NOWHERE TO HIDE

Privacy Risks and Policy Implications of AI Geolocation

February 2026

privacyinternational.org

ABOUT PRIVACY INTERNATIONAL

Governments and corporations are using technology to exploit us. Their abuses of power threaten our freedoms and the very things that make us human. That's why Privacy International campaigns for the progress we all deserve. We're here to protect democracy, defend people's dignity, and demand accountability from the powerful institutions who breach public trust. After all, privacy is precious to every one of us, whether you're seeking asylum, fighting corruption, or searching for health advice.

So, join our global movement today and fight for what really matters:
our freedom to be human.

Photo by Fred Moon on Unsplash

Privacy International
62 Britton Street, London EC1M 5UY, United Kingdom
Phone +44 (0)20 3422 4321

privacyinternational.org

Privacy International is a registered charity (1147471), and a company limited by guarantee registered in England and Wales (04354366).

About the PRIV-LOC Project

Assessing and Mitigating Privacy Risks of Vision-Language Models in Image-based Geolocation Systems (PRIV-LOC) was a short project (2025-2026) led by Shoaib Ehsan at the University of Southampton, and funded by the UK AI Security Institute (AISI) that aimed to understand the geolocation capabilities of advanced AI models and the privacy risks arising from these. It asked four questions:

- How accurately are vision-language models able to identify locations from images?
- How much do we know and can we know about how these models are able to do this?
- What are the privacy implications of these capabilities and the incentives of the actors involved?
- How can these risks be mitigated?

The project was a collaboration between researchers at the University of Southampton, University College London, Queensland University of Technology, and Privacy International. Additional support came from the UKRI Responsible AI programme and the UKRI Generative AI Hub.

Authors

This report was compiled by Radhika Jhalani, Sarah Kiden, Karla Prudencio Ruiz, Julie Reintjes, Ilia Siatitsa, Jack Stilgoe, and Sania Waheed.

Acknowledgements

This report benefitted from a workshop in September 2025 at the UCL AI Centre with the following expert stakeholders: Mahsa Alimardani (Witness), Lucie Audibert (AWO), Anna Bacciarelli (Human Rights Watch), Tanu I (Article 19), Fanny Hidvegi (AI Collaborative), Kimberley Mai (AI Security Institute), Matt Mahmoudi (Amnesty International), Professor Lorna Macgregor (University of Essex), Foeke Postma (Bellingcat), Sir Bernard Silverman (former chief scientific adviser, UK Home Office), and Andrea Walker (BBC). The ideas and views in this report are our own and do not necessarily represent the views of the workshop attended by the expert stakeholders listed.

Table of Contents

Executive Summary

One of the surprising and concerning capabilities of the newest Artificial Intelligence (AI) models is geolocation from images. AI systems using Vision-Language Models (VLMs) have become capable of determining where in the world a photo is taken. Most people will be unaware that easily-accessible current AI systems can identify with speed and accuracy the location of their travel photos, even if those photos have their location information removed.

The ability to infer location from pictures without Global Positioning System (GPS) metadata has some benefits for those developing new robotics systems or conducting investigative journalism, but it also presents serious risks to privacy and other human rights. VLMs can transform an ordinary photograph into a source of personal information. There are immediate personal risks of covert surveillance and doxxing, discriminatory policing, or profiling. Beyond these individual harms, there are wider risks of chilling effects on freedom of assembly and enable social media platforms to monetise this location data.

This report comes from a project, PRIV-LOC, that asks: How good are the latest AI systems at geolocation? How much do we know about how these capabilities work? What are the risks? How much can/should these risks be mitigated? This report focuses on the risks. We explain our latest research on the geolocation capabilities of VLMs. We then explore, the risks posed by these, whether in the hands of expert users or ordinary members of the public. There is still a lot that we do not know. VLMs infer information from visual clues in images but, especially with closed-source, 'black box' AI systems, it is hard to know how. However, we have clear evidence of geolocation capabilities that raise both immediate and long-term concerns.

In this report, we identify some of the privacy risks that VLMs bring and organise them into those of direct identification and re-identification, tracking and location inference, training data and model release risks, memorisation and data

leakage, membership inference, model inversion and reconstruction, authoritarian surveillance, gendered threats, commercial exploitation, structural data biases, psychological and social impacts, dual use risks. In the last part, we review the legal and regulatory landscape and identify accountability gaps relevant to VLM regulation. Without harmonised approaches in understanding of the risk, challenges and solutions, the present uncertainty risks leaving people who currently share images online with nowhere to hide.

Introduction

Artificial Intelligence (AI) has evolved from text-generation models to integrate vision and natural language. Seeking to mimic humans' ability to make inferences from different sensory inputs, Vision-Language Models (VLMs) are becoming an increasingly common tool for understanding information.[1] By jointly processing both image and text data, VLMs are starting to show extraordinary capabilities in visual recognition, Visual Question Answering (VQA) and generating context-aware descriptions. Innovators are excited by the potential of VLMs in various applications, including robotics, medical imaging analysis, geospatial analysis, autonomous driving,[2] and visual accessibility for visually impaired individuals.[3] In addition, VLMs are also discussed as tools for investigative journalism, where they may assist in analysing images, verifying locations, and contextualising visual evidence.

However, these new models also raise pressing concerns beyond the risks that are becoming familiar with more common Large Language Models (LLMs). Visual data in images and videos contains identifying information and markers that can potentially reveal far more details than the users intended to share. When VLMs are combined with the powerful inference capabilities of other multimodal systems, they create new forms of privacy risks. The ability of VLMs to geolocate – to identify where in the world someone or something is – from sparse data has taken even experts by surprise.

---

1 Sim M. Y., Zhang W. E., Dai X., & Fang B., 'Can VLMs actually see and read? A survey on modality collapse in vision-language models', in: *Findings of the Association for Computational Linguistics* (Association for Computational Linguistics (ACL) 2025) 24452–24470 https://aclanthology.org/2025.findings-acl.1256.pdf

2 Kaduri O., Bagon S., & Dekel T., What's in the image? *A deep-dive into the vision of vision-language models*. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR 2025) https://openaccess.thecvf.com/content/CVPR2025/papers/Kaduri_Whats_in_the_Image_A_Deep-Dive_into_the_Vision_of_CVPR_2025_paper.pdf

3 Liu L., Yang D., Zhong S., Tholeti K. S., Ding L., Zhang Y. & Gilpin L. H., *Right this way: Can VLMs guide us to see more to answer questions? In Advances* in *Neural Information Processing Systems* (NeurIPS 2024) https://proceedings.neurips.cc/paper_files/paper/2024/file/efe4e50d492fedc0dfd2959f3320a974-Paper-Conference.pdf

New multimodal AI models with easy, free access give any user new access to rich information about location from images, inferring from subtle cues that computer scientists do not yet fully understand. The privacy risks, particularly to those vulnerable to persecution, discrimination, stalking or other forms of abuse, need careful exploration. And in a world in which social media platforms and other companies have access to billions of images and videos, the potential for automation of geolocation as part of a new mode of 'surveillance capitalism'[4] demands scrutiny. Our report is intended for researchers and policymakers who are engaged in emerging privacy challenges at the intersection of technology and governance, particularly in the field of VLMs.

The report comes from a project, PRIV-LOC, that asked: How good are the latest AI systems at geolocation? How much do we know about how these capabilities work? What are the risks? How much can these risks be mitigated? We focus on the privacy risks associated with VLMs. We begin by outlining our latest research on the geolocation capabilities of VLMs (Section 1). We then reflect on their broader transformative potential (Section 2) before examining the risks they pose, whether used by expert practitioners or members of the public (Section 3). Recognising that these risks do not arise in a regulatory vacuum, we identify some of the key legal and regulatory frameworks relevant to policymaking in this area (Section 4). Finally, we set out a series of forward-looking conclusions (Section 5).

4 Zuboff S., 'Big other: surveillance capitalism and the prospects of an information civilization', 30(1) *Journal of information technology* (2015) 75–89.

1. VLMs: An Overview

VLMs are AI models that blend computer vision and natural language processing (NLP) capabilities.[5] These models enable the analysis of images and videos and the generation of text-based outputs. A typical VLM, combining vision encoders with LLMs, can process and provide an advanced understanding of an image, video, or text input, producing descriptive captions, and answering questions about the content of the multimedia.[6] VLMs can identify objects in a photo, read text from images, recognise landmarks, and summarise information in text form.

