Life of Research · Published 20 November 2026 · DOI 10.5281/zenodo.24382475

A Benchmark Hit Is Not an Exposure: What Published Vision-Language Geolocation Accuracy Says About Location Privacy, Why One Model Locates 61 Percent of One Photo Set to a Kilometre and Fewer Than 12 Percent of Another to a Hundred, and the Exposure Estimand a Privacy Claim Needs

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Since 2024 a growing literature has argued that general-purpose vision-language models can work out where a photograph was taken well enough to make image geolocation an immediate privacy risk, and a second literature has begun building defences against it. In this analysis the author reviews and synthesises published results to show that the direction of that claim survives the evidence but its usual support does not: the accuracies quoted for it are properties of particular photo sets, adversaries and distance thresholds, and none of them is the quantity a privacy claim is about. The same model places 61 percent of a curated, text-rich stock-photo set within one kilometre, about 23 percent of a public Flickr benchmark, and fewer than 12 percent of sampled street-level images within one hundred kilometres. One video-frame benchmark reports 65.6 percent within a kilometre for clips from one platform and 6.4 percent for clips from another. The obvious deflation, that benchmark images were seen in training, does not explain the capability: on photographs collected by the authors and not drawn from any public source, frontier models still name the city about 60 percent of the time, and freshly taken photographs of Seoul are located about as well as Street View images of the same city. The author separates five variables that published results leave unfixed (image population, harm radius, adversary configuration, counterfactual baseline, and the unit of exposure), shows that defences are mostly scored against a fixed adversary although a published photograph must resist every later one, and proposes an exposure protocol that states each variable before a number is reported.

Keywords: image geolocation; vision-language models; location privacy; threat models; benchmark validity; contextual integrity; adversarial perturbation; privacy evaluation

1. Introduction

Working out where a photograph was taken from its pixels alone is an old problem in computer vision. Hays and Efros (2008) treated it as scene matching against six million geotagged web images; PlaNet (Weyand et al., 2016) recast it as classification over cells of the Earth's surface; hierarchical cells and scene classes followed (Muller-Budack et al., 2018), and so did retrieval over learned image-to-GPS embeddings (Vo et al., 2017; Vivanco Cepeda et al., 2023). For most of that history a strong system was a specialised one, trained on millions of geotagged images and evaluated on test sets drawn from the same photo-sharing sites. PIGEON (Haas et al., 2023), trained on data from the game GeoGuessr, put more than 40 percent of its guesses within 25 kilometres and beat one of the game's best professional players in six of six matches.

What changed in 2024 was who could do this. Mendes et al. (2024) prompted GPT-4V on the 237-image IM2GPS test set and found that it placed 24 percent of the images within a kilometre, with a median error of 13 kilometres, the lowest median error among the specialised systems they compared it with. Their abstract introduces the finding with the words "somewhat surprisingly" and concludes that geolocation by vision-language models is "an immediate privacy risk, rather than merely a theoretical future concern". A run of evaluations followed, each reaching a version of that conclusion: on Street View images (Jay et al., 2025; Zhang and Cheng, 2025), on photographs taken in private homes in California (Luo et al., 2025), on Korean streets (Wang et al., 2025), across 25 models and four datasets (Grainge et al., 2025), and with agents that search the web, consult maps or walk through Street View (Zhang, X. et al., 2025; Ji et al., 2026; Mukherjee et al., 2026). Alongside them, benchmark and method papers built on the same models, among them GeoLocator (Yang et al., 2023), LLMGeo (Wang et al., 2024) and GeoComp with its chain-of-thought method (Song et al., 2025), kept raising the accuracy available to anyone with an API key. A second literature began building defences: conversation moderation (Mendes et al., 2024), context-sensitive disclosure (Yang et al., 2026; Chen et al., 2026), and image-side protection by adversarial perturbation or editing (Liu et al., 2025; Wang et al., 2026; Wu, Y. et al., 2026).

A reader who wants to know how large the risk is meets a puzzle. Grainge et al. (2025) report that the strongest model they tested places 61 percent of one dataset within a kilometre, and they also report that current models "perform poorly on generic street-level images". Both statements are true of their own table. The two

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

numbers are not a contradiction about capability; they are measurements of two different things that share a name. One is accuracy on photographs chosen by annotators because they looked like social-media posts and carried legible text. The other is accuracy on street-level images sampled from a dataset that spans 225 countries.

This paper's thesis is that published geolocation accuracy, however carefully measured, is not yet evidence of the size of a privacy risk, because a privacy risk is a property of a population of shared photographs and an adversary, and the published numbers fix neither. Five variables decide what an accuracy means for privacy, and the literature varies all five without reporting them as variables. The image population sets the base rate of locatable photographs. The harm radius decides which errors still count as exposure. The adversary configuration sets how much effort and tooling is applied to each photograph. The counterfactual baseline decides whether the model adds anything over what a person with a search engine could already do. And the unit of exposure decides whether the risk is counted per photograph or per person, who usually posts many.

