cities.

(a) Auckland, New Zealand

(b) Funafuti, Tuvalu

(c) Suva, Fiji

(d) Sydney, Australia

Fig. 11. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in major cities across Oceania.

(a) Asunción, Paraguay

(b) Buenos Aires, Argentina

(c) Lima, Peru

(d) Montevideo, Uruguay

(e) Quito, Ecuador

(f) Rio de Janeiro, Brazil

(g) Santiago, Chile

Fig. 12. Frames from Veo 3-generated dashcam videos depicting daytime urban traffic scenes in major South American cities.