Without having been explicitly designed for the purpose, VLMs are also demonstrating an emerging capability of image-based geolocation, inferring a photo's location by understanding and analysing visual cues present in it.[7] Images often contain contextual information that is correlated with specific places, such as architectural styles, road layouts, vegetation, climate-related features, signage, and cultural artefacts. During training, models implicitly learn to associate these visual cues and their relevant textual descriptions with geographic locations.[8] As a result, models can make accurate location inferences by leveraging subtle visual details.[9] For example, a model may be able to infer a place and even the angle from which an image is shot by recognising the position of the sun or analysing the shadow of the subject.[10]

---

5 Caballar R. D., & Stryker C., 'What are vision language models (VLMs)?' (IBM Think, 2024) https://www.ibm.com/think/topics/vision-language-models

6 Li Y., Lai Z., Bao W., Tan Z., Dao A., Sui K., Shen J., Liu D., Liu H. & Kong, Y., 'Visual large language models for generalized and specialized applications' (Version 1) (arXiv, 2025) https://arxiv.org/html/2501.02765v1; Radford A., Kim, J. W., Hallacy, C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., & Sutskever, I, 'Learning transferable visual models from natural language supervision' (International Conference on Machine Learning, 2021) https://arxiv.org/pdf/2103.00020v1

7 Wu M., & Huang Q., 'IM2City: Image geo-localization via multi-modal learning' Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, (2022) 50–61, https://doi.org/10.1145/3557918.3565868

8 Vivanco V., Nayak G. K. & Shah M., 'GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization', Advances in Neural Information Processing Systems (2023) https://arxiv.org/abs/2309.16020v2

9 *ibid*

10 Huang J., Huang J.-t., Liu Z., Liu X., Wang W. & Zhao J., 'VLMs as GeoGuessr masters — Exceptional performance, hidden biases, and privacy risks: Mind the photos you post: AI knows where you are!' (Version 2), ar5iv (arXiv) (2025) https://ar5iv.labs.arxiv.org/html/2502.11163v2

Conventional image geolocation relies on supervised learning approaches trained on large geo-tagged datasets or on large-scale image retrieval. These methods typically require extensive domain-specific training. Recent research[11] shows that general-purpose VLMs often match or exceed the performance of specialized geolocation models. As part of the PRIV-LOC project, our research team evaluated 25 state-of-the-art generative VLMs on standard geolocation benchmarks across multiple data distributions.[12] The results show that VLMs can achieve up to 60% accuracy at street-level localisation when evaluated on images resembling social media content. However, this performance degrades substantially when the data distribution shifts, such as when models are tested on Streetview-type images. We are starting to understand where and how these systems are particularly accurate geolocators, but there are still large uncertainties.

This ability of current VLMs to geolocate introduces new forms of imminent privacy risk. Unlike previous geolocation tools, which required specialised datasets, domain expertise, and controlled deployment, VLMs are widely accessible through AI interfaces such as Gemini, ChatGPT and Co-Pilot and they are general-purpose, with few checks on who can use them or how. Their broad availability, combined with high performance, means that sensitive information can be inferred at scale from everyday images, raising concerns about unintended surveillance, stalking, or misuse.[13] Additionally, many VLMs are closed 'black-boxes' (systems whose internal workings are not fully transparent or understandable), so while we can test their inferences, our understanding of their internal workings, training data, and biases will always be limited, restricting the potential for users, researchers, regulators or others to control or mitigate their behaviour.

11 Mendes E., Chen Y., Hays J., Das S., Xu W. & Ritter A., 'Granular privacy control for geolocation with vision language models' in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2024) 17240–17292, https://aclanthology.org/2024.emnlp-main.957/

12 Grainge O., Waheed S., Stilgoe J., Milford M. & Ehsan S., 'Assessing the geolocation capabilities, limitations and societal risks of generative vision-language models' in: *AAAI 2025 Fall Symposium Series: AI Trustworthiness and Risk Assessment for Challenged Contexts* (ATRACC) (AAAI Press, 2025) https://arxiv.org/abs/2508.19967

13 Mendes E., Chen Y., Hays J., Das S., Xu W. & Ritter A., 'Granular privacy control for geolocation with vision language models' in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2024) 17240–17292 https://aclanthology.org/2024.emnlp-main.957/

GPT GeoChat OSV5

Images from two datasets with different data distributions. The left column shows samples from the GPTGeoChat dataset, and the right column shows samples from the OSV5 dataset. Geolocation results are substantially more accurate for GPTGeoChat images, likely due to greater overlap between GPTGeoChat's data distribution and the VLMs' training data.

How VLMs geolocate

Most state-of-the-art VLMs are closed-source, which limits direct access to their internal representations, training data, or decision mechanisms. As a result, evaluation is restricted to inference-based analysis. Some newer AI models include elements sometimes referred to as 'reasoning' or 'chain-of-thought' and are able to produce forms of explanation for their responses (although such explanations should not be taken as perfect representations of how the models work). We analysed the text rationales produced by the models alongside their geo-localisation predictions. These rationales provide a useful proxy for understanding which semantic and spatial features are used for localisation.

We aggregated these rationales across multiple datasets using word clouds (Fig. 2) to identify recurring patterns. We also developed heatmaps (Fig. 3) highlighting regions of the image associated with these rationales. We observed that man-made structures, landmarks, signage, and symbolic elements such as flags are consistently highlighted, whereas natural scenes, such as vegetation, trees, and generic landscapes, contribute less towards localisation. This pattern aligns with prior findings in visual memorability[14] and place recognition[15], which show that distinctive, semantically meaningful elements play a central role in both human recall and place recognition. These results suggest that the geo-localisation capability of VLMs develops due to image-text alignment learned during training: by associating semantic visual cues with human generated textual descriptions. These findings indicate that VLM-based geo-localisation is driven primarily by semantic distinctiveness and language-alignment.

14 Isola P., Xiao J., Torralba A. & Oliva A., 'What makes an image memorable?' in: *Proceedings of the 2011 IEE Conference on Computer Vision and Pattern Recognition* (CVPR) (IEEE, 2011) 145-152, https://ieeexplore.ieee.org/document/5995721

15 Zaffar M., Ehsan S., Milford M. & McDonald-Maier K.D., 'Memorable Maps: A framework for re-defining places in visual place recognition', 22(12) *IEEE Transactions on Intelligent Transportation Systems* (2021) 7355-7369, https://doi.org/10.1109/TITS.2020.3001228

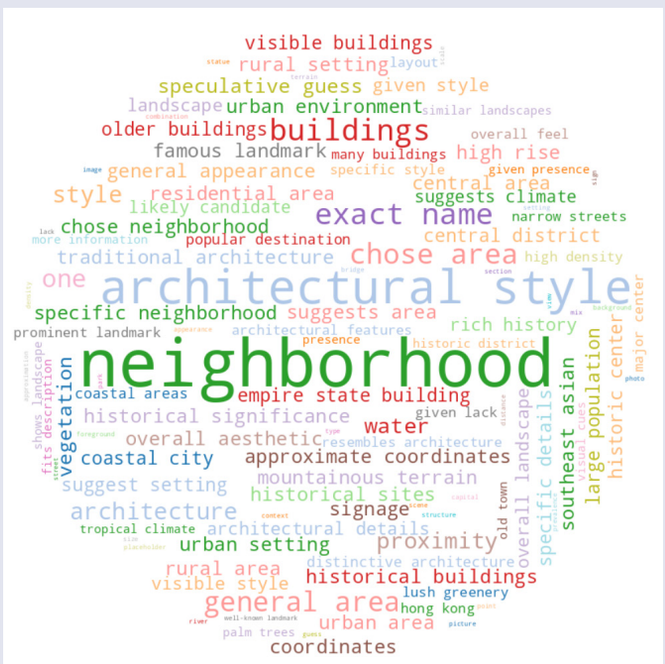

Fig 2: Word cloud of the text rationales produced by VLMs across multiple datasets. The size of each word reflects how frequently the model cited the corresponding visual cue when predicting image locations.

Heatmap1: Historic downtown district, Central square, Architectural style, American Flag, Clock tower

Heatmap2: Stadium structure, Green color, Baseball parks

Heatmap3: Bright building color, Sidewalk, Cultural vibes, Music venue, English text

Fig 3: Heatmaps highlighting regions of images that the VLM identifies as important for geolocation. Areas with higher intensity (red) indicate regions the model focuses on more when making predictions, while lower-intensity (blue) regions contribute less..

VLM geolocation v metadata geolocation

VLMs often have emergent capabilities for which they were not specifically programmed. For example, even if a model was trained merely to caption images, it may eventually learn to identify flags, landmarks, or understand scripts on signs as part of generating captions, which will enable it to infer location implicitly.