The analysis does not conclude that the risk is overstated. The deflationary reading that circulates alongside the alarm, that the accuracies are inflated because benchmark images sit in the training data, fails on the evidence the field already has: on 220 photographs collected by the authors and not drawn from any public source, o3 names the right city 60.8 percent of the time (Li et al., 2025), and KoreaGEO's freshly taken photographs of Seoul are located about as accurately as Street View images of the same city (Wang et al., 2025). What the evidence complicates is the shape of the risk. Capability is concentrated in photographs that carry legible text, landmarks or dense urban structure, and it falls off steeply elsewhere. A per-photograph average over a benchmark therefore misstates the risk in both directions: it overstates it for a person whose photographs show fields and kitchens, and understates it for a person who posts many photographs of the same neighbourhood, one of which shows a shop sign.

The same five variables explain why the defence literature is harder to read than it looks. Most defences are scored against one adversary chosen at evaluation time. A refusal policy binds only the users of the model that carries it. A perturbation added to an image before posting has to resist every model that will ever be run on that image, an asymmetry that Radiya-Dixit et al. (2021) showed undermines perturbation-based protection against facial recognition and that Honig et al. (2024) showed for protection against style mimicry. The defences that do not depend on outrunning future adversaries are the ones that remove the information itself, such as occluding text (Wang et al., 2025) or editing out peripheral clues (Wu, Y. et al., 2026), and they are paid for in the utility of the photograph.

The contributions are four. First, an account of what the anchor result measured and what it has been cited for (Section 2). Second, a decomposition of a geolocation result into five variables, each worked through against the published evidence (Sections 3 to 9), including the contamination question, on which the evidence cuts against the sceptical reading (Section 5). Third, an analysis of defences by which adversary they bind (Section 10), with the measured values consolidated in Table 1 and Figure 1 (Section 11). Fourth, an exposure protocol, Algorithm 1, that states the five variables before reporting a number and evaluates a defence against an adversary ladder rather than a single attacker (Section 12). Section 13 states the strongest case against this paper and Section 14 what is not known.

A note on scope. The paper concerns inference of where a photograph was taken from its visual content, by general-purpose or fine-tuned vision-language models and pipelines built on them. It does not treat location leaked through metadata, through the network, or through text alone, except where text accompanies the image, and it does not treat inference of other personal attributes except where that bears on linkage. Several sources are recent preprints from small teams; the paper relies on them for what they measured, and says where a single study carries a conclusion.

2. What the anchor measured, and what it was cited for

Mendes et al. (2024) did two things, and the field has mostly cited the first.

The first was a capability demonstration. GPT-4V, prompted with a geographic least-to-most scheme, was asked for coordinates for each of the 237 images in the IM2GPS test set (Hays and Efros, 2008). It placed 24 percent within a kilometre and reached a median error of 13 kilometres. One image was refused by a content filter and was scored as an infinite error. The comparison systems, GeoDecoder and PIGEOTTO, were taken from the PIGEON paper's reported results (Haas et al., 2023). On this basis the abstract states that current models are very capable geolocators and that the risk is immediate.

The second, which is most of the paper, was a moderation benchmark. GPTGeoChat consists of 1,000 images and 1,000 multi-turn conversations in which an annotator tried to get GPT-4V to reveal where each image was taken, with every turn annotated for the finest granularity revealed so far: country, city, neighbourhood, exact location name, or exact coordinates. Double annotation of 100 dialogues gave a Fleiss kappa of 0.76. The task for a moderation agent is to decide, turn by turn, when a response reveals more than a configured granularity. The fine-tuned agents and the prompted GPT-4V agent reached message-level F1 of roughly 0.8 at the country, city and neighbourhood levels; at the exact-location-name level fine-tuned agents beat the prompted GPT-4V agent by 10 to 15 points. When the location information left in moderated conversations was passed to a geocoding service, the best