The image-based geolocation is different from traditional geotag or metadata-based location used by Open-source intelligence (OSINT) researchers. VLMs can infer location without resorting to GPS metadata, from the pixels and context of the image itself. When someone takes a photo, modern cameras and smartphones embed GPS coordinates, also known as geotags, which are a source of location information. Privacy-conscious users may either remove or obfuscate such metadata attached to multimedia content before sharing it.

With VLMs' inherent ability to geolocate from the contents of an image, it has become possible to trace to the source location even in the absence or stripping of the metadata. That means, image-based geolocation by VLMs is a form of derived data based on observation and inferences made by analysing the content of a visual. This derived data, based on inferences, stands in contrast to the provided data, which is either directly provided by a user or comes automatically attached to the content. Thus, VLM's capability to perform image-based geolocation effectively makes the entire visual content into personal location data.

2. The transformative potential of VLMs

VLMs can support complex tasks that combine visual and textual data. For example, they may assist open-source investigations to analyse images and link them with textual information. Investigative journalism groups begun to experiment with AI-assisted tools to support tasks such as identify leads, suggesting that these systems could assist in certain processes.[16] As we highlight in the next section, while such applications appear promising, they also raise questions about reliability, bias, and accountability that cannot be overlooked.

Similarly, embedding VLMs into drone navigation and robotic systems has been proposed to theoretically improve autonomy and context awareness.[17] Multimodal commands, combining text and camera input, could enable drones to respond to dynamic environments with improved situational understanding. These could potentially improve drone operations without needing extensive retraining.[18] But they also introduce concerns around misinterpretation, unintended behaviour, and ethical use in sensitive contexts.

VLMs have been tested in disaster scenarios to analyse aerial images to support search-and-rescue operations. Some studies report high accuracy rates in matching images to textual descriptions and identifying survivors under challenging conditions.[19] Similarly to the above however, while these figures suggest potential benefits for resource allocation and response speed, they do not eliminate risks

16 Belisario A., 'Easy AI: Classifying images with AI models [How-to guide]', Bellingcat (15 August 2024) https://www.bellingcat.com/resources/how-tos/2024/08/15/easy-ai-zero-shot-ai-image-classification-smart-image-sorter/

17 Ranasinghe Y., Vibashan V. S., Uplinger J., De Melo C. & Patel V. M., 'Zero-Shot Scene Understanding for Automatic Target Recognition Using Large Vision-Language Models' (arXiv:2501.07396) [Preprint] arXiv:2501.07396v1 [cs.CV] 13 January 2025, https://doi.org/10.48550/arXiv.2501.07396

18 Krupáš M., Urblík Ľ. & Zolotová I., 'Multimodal AI for UAV: Vision–Language Models Human–Machine Collaboration', 14(17) Electronics (2025) 3548 https://doi.org/10.3390/electronics14173548

19 For instance, according to a recent study, VLM-equipped UAV frameworks have achieved a 93% accuracy rate in matching images to textual descriptions and a precision of 90% in spotting survivors and recall of 85%, even under challenging conditions such as smoke and debris in high-risk zones. While these figures indicate promising performance, they should be interpreted cautiously given the complexity and variability of real-world environments. Khan F. A. & Shin S. Y., 'Real-time multimodal analysis for disaster management using UAVs and vision–language models' [Conference paper] (ResearchGate, February 2025) https://www.researchgate.net/publication/392940760

of false positives, automation bias (over-reliance on automated systems), or the implications of deploying AI in high-stakes environments.

Beyond emergency contexts, VLM-based systems are being explored for infrastructure monitoring and public safety.[20] They could in theory detect anomalies like intrusions or hazards and allow natural language queries of video feeds, reducing manual review time. For instance, there have been claims that such a system could be queried by asking it to show footage of people entering a restricted area after opening hours or get coordinates of potholes on a highway, broken guardrails, etc. Yet, these efficiencies come with challenges, including privacy concerns, susceptibility to adversarial inputs, and the possibility of reinforcing surveillance practices without adequate safeguards.

Finally, assistive technologies for people with visual impairments illustrate another area of potential benefit. VLMs could for example help interpret real-time video streams to warn users of obstacles or signage.[21] However, such applications depend on managing bias, ensuring accuracy, and addressing privacy risks—issues that remain unresolved.

In short, while VLMs may offer some advantages across domains such as journalism, conflict monitoring, disaster management, and accessibility, these benefits are neither guaranteed nor risk-free. Each example carries its own set of technical, ethical, and societal challenges, which must be critically examined alongside any claims of benefits. It is therefore important to remain cautious of overly positive advocacy that frames these technologies as inherently transformative, as such narratives can obscure limitations and amplify blind spots. This awareness provides a necessary foundation for the next section, which explores the taxonomy of threats and risks associated with VLMs.

---

20 Torneiro A., Monteiro D., Novais P., Rangel Henriques P., Rodrigues NF, 'Towards General Urban Monitoring with Vision-Language Models: A Review, Evaluation, and a Research Agenda' arXiv:2510.12400v1 [cs.CV] 14 October 2025, https://arxiv.org/html/2510.12400v1?utm

21 Yuan Z., Zhang T., Deng Y., Zhang J., Zhu Y., Jia Z., Zhou J. & Zhang J., 'WalkVLM: Aid visually impaired people walking by vision language model', arXiv:2412.20903 [cs.CV] submitted on 30 December 2024 (v1), last revised 4 March 2025 (this version, v4) https://doi.org/10.48550/arXiv.2412.20903.

How Accurate Are VLMs at Geolocation?

A model's geolocation ability is evaluated by the distance between the predicted coordinates and the true location of the subject. If a prediction falls within a certain error radius, it is deemed as "accurate". Most researchers define "accuracy" as the percentage of test photos for which a model can best guess within a radius of five spatial scales– street-level (1 km), city-level (25 km), region-level (200 km), country-level (750 km), and continent level (2,500 km).

A 2025 study evaluated the potential of the state-of-the-art VLMs in performing geolocation tasks. The models tested on a custom-built reasoning framework called ETHAN for improved performance could only guess 29% of the time the place within 1 km of the true spot. In other words, they missed the location by 71% of a 1km radius, by being off by an average of 499 kilometres.[22]

Even the existence of the present capability of VLMs wherein a model can predict accurate location with mere 20-30% probability is non-negligible, especially at Internet scale. A low accuracy rate still means millions of images could be accurately geolocated.

The results also revealed that these models were good at geolocating images resembling social media posts but poor on street-view or low-quality imagery. This asymmetry is likely to have stemmed from training data bias, as these models have been trained heavily on Internet-scrapped datasets and photos of famous sights or user-generated content compared to generic street panoramas. VLMs are trained to learn associations between images and text pairs scrapped from the Internet. Platforms such as Instagram, Flickr, and travel blogs provide rich training grounds for VLMs because they contain semantically dense, contextually anchored images with captions, hashtags, and engagement metadata. In contrast, street-view imagery lacks both semantic annotation and human-centred framing to draw associations with.

22 Liu Y., Deng D., Ding J., Li Y., Zhang T., Sun W., Zheng Y., Ge J., 'Mission Impossible- Image Based Geolocation with Large Vision Language Models', *Proceedings on Privacy Enhancing Technologies* 2025(4), 410-428, https://doi.org/10.56553/popets-2025-0137

3. Threat Taxonomy for VLMs: Privacy and beyond

VLMs bring many of the concerns that have already arisen with LLMs. As with other AI systems, LLMs do not just make mistakes, they are trained to please users. These systems tend to gauge a user's implied viewpoint and, as soon as prompts carry a nudge, they start leaning towards it, prioritising agreement over accuracy, a phenomenon now referred to as sycophancy.[23] The sycophancy in large and multimodal AI models arises because these systems are trained to maximise human approval signals rather than epistemic accuracy. A model that parses images or location can also exhibit the same optimisation. For instance, if along with a photo that carries coordinates of a predominantly Black neighbourhood, a user frames a narrative that goes, "this protest looks violent, right?" a model would tend to agree with it and reinforce the bias rather than testing the claim against evidence.[24] While this illustration is specific to contexts where the systemic racialisation of Black communities is predominant, this interpretive bias and the underlying mechanism applies to all social and spatial inequalities and stereotypes.

Furthermore, VLMs, like LLMs, are trained on vast datasets, which are often scraped from the Internet. These datasets may contain personal and sensitive information, such as pictures, videos, and personal identifiers. As a result, VLMs replicate—and even intensify—the same privacy and bias problems seen in large language models. Their dependence on massive datasets that encompass personal data renders serious privacy risks inherent to their operation.[25] The integration of visual inputs expands these concerns into new territory, creating distinct and increasingly complex privacy risks that are intrinsic to visual data processing. In ways, VLMs powered on datasets containing critical personal identifiers can be even more

23 Malmqvist L., 'Sycophancy in large language models: Causes and mitigations' (Version 1) arXiv:2411.15287v1 [cs.CL] 22 November 2024, https://arxiv.org/html/2411.15287v1

24 Article 19, 'Algorithmic people-pleasers: Are AI chatbots telling you what you want to hear?' (20 May 2025) https://www.article19.org/resources/algorithmic-people-pleasers-are-ai-chatbots-telling-you-what-you-want-to-hear/

25 Tömekçe B., Vero M., Staab R., & Vechev M., 'Private attribute inference from images with vision-language models', arXiv preprint arXiv:2404.10618, 4 November 2024, https://arxiv.org/pdf/2404.10618

privacy-invasive than LLMs, since even a single picture can expose a person by revealing their face, surroundings, and other contextual clues.

3.1. Illustrative Privacy Risks in the VLM Context

3.1.1. Direct Identification and Re-Identification Risks

Although VLMs are not equivalent to facial recognition technology (FRT), their ability to identify individuals introduces FRT-adjacent risks, often with broader and less predictable implications.[26] FRTs are typically understood as one-to-many identification or one-to-one verification systems: they match faces to names or confirm whether two faces belong to the same person.[27] VLMs, by contrast, can generalise across modalities and perform many-to-many inference tasks, potentially linking a person's image to a wider set of personal information available online.[28] Even background details captured in an image such as a painting, a certificate on a wall, or a bookshelf– may operate as quasi-identifiers that VLMs can use to infer or re-identify an individual. This shift from deterministic, purpose-specific technology to probabilistic, open-ended models raises important epistemic and regulatory questions, as VLMs shift from identity-based verification to open-ended interpretability.

The wider accessibility of VLMs considerably amplifies the risks for abusive uses of re-identification capabilities. Historically, advanced FRT has been tightly controlled by companies and state agencies;[29] by contrast, VLMs with image-analysis and location-inference capabilities can be widely accessible via public Application Programming Interface (APIs) or open-source releases. Individuals

---

26 Tömekçe B., Vero M., Staab R. & Vechev M., *'Private attribute inference from images with vision-language models', in: Advances in Neural Information Processing Systems* (NeurIPS 2024) https://proceedings.neurips.cc/paper_files/paper/2024/file/bb97e9a7c811904c9b01f51fde66edcf-Paper-Conference.pdf

27 National Institute of Standards and Technology (n.d.), '*Facial Recognition Technology (FRT)*' (6 February 2020) https://www.nist.gov/speech-testimony/facial-recognition-technology-frt-0

28 Sun W., Fan Y., Guo J., Zhang R. & Cheng X., *'Visual named entity linking: A new dataset and a baseline'*, arXiv preprint arXiv:2211.04872 [cs.CV] 9 November 2022 https://arxiv.org/pdf/2211.04872

29 For example, companies like Clearview AI claim that their services are restricted to law enforcement. Privacy International, Challenge against Clearview AI in Europe, https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe

– with little or no specialised expertise – can now automate interference at scale, including identity adjacent or geolocation predictions, using cues from background details (e.g. certificates, artwork, or bookshelves) as quasi-qualifiers.[30] This wider accessibility of tools both increases the opportunities for misuse and complicates regulatory oversight, underscoring the pressing need for governance frameworks.

As such tools become embedded in routine consumer applications, risks extend from interpersonal harms (e.g., stalking or evading abusive partners) to large-scale commercial profiling, where platforms infer sensitive attributes from user-generated images and monetise derived insights". For instance, social media companies may monetise location-based inference technologies without being classified as high-risk entities under most regulatory frameworks. Here, a structural gap emerges between the magnitude of potential harms and current legal frameworks: these deployments may not be classified as "risky" in law, yet they operate at the borders of risk-intensive inference, profiling, and behavioural prediction—underscoring the need for governance calibrated to accessibility, scale, and cross-modal inference.

3.1.2. Location Inference and Tracking Risks

A prominent privacy risk with VLM capabilities is that a user's location can be covertly obtained from innocuous images they share, even if they never intended to disclose it in the first place. For example, a person might upload vacation photos or daily life snapshots to a social media platform, thinking that they do not contain any personally identifiable information (PII). However, even seemingly harmless posts can be mined to reveal sensitive PII, putting individuals at risk without their knowledge. Location tracking capabilities have already been abusively exploited. For example, an investigation by the New South Wales Crime Commission revealed that tracking devices are used to monitor and harass victims of domestic and

30 Liu Y., Ding J., Deng G., Li Y., Zhang T., Sun W., Zheng Y., Ge J., & Liu Y., 'Image-based geolocation using large vision-language models', arXiv preprint arXiv:2408.09474 [cs.CR] 18 August 2024 https://arxiv.org/pdf/2408.09474

family violence.[31] There are also reports where devices like Apple AirTags have been used to track locations without the victim's knowledge.[32] Unlike GPS tracking, which usually requires access to a person's device, image-based tracking can be performed covertly on publicly available data.

VLM-based tracking, however, can achieve comparable intrusions without needing any direct access to a device or explicit data collection, such as real-time coordinates or embedded geotags in photos. A VLM deployed by an organisation or a third party could scrape images from the Internet and analyse them to extract precise or approximate location stamps. In doing so, a human or an AI model could aggregate inferred locations from multiple pictures posted over time by the user to map out the person's movements or routine. The systematic collection and aggregation of personal data, even from public sources, when done with the intent of identifying a person, amounts to an interference with the right to privacy.[33] Even if they individually are innocuous or inaccurate, there could be a 'mosaic effect' in which data points from VLM inferences are used alongside other forms of data to build up a privacy-infringing picture.

3.2. Training Data and Model Release Risks

3.2.1. Memorisation and Data Leakage

VLMs are trained on massive datasets scraped from the Internet. This risks cementing and reproducing personal data which could lead to privacy breaches through memorisation and training-data leakage.[34] Retention of personal data, especially images of identifiable persons, makes them effectively retain fragments of private life within their architecture and amount as such to interferences with the rights

31 New South Wales Crime Commission, 'Project Hakea: Criminal use of tracking and other surveillance devices in NSW' (2024) 26; Al-Khouri C., & Johnson K., 'NSW Crime Commission report finds tracking devices being used by domestic violence offenders', ABC News (24 June 2024) https://www.abc.net.au/news/2024-06-25/tracking-devices-domestic-violence-nsw-crime-commission-report/104017466

32 Crawford D., 'What is AirTag stalking and how to protect against it', Proton VPN (3 May 2024) https://protonvpn.com/blog/airtag-stalking

33 European Court of Human Rights, Rotaru v Romania, App No 28341/95, Judgment, 4 May 2000, https://hudoc.echr.coe.int/eng?i=001-58586

34 Jayaraman B., Guo C. & Chaudhuri K., 'Déjà Vu Memorization in Vision-Language Models', arXiv preprint arXiv:2402.02103 [cs.CV] 28 October 2024, https://arxiv.org/pdf/2402.02103

to privacy and data protection.[35] Information shared online or captured in public spaces may be publicly available, yet is still covered by privacy protections.[36] A person may willingly share private pictures with their followers on social media and yet reasonably expect that these pictures will not be repurposed by third parties to infer more private attributes, such as identity, location, or behaviour through VLMs. Essentially, public availability of data does not give presumption of consent for such secondary, unforeseen, and large-scale processing.

LLMs have also been shown to 'ingest and memorise' parts of their training data verbatim; hence, similar risks extend to multimodal AI models.[37] A public VLM could potentially be queried by prompt engineering by a malicious actor in a way to output a specific image or details related to a person.[38] A picture being publicly available should not give carte blanche to VLM developers to scrape, process or infer personal details from it in complete disregard to privacy norms, alongside intellectual property rights.[39]

3.2.2. Membership Inference Attacks

Even if the model has not fully memorised an image, malicious actors can attempt to detect if a specific image or data was used during its training. This is known as Membership Inference Attack. Such an attack comprises an attacker querying the model with a target image or a cropped part with recognisable features and observing how it responds.[40] If the model suggests a higher degree of familiarity or gives more precise outputs, that implies the image was indeed seen during the training.[41]

35 Kowalczuk A., Dubiński J., Boenisch F. & Dziedzic A., 'Privacy attacks on image autoregressive models', arXiv:2502.02514v1 [cs.CV] [Submitted on 4 February 2025 (this version), latest version 24 June 2025 (v4)] https://arxiv.org/abs/2502.02514v1

36 See section below Section 4 of this report. Also for further references see Privacy International, PI's Guide to International Law and Surveillance, 4th edition (March 2024) https://privacyinternational.org/report/5403/pis-guide-international-law-and-surveillance

37 Huang J., Yang D. & Potts C., 'Demystifying verbatim memorization in large language models' arXiv:2407.17817 [cs.CL] 25 July 2024 https://doi.org/10.48550/arXiv.2407.17817

38 Steele A., 'Understanding model memorization in machine learning', *Tonic.ai* (2024) https://www.tonic.ai/guides/understanding-model-memorization-in-machine-learning

39 Getty Images (US), *Inc. v Stability* AI, Inc., No. 1:23-cv-00135 (D. Del. Filed 3 February 2023).

40 Li Z., Wu Y., Chen Y., Tonin F., Rocamora E. A. & Cevher V., '*Membership inference attacks against large vision-language models*', arXiv:2411.02902v1 [cs.CV] 5 November 2024, https://arxiv.org/pdf/2411.02902

41 *ibid.*

Example

A photographer documents a political rally and posts an image of a protester holding a placard. Months later, they worry the image may have been scraped and used to train a VLM without consent. When they upload the photo—and cropped portions of it—to the model, it returns specific details such as the protester's name, the event's location, links to news reports, and inferred political views.

Such outputs suggest that the model has memorised information from its training data, including the original image and associated metadata. This means sensitive information—such as someone's presence at a political event—has been stored and replicated by an AI system without their knowledge. For individuals in fragile political contexts, these unintended disclosures can translate into real-world harms, including profiling, surveillance, harassment, or retaliation.

3.2.3. Model Inversion and Reconstruction

Another related threat with training data is model inversion, where an attacker having access to the model can reconstruct inputs on which the model was trained initially by exploiting its learned representations.[42] By analysing the model's internal activations or outputs in response to various prompts, one might reconstruct an image as closely as the model remembers.[43]

42 Fredrikson M., Jha S. & Ristenpart T., 'Model inversion attacks that exploit confidence information and basic countermeasures' in: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2015) 1322–1333, https://rist.tech.cornell.edu/papers/mi-ccs.pdf?

43 Zhou Z., Zhu J., Yu F., Li X., Peng X., Liu T. & Han B., 'Model inversion attacks: A survey of approaches and countermeasures' (Version 1.0) arXiv:2411.10023v1 [cs.LG] 15 November 2024 https://arxiv.org/html/2411.10023v1

Example

A hospital trains a diagnostic model on confidential patient skin-lesion images, ensuring the original photographs remain securely stored and never disclosed. Clinicians later access the system through an interface that provides confidence scores for different diagnostic labels.

An attacker, with access to this interface, uploads synthetic images and iteratively adjusts them to maximise the model's confidence for a particular patient label. After repeated iterations, they produce an image that closely resembles a real patient's lesion from the training set. Even without access to the underlying dataset, the model's learned representations allow sensitive medical information to be reconstructed—demonstrating how model inversion can compromise data confidentiality despite secure data handling practices.

3.3. Broader Misuse Risks of VLMs Pervasive Monitoring

People engaged in the ordinary course of public activities are routinely captured, and their information is, as a result of today's use of various data-driven technologies, permanently stored, monitored, and analysed. Empirical analysis of the computer vision field revealed that human-centred data extraction is not incidental but overwhelmingly dominant. According to a study, 90% of papers and 86% of downstream patents involved extracting human data, and the vast majority focus specifically on human bodies, body parts, and socially salient information. Much of this information is collected without the consent or awareness of the users.[44] This is done by obscuring the fact that humans are being treated as objects within these systems, as these systems describe their inputs in generic

44 Kalluri P. R., Agnew W., Cheng M., Owens K., Soldaini, L. & Birhane, A., 'Computer-vision research powers surveillance technology' 643(8070) *Nature* (2025) 73–79 https://www.nature.com/articles/s41586-025-08972-6

terms such as "images" or "objects".[45]

Such widespread collection, storage and aggregation of human centred visual data results in diminished autonomy of those depicted and limited ability to meaningfully opt out from these pervasive monitoring systems. Since VLMs are trained on large image and text datasets, they may inherit the same patterns of large-scale human data collection and result in broader risks such as unwanted identification, surveillance, gendered threats, doxxing, and exposure of other sensitive contexts.

3.3.1. Authoritarian Surveillance

The potential use of VLMs becomes particularly alarming in the case of governments and states regimes with authoritarian practices, where the fundamental human rights values and the rule of law are under attack. For example, a regime could exploit VLM-powered image analysis to comb social media for pictures from protests and civil society action, leading to automatic identification of persons involved and their locations. The chilling effect of such abuse on freedom of expression and assembly would be severe.[46] Concerns have already been raised about the use of facial recognition tools and other indiscriminate surveillance tools, such as International Mobile Subscriber Identity (IMSI) catchers[47] to identify who was attending a protest by authorities at protest sites.[48] The use of VLMs in

45 *ibid.*

46 UN Human Rights Committee, General Comment No 37 (2020) on the Right of Peaceful Assembly (Article 21), UN Doc CCPR/C/GC/37 (17 September 2020) para 10 https://tbinternet.ohchr.org/_layouts/15/treatybodyexternal/Download.aspx?symbolno=CCPR/C/GC/37&Lang=en; Report of the United Nations High Commissioner for Human Rights, Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, Including Peaceful Protests, UN Doc A/HRC/44/24 (24 June 2020) para 35 https://www.undocs.org/Home/Mobile?FinalSymbol=A%2FHRC%2F44%2F24&Language=E&DeviceType=Desktop&LangRequested=False

47 Privacy International, 'How IMSI catchers can be used at a protest' (5 May 2021) https://privacyinternational.org/explainer/4492/how-imsi-catchers-can-be-used-protest; Privacy International, 'IMSI catchers: facilitating indiscriminate surveillance of protesters' (19 June 2020) https://privacyinternational.org/news-analysis/3948/imsi-catchers-facilitating-indiscriminate-surveillance-protesters

48 Gabrielli G. 'The use of facial recognition technologies in the context of peaceful protest: The risk of mass surveillance practices and the implications for the protection of human rights', 16(2) *European Journal of Risk Regulation*, (2025) 514–541, https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/use-of-facial-recognition-technologies-in-the-context-of-peaceful-protest-the-risk-of-mass-surveillance-practices-and-the-implications-for-the-protection-of-human-rights/A4B2FABA8F32DDBC0217C86837CDBAC6

combination with existing facial recognition technologies could increase such risks, as people may share images themselves, not knowing that their geolocation may be obtained from these images.

3.3.2. Gendered Threats to Privacy and Safety

VLMs with geolocation capabilities pose acute risks of technology-facilitated gender-based violence (TFGBV). Ordinary or seemingly innocuous social media images can be processed with malicious intent, enabling doxxing, stalking, or real-time tracking.[49] Research has shown that even stock photographs can be precisely geolocated by GPT-4V, significantly heightening the risk of physical intrusion and other forms of targeted harm.[50] These dangers are not experienced equally. Groups already subject to widespread online abuse—including women, LGBTQ people, and other marginalised communities—face a disproportionate burden.

Unregulated geolocation via VLMs could compound these abuses by making them more personal, targeted, and harder to avoid. The privacy harm by VLMs gets further compounded by gendered context, as these technologies can intersect with existing gendered patterns of online abuse, harassment, or discrimination. According to the World Bank, fewer than 40% of countries have laws addressing cyberstalking and online harassment against women.[51]
A targeted survey by Amnesty International USA, Gay & Lesbian Alliance Against Defamation (GLAAD), and the Human Rights Campaign found that all LGBTQ respondents reported exposure to hateful or abusive speech online, with 60%

49 Mendes E., Chen Y., Hays J., Das S., Xu W. & Ritter A., 'Granular privacy control for geolocation with vision language models' in: Al-Onaizan Y., Bansal M. & Chen Y. (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, November 2024) 17240–17292 https://aclanthology.org/2024.emnlp-main.957/

50 Mendes E., Chen Y., Hays J., Das S., Xu W. & Ritter A., 'Granular privacy control for geolocation with vision language models', arXiv:2407.04952v1 [cs.CL] 06 July 2024 https://arxiv.org/html/2407.04952v1

51 UN Women, 'Digital violence is intensifying, yet nearly half of the world's women and girls lack legal protection from digital abuse', Press release (November 2025) https://www.unwomen.org/en/news-stories/press-release/2025/11/digital-violence-is-intensifying-yet-nearly-half-of-the-worlds-women-and-girls-lack-legal-protection-from-digital-abuse

saying this had a chilling effect on their social media use.[52] Further illustrating this risks, Human Rights Watch has documented how authorities in the Middle East and North Africa have used social media and dating apps to digitally target LGBTQ individuals.[53]

In this context, VLM-enabled geolocation could effectively weaponise personal images as tracking beacons—often without individuals realising that everyday photos may reveal their whereabouts or identities. For instance, there have been several documented instances of authorities in Egypt exploiting metadata and location tracking from dating apps to entrap and arrest gay and transgender users.[54] With VLMs, a photo of a pride march or an identified queer-safe space could be processed similarly to disclose someone's sexual orientation or gender identity against their will and expose them to targeted violence. Similarly, such models can infer a neighbourhood or home address of survivors of abuse, further endangering their safety.

3.3.3. VLM-Driven Harms: From Doxxing to Disinformation

VLMs can facilitate harassment, doxxing, and disinformation campaigns. The combination of deepfakes with VLM-based verification may be exploited for blackmail or to discredit individuals. VLMs can fuel misinformation through probabilistic inferences. Fake or AI-enhanced images could appear credible when a VLM links them to real locations or contexts.[55] Because VLMs analyse all visual elements–not only faces– they enable new forms of harm, and their widespread availability risks further normalising surveillance. This makes behaviours such as

52 Amnesty International, 'Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk' (9 February 2023) https://www.amnesty.org/en/latest/news/2023/02/hateful-and-abusive-speech-towards-lgbtq-community-surging-on-twitter-surging-under-elon-musk/

53 Human Rights Watch, '"All this terror because of a photo": Digital targeting and its offline consequences for LGBT people in the Middle East and North Africa' (21 February 2023) https://www.hrw.org/report/2023/02/21/all-terror-because-photo/digital-targeting-and-its-offline-consequences-lgbt

54 Privacy Guides, 'Queer dating apps: Beware who you trust with your intimate data' (24 June 2025) https://www.privacyguides.org/articles/2025/06/24/queer-dating-apps-beware-who-you-trust/

55 Huang T., Liu Z., Wang R., Zhang Y. & Jing L., 'Visual hallucination detection in large vision-language models via evidential conflict', *International Journal of Approximate Reasoning* (2025) 186. https://www.sciencedirect.com/science/article/abs/pii/S0888613X25001483

stalking, fabricating evidence, or conducting intrusions both easier to carry out and harder to detect.

VLMs also amplify the threat of doxxing by extracting or inferring sensitive, identifying details from images that people share online without realising the hidden cues they contain. By interpreting backgrounds, landmarks, objects, and subtle contextual information, a VLM can infer a person's location, workplace, daily routines, or relationships. This makes it significantly easier to unmask individuals or expose private information that may not be explicitly available anywhere else.[56] In effect, VLMs can transform ordinary photos into rich data sources that can be exploited for targeted harm.

3.3.4. Structural Biases in VLM Datasets

Large Internet-scraped datasets over-index on the Global North, affluent communities, and widely photographed landmarks. VLMs trained on such data inherit these skews, producing higher accuracy where imagery is plentiful and culturally dominant, and lower accuracy where it is sparse. This is not merely a technical artefact; it has distributive effects on those who benefit from reliable outputs and those who do not. Places in the Global Majority are often under-mapped and under-recognised, producing systematically poorer model performance.[57]

A different dynamic arises where communities are over-surveilled (e.g., historically targeted neighbourhoods).[58] Here, data abundance may reflect scrutiny rather than privilege. VLMs can perform better in these contexts not because they are better represented in a neutral sense, but because policing and monitoring practices

56 Mendes E., Chen Y., Hays J., Das S., XU W., Ritter A., 'Granular Privacy Control for Geolocation with Vision Language Models', arXiv:2407.04952v2 [cs.CL] 17 Oct 2024, https://arxiv.org/pdf/2407.04952; Luo W., Qiming Z., Lu T., Liu X., Zhao Y., Xiang Z., Xiao C., 'Doxing via the Lens: Revealing Privacy Leakage in Image Geolocation for Agentic Multi-Modal Large Reasoning Model', arXiv:2504.19373v1 [cs.CR] 27 Apr 2025, https://arxiv.org/html/2504.19373v1

57 Ramaswamy V. V., Lin S. Y., Zhao D., Adcock A. B., van der Maaten L., Ghadiyaram D. & Russakovsky O., 'GeoDE: A geographically diverse evaluation dataset for object recognition', arXiv:2301.02560v4 [cs.CV] 11 Sep 2025, https://arxiv.org/pdf/2301.02560

58 Ensign D., Friedler S. A., Neville S., Scheidegger C., Venkatasubramanian S., 'Runaway Feedback Loops in Predictive Policing', arXiv:1706.09847 [cs.CY] submitted on 29 June 2017 (v1), last revised 22 December 2017 (this version, v3)], https://arxiv.org/abs/1706.09847

generate dense, repeated coverage—risking feedback loops that normalise intensified attention and error-prone inference about already marginalised groups.

Importantly, simply "improving accuracy" does not by itself remove bias: in over-surveilled areas it can deepen unequal enforcement; in under-represented regions it can entrench inequities if the added data reflect dominant perspectives or extractive collection practices.[59] These distinct biases—privileged over-representation, targeted over-surveillance, and Global Majority under-representation—operate differently but interact in practice. Without safeguards, VLM-driven inferences risk reinforcing stereotypes, allocating errors unevenly, and directing suspicion towards marginalised communities.[60] Mitigations must therefore be tailored to each bias pattern (e.g., representational rebalancing, audit for surveillance feedback loops, and participatory data governance for under-represented regions) to uphold equality and non-discrimination.

3.3.5. Chilling Impacts of VLMs

The knowledge that any photograph may be analysed by a VLM for otherwise imperceptible information can prompt individuals to self-censor or become hyper-vigilant in their online expression. People may begin avoiding taking photographs in any recognisable locations or altering their behaviour to prevent unwarranted inferences, carrying the constant burden of anticipating how an AI system may interpret aspects of their lives. This anticipatory concern—whether grounded in the risk of misinterpretation or the simple possibility of being analysed—creates a chilling effect on individual freedoms, such as freedoms of expression, assembly and association, as well as personal autonomy.[61]

---

59 European Union Agency for Fundamental Rights, 'Bias in algorithms – Artificial intelligence and discrimination' (Publications Office of the European Union, 2022) https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

60 Li Y., Lai Z., Bao W., Tan Z., Dao A., Sui K., Shen J., Liu D., Liu H. & Kong Y., 'Visual large language models for generalized and specialized applications', arXiv:2501.02765v1 [cs.CV] 06 January https://arxiv.org/html/2501.02765v1

61 Report of the Office of the United Nations High Commissioner for Human Rights, The right to privacy in the digital age, UN Doc A/HRC/51/17 (4 August 2022), para 27 https://undocs.org/Home/Mobile?FinalSymbol=A%2FHRC%2F51%2F17&Language=E&DeviceType=Desktop&LangRequested=False

These chilling effects are compounded by uncertainty about who may later access an image or the inferences derived from it, how long such data will be retained and for what secondary purposes it may be used. As a result, individuals may refrain from capturing or sharing moments from meaningful places or contexts, not only because of how an AI system might interpret them, but because they cannot realistically predict where their images or inferred data may circulate, be reused or repurposed. The privacy risks posed by VLMs therefore extend far beyond isolated data points: they reshape the broader environment of surveillance, undermining a person's ability to exercise their human rights and participate freely in public and digital life. As explained before, this chilling effect can also exacerbate more targeted harms—such as doxxing, discrimination, and gender-based targeting—further eroding the security and dignity of those affected.

3.3.6. Commercial Exploitation

Corporations can deploy VLMs to profile consumers by inferring demographics, interests, purchasing habits, or lifestyle characteristics from images, and then use these inferences for targeted advertising. In October 2025, Meta announced that it would use users' interactions with its AI features to refine ad targeting.[62] Although the announcement focused on text and voice interactions, it signals a future in which photo content may similarly be mined to generate behavioural or interest-based signals.

Example

A social media platform introduces an "auto-tag" feature for photos. A VLM trained on large Internet-scraped datasets analyses a user's image of herself hiking near her home. The model labels her an "outdoors enthusiast," identifies nearby landscapes, and links this information to her advertising profile. She soon begins receiving targeted ads for camping gear and premium travel packages—even though she never disclosed these interests.

62 Meta, 'Improving your recommendations on our apps with AI at Meta' (1 October 2025) https://about.fb.com/news/2025/10/improving-your-recommendations-apps-ai-meta/

These developments raise significant concerns around data privacy, fairness, and meaningful consent. Companies are required to limit data collection to what is necessary and to inform users about how their data is processed. Yet inferences derived from images often bypass these safeguards, as inferred data is frequently treated differently from explicitly provided data. With users becoming more aware of some privacy risks, companies may increasingly rely on inference-based profiling from visual content. Monetisable geolocation or behavioural predictions may seem less immediately harmful than interpersonal misuse, but at scale they can create new forms of exploitation—turning ordinary images into sources of commercially valuable surveillance.

3.3.7. Dual-Use and Military Repurposing

The geolocation capabilities of VLMs make them a dual-use technology. Tools developed for benign, commercial or civilian purposes can be readily repurposed within military and intelligence contexts.[63] VLMs are increasingly integrated into defence and security systems,[64] where their ability to extract location and contextual information from images supports surveillance, targeting, and operational planning. This forms part of a broader trend of militarising data-intensive technologies, in which civilian digital infrastructure—including social-media platforms and cloud services—is intertwined with military operations,[65] enabling new modalities of warfare and real-time battlefield intelligence.[66]

Embedding VLMs into military-national security systems can facilitate automated threat detection, population-level profiling and drone-targeting workflows. AI systems, including VLMs, are reshaping the conduct of warfare by dramatically expanding surveillance capacities, accelerating intelligence analysis and

63 Sayler K. M., 'Artificial intelligence and national security' (CRS Report No. R45178), Congressional Research Service (10 November 2020) https://www.congress.gov/crs-product/R45178

64 Graham D., 'Inside Project Maven, the U.S. military's AI project', Bloomberg (20 February 2024) https://www.bloomberg.com/news/newsletters/2024-02-29/inside-project-maven-the-us-military-s-ai-project

65 Privacy International, 'How data drives the militarisation of tech' (12 September 2025) https://privacyinternational.org/long-read/5667/how-data-drives-militarisation-tech.

66 Simpson K. H., Paquette S., Racicot R. & Villanove S., 'Militarising AI: How to catch the digital dragon?', Centre for International Governance Innovation (26 February 2025) https://www.cigionline.org/articles/militarizing-ai-how-to-catch-the-digital-dragon/

identifying patterns across vast civilian datasets.[67] These systems are designed to process large volumes of imagery and metadata, producing "actionable insights" that can influence decisions with profound consequences for human rights.

This blurring of civilian and military uses raises significant concerns for privacy and data protection.[68] The principle of purpose limitation—requiring personal data to be used only for the purposes for which it was collected—becomes difficult to uphold when technologies developed for civilian applications are transferred into military systems without the original safeguards or consent frameworks. As civilian-generated data becomes embedded in security infrastructures, surveillance risks are normalised and individuals' rights to privacy, autonomy and freedom from arbitrary interference are further diluted.

67 Privacy International, 'Big Tech's bind with military and intelligence agencies' (1 October 2025) https://privacyinternational.org/long-read/5683/big-techs-bind-military-and-intelligence-agencies

68 Privacy International, 'Blurring the Line: How Militarisation of Tech is Reshaping our Town Squares' (12 September 2025) https://privacyinternational.org/long-read/5669/blurring-line-how-militarisation-tech-reshaping-our-town-squares

4. Legal and Policy Landscape

The risks associated with VLMs—including identity-adjacent inference, location derivation, re-identification, and discriminatory profiling—do not arise in a regulatory void. Although VLMs represent new technological capabilities, their deployment is already governed by well-established legal frameworks in the United Kingdom, Europe and internationally. In the UK, the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018[69] constitutes the core data-protection regime, closely aligned with the EU GDPR.[70] These sit alongside privacy rights under Article 8 of the European Convention on Human Rights (ECHR)[71] and broader international human-rights obligations under the International Covenant on Civil and Political Rights (ICCPR),[72] which prohibit arbitrary or unlawful interference with private life. Together, these frameworks regulate personal-data processing VLMs perform when it affects privacy, equality, non-discrimination or other fundamental rights.

4.1. Data-Protection Frameworks Governing VLM Data

Under the UK GDPR (mirroring Article 4(1) GDPR), personal data includes any information relating to an identified or identifiable natural person.[73] This scope includes VLM inputs, outputs, and inferred data. Regulators have consistently held that information qualifies as personal data even where different pieces must be

69 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (United Kingdom General Data Protection Regulation) (UK version, as amended) https://www.legislation.gov.uk/eur/2016/679/contents; Data Protection Act 2018 c. 12, https://www.legislation.gov.uk/ukpga/2018/12/contents

70 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), [2016] OJ L119/1, https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (hereinafter EU GDPR)

71 Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended by Protocols Nos 11 and 14, opened for signature 4 November 1950, entered into force 3 September 1953) 213 UNTS 221, https://www.echr.coe.int/documents/d/echr/convention_ENG

72 International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171, https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

73 See Articles 2(1), 4(1) GDPR and Recitals 15, 26, and 30 GDPR, note above.

combined to identify a person.[74] The UK Information Commissioner's Office (ICO) emphasised that identifiability may be direct or indirect depending on context and available auxiliary information.[75] As such, inferred data is protected: EU regulators consider inferences and predictions to be personal data whenever they relate to an identifiable person.[76]

Both the UK and EU legal frameworks, explicitly list location data as a form of identifier, reinforcing that location inference constitutes personal data when tied to an identifiable person.[77] Additionally, VLMs outputs become biometric data—a special category requiring enhanced protection—when they derive, verify or confirm identity using bodily characteristics. For example, VLM-generated "faceprints" used to recognise individuals fall within GDPR's biometric-data definition.[78]

74 European Data Protection Board, 'What is personal data' https://www.edpb.europa.eu/sme-data-protection-guide/faq-frequently-asked-questions/answer/what-personal-data_en; European Data Protection Supervisor (EDPS) & Agencia Española de Protección de Datos (AEPD), '10 Misunderstandings Related to Anonymisation' (27 April 2021) 5–7, https://www.edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf

75 ICO, 'Can we identify an individual indirectly from the information we have (together with other available information)?' https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/can-we-identify-an-individual-indirectly/

76 Article 29 Working Party, Opinion 4/2007 on the Concept of Personal Data (WP136, 20 June 2007) 6–10; European Data Protection Supervisor (EDPS) & Agencia Española de Protección de Datos (AEPD), '10 Misunderstandings Related to Anonymisation' (27 April 2021) 5–7; European Data Protection Board (EDPB), Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (endorsed 25 May 2018).

77 ICO, 'What are identifiers and related factors?', https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-are-identifiers-and-related-factors/; International definitions follow a similar approach, including India's Digital Personal Data Protection Act and Brazil's LGPD, both of which cover identifiable data and sensitive categories such as biometrics. The Digital Personal Data Protection Act, 2023 (No. 22 of 2023) https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf; Brazilian Data Protection Law (LGPD) (As amended by Law No. 13,853/2019) https://www.gov.br/anpd/pt-br/centrais-de-conteudo/outros-documentos-e-publicacoes-institucionais/lgpd-en-lei-no-13-709-capa.pdf

78 ''biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data;' Article 4(14) GDPR, note above.

When is an image "personal data"? Context is decisive. A landscape photo with no person and no unique marker is typically not personal data. However, an image depicting a person (even partially), or one that can reasonably be linked to a particular individual through other available information (such as a car number plate or a distinctive home frontage), is personal data.

As a result, images can constitute personal data even in the absence of a clearly visible face, if individuals are recognisable through bodily features, surroundings or other contextual cues. Location information also qualifies as personal data when it can be linked with reasonable specificity to an individual, whether taken directly from metadata or inferred by a VLM. Consequently, processing such outputs requires a lawful basis and compliance with transparency, purpose-limitation and, where relevant, sensitive-data restrictions.

Example

A travel blogger posts a sunset photo taken from her apartment balcony. The image contains no geotag or metadata, and she assumes it is anonymous. A VLM analyses the photo, identifies distinctive skyline contours and environmental cues, and infers precise GPS coordinates corresponding to her apartment. Although the picture contains no explicit identifier, the VLM's inference reveals her home location. This output qualifies as personal location data because it indicates where the photo was taken and can be linked to an identifiable person.

4.2. Privacy and other Human-Rights Constraints on VLM Inference

International human-rights protections complement data-protection obligations. Article 17 ICCPR[79] and Article 8 ECHR[80] require states to prevent unlawful or disproportionate privacy interferences, and the European Court of Human Rights has held that repurposing publicly available images can still engage privacy rights.[81] These principles apply directly to VLMs: mass scraping of images, extraction of biometric identifiers, geolocation inference and secondary analysis of public photos may infringe privacy unless justified as lawful, necessary and proportionate. Regulatory actions against Clearview AI across several European jurisdictions illustrate these limits.[82]

A privacy lens is necessary but not sufficient for regulating VLMs. Beyond this, other human rights standards become relevant for VLMs regulation. A rights-based approach should address wider impacts: freedom of expression and association, affected by content generation, moderation and recommendation; equality and non-discrimination, implicated by biased outputs; and access to information and participation, shaped by systems that influence visibility and discoverability, among others. Under the UN Guiding Principles on Business and Human Rights, VLM providers should conduct ongoing human-rights due diligence, ensure meaningful transparency and offer effective remedies for harms.[83] Embedding

79 ICCPR, note above.

80 ECHR, note above.

81 The European Court of Human Rights (ECtHR) has consistently held that privacy rights apply even where information originates from public environments, and that the re-use of public images for new, more intrusive purposes can still constitute an interference. ECtHR, *Peck v The United Kingdom*, App nos 44647/98, Judgment, 28 January 2003 https://hudoc.echr.coe.int/eng?i=001-60898; ECtHR, *Perry v The United Kingdom,* App no 63737/00, Judgment 17 July 2003, https://hudoc.echr.coe.int/eng?i=001-61228. In another case now pending before the same court, the Court is posed with the legal question whether the mere processing of a person's social media data which is publicly available by public authorities constitutes an interference with private life. ECtHR, BUTT v The United Kingdom, App no 32946/20, pending, https://hudoc.echr.coe.int/eng?i=001-205953. See also Privacy International, Third-party intervention in Butt v. The United Kingdom (App no 48821/15) https://privacyinternational.org/sites/default/files/2021-06/PI%20Intervention%20-%20Butt%20v.%20UK-Final.pdf

82 European regulators have repeatedly ruled that scraping publicly available images for biometric profiling is unlawful without a valid basis, as illustrated by enforcement actions against Clearview AI in five jurisdictions, including the UK, Austria, Greece, Italy, and France, which found violations of data-protection principles. For an overview of cases and updates, see Privacy International, 'Challenge against Clearview AI in Europe', https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe

83 UN Human Rights Council, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (21 March 2011) UN Doc A/HRC/17/31, https://www.ohchr.org/sites/default/files/Documents/Issues/Business/A-HRC-17-31_AEV.pdf

these duties, alongside legality, necessity and proportionality requirements, supports governance frameworks that address systemic—not merely technical—risks.

4.3. Interaction with Emerging AI Regulation

Emerging AI-specific regulation adds further obligations. The EU Artificial Intelligence Act (2024) introduces a risk-based regime requiring high-risk systems to implement risk management, documentation, human oversight and transparency measures.[84] It also mandates labelling for AI-generated or manipulated content and requires fundamental-rights impact assessments for certain high-risk deployments.[85] The UK has opted against a unified AI Act, instead relying on sectoral regulators—including the ICO, EHRC, Competition and Markets Authority (CMA) and Ofcom—to apply existing rights-based laws.

4.4. Accountability Gaps

There are significant difficulties in assigning responsibility for VLM governance because these systems are frequently developed and deployed through a fragmented "patchwork" of actors.[86] Regulatory approaches assume clear lines between developers, providers and users, but VLMs—particularly open-source models—circulate through far more diffuse chains. VLMs with geolocation capabilities may be downloaded as open-source code and fine-tuned by intermediaries and embedded into countless downstream applications.[87] This diffusion of agency makes accountability elusive: once model weights are released, they can be reused, forked, or combined with external datasets.

---

84 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689.

85 Section 2: Requirements for High-Risk AI Systems, EU AI, ibid.

86 National Telecommunications and Information Administration, '*Ensure accountability across the AI lifecycle and value chain.* U.S. Department of Commerce' (27 March 2024) https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/requisites-for-ai-accountability-areas-of-significant-commenter-agreement/ensure-accountability-across-the-AI-lifecycle-and-value-chain

87 National Telecommunications and Information Administration, 'AI open models: Opportunities and risks' (U.S. Department of Commerce, 2023) https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

This has been described as both a "responsibility gap" and the "problem of many hands": harmful outcomes arise from the cumulative actions of multiple actors, yet no single individual or organisation can be readily held accountable.[88] Research on algorithmic supply chains shows that responsibility flows across the full lifecycle—from those who build and pre-train foundation models, to actors who fine-tune, deploy, and operate them—mirroring the movement and recombination of data and model components.[89] In open-source VLMs, deployers may also attempt to deflect responsibility precisely because the model is "open," even when they shape the context of use.

Regulatory initiatives such as the EU AI Act attempt to clarify the division of responsibilities by imposing distinct duties on providers and deployers—covering risk management, documentation, post-market monitoring, logging, and incident reporting for high-risk systems.[90] Yet gaps remain in terms of enforceability, proportionality and auditability, especially when AI components are reused across borders and integrated into new systems without formal notification or oversight.[91]

Compounding this, accuracy failures alone rarely discourage adoption. Public and private actors often procure recognition or inference tools despite high error rates, tolerating inaccuracy until visible harm materialises.[92] With open-source VLMs, diffuse responsibility across training, fine-tuning and deployment makes it even less likely that any single vendor can be held accountable. Strengthened governance therefore requires more robust procurement rules and minimum safeguards, including disclosures about models and datasets, error rates and precision limits, and auditable documentation capable of tracing how system components are assembled and used.

---

88 Königs P., 'Artificial intelligence and responsibility gaps: What is the problem?' 24 *Ethics and Information Technology* 36 (2022) https://doi.org/10.1007/s10676-022-09643-0

89 Cobbe J., Veale M. & Singh J., 'Understanding accountability in algorithmic supply chains' in: Proceedings of the 2023 *ACM Conference on Fairness, Accountability, and Transparency* (FAccT'23) (ACM, 2023) 1186–1197, https://doi.org/10.1145/3593013.3594073

90 Article 26: Obligations of deployers of high-risk AI systems, EU AI Act, note above.

91 Finch W. W. & Butt M, 'Gaps in AI-Compliant Complementary Governance Frameworks' Suitability (for Low-Capacity Actors), and Structural Asymmetries (in the Compliance Ecosystem)—A Systematic Review', 5(4) *Journal of Cybersecurity and Privacy* (2025) 101, https://doi.org/10.3390/jcp5040101

92 Hickok M. & Hu E., 'Don't let governments buy AI systems that ignore human rights', 40(3) *Issues in Science and Technology* (2024) 37–41, https://issues.org/government-procurement-ai-systems-human-rights-hickok-hu/

5. Concluding thoughts

Careful study of VLMs as geolocators reveals some clear and surprising threats to privacy. These AI systems are often better at inferring location than purpose-built software. Widespread public access to these technologies creates immediate risks to privacy. There are clear incentives for bad actors, whether authoritarian or other governments, abusive partners or cybercriminals, to use them to gather additional information about people and their movements. And the potential use of these systems by military and security services or tech companies, such as social media companies for surveillance demands not only ongoing attention and scrutiny, but appropriate regulation and oversight.

We need to understand more about how these systems work, in order to calculate how confident we can be about particular geolocation predictions. But technical research on its own will not be able to explain the risks. Our project was a university/civil society collaboration, which benefited from on a wide range of expertise. We would encourage this approach to understanding the myriad risks of AI systems, especially when those risks are uncertain.

There are clear needs and opportunities for both policy measures and technical interventions to help mitigate these risks. We need to understand how widespread the uses of these technologies are becoming, so that users own privacy can be protected. Technical approaches to privacy-enhancement should be possible, such as adding 'noise' to pictures before they are shared. During the course of our year-long project, the capabilities of the most advanced AI systems have grown rapidly. If policymakers are to understand and keep ahead of the risks, continued collaboration between researchers, civil society and AI companies remains vital.

Privacy International
62 Britton Street
London EC1M 5UY
United Kingdom

+44 (0)20 3422 4321

privacyinternational.org