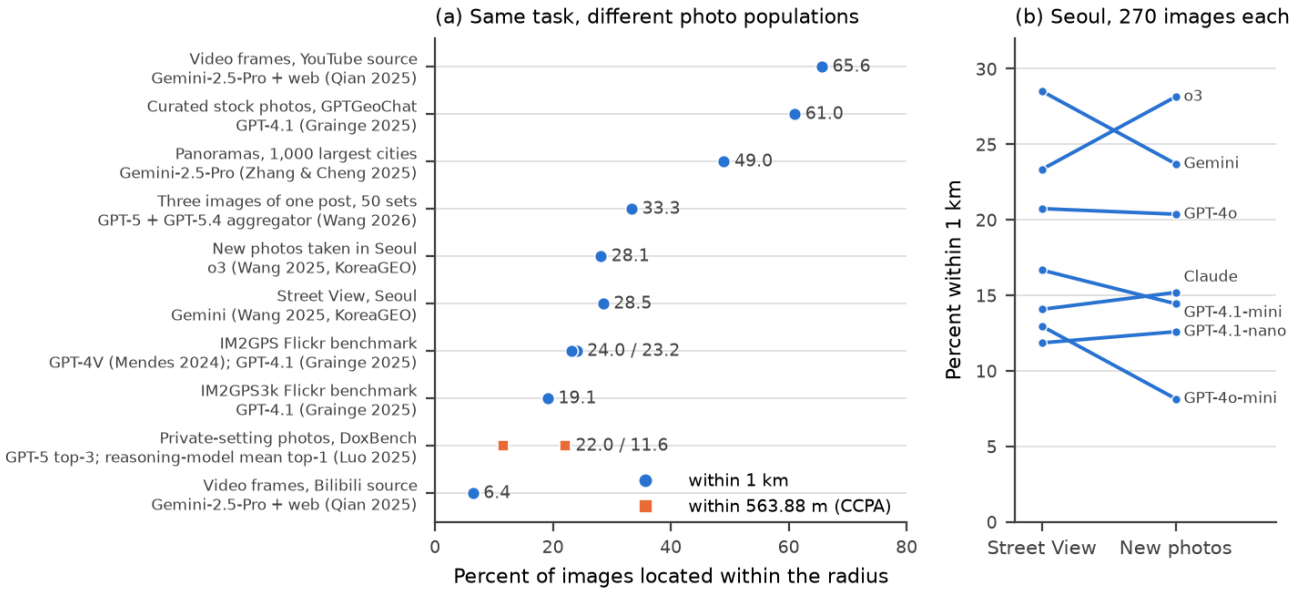


Figure 1. (a) Published accuracies of vision-language geolocation by photo population: share of images located within 1 km (circles) or within the 563.88 m statutory radius (squares, DoxBench), each point one value reported by the source named in its label (Qian et al., 2025; Grainge et al., 2025; Zhang and Cheng, 2025; Wang et al., 2026; Wang et al., 2025; Mendes et al., 2024; Luo et al., 2025). (b) KoreaGEO's leakage check (Wang et al., 2025, Table 10): accuracy within 1 km for seven models on 270 Street View images of Seoul and on 270 photographs the authors took there with the same place-type sampling.

prompted agent, configured to reveal only the country, still allowed 3 percent of images to be placed within 20 kilometres.

Two details of the second part matter for reading the first. The GPTGeoChat images were bought from a stock-photo provider by annotators who were told to prefer everyday scenes over iconic landmarks, and to prioritise images that resembled personal social-media photographs and that contained text on signs or posters. By the authors' own OCR count, 85 percent of the images contain text, against 19 percent of the IM2GPS test set and 25 percent of the 3,000-image IM2GPS3k set. And the authors say plainly why they did not use IM2GPS for the moderation benchmark: geolocating images in such benchmarks "tends to devolve into an object recognition task that does not require reasoning"; an image with a recognisable landmark is easy and one without is usually very hard.

That sentence is the key to the anchor. The 24 percent within a kilometre on IM2GPS is, by the anchor's own description of the benchmark, substantially a landmark-recognition rate on photographs from a photo-sharing site, collected for a 2008 paper. It establishes that a general-purpose model recognises famous places as well as a specialised geolocator does. It does not establish how often an ordinary photograph posted by an ordinary person can be placed within a kilometre of where it was taken, and the paper does not claim that it does. The claim that has travelled is the stronger one, carried by the word "immediate".

Nor is the moderation benchmark a measurement of exposure. It measures whether a second model can stop a first model from saying too much in a conversation. That is a defence against one adversary, a user of a moderated assistant who asks in conversation. The image is still the image. An adversary who passes it to a different model, or to the same model without the moderator, is not bound by it, a limit the authors acknowledge when they test what remains recoverable with external tools. Section 10 returns to what that means for defences generally.

3. Five variables a geolocation result leaves unfixed

This section states the decomposition the rest of the paper uses. The definitions are stipulations, and are stated flat.

A geolocation result, as published, is an estimate of the probability that the error of a predicted location falls below a radius r, for images drawn from some benchmark distribution D, under some attack configuration A. Write it acc(r; D, A). Benchmarks usually report it at several radii (1, 25, 200, 750 and 2,500 kilometres is the standard ladder), and sometimes as a median or mean error.

A privacy claim is about a different quantity. Call it exposure. For a population P of photographs that people actually share, a harm radius h set by what an adversary would do with the location, and an adversary class A* that describes who might attempt the inference with what tools, exposure is the probability that the adversary places a photograph, or the person behind it, within h of a place the person did not mean to reveal. The privacy-relevant change brought by a new model is not exposure itself but the uplift: exposure with the model minus exposure under the counterfactual adversary who has

everything except the model.

Five variables separate the published quantity from the needed one.

V1, the image population. D is a benchmark; P is what people post. The two can differ in the share of photographs that carry legible text, landmarks, dense urban structure or any clue at all. Section 4 shows that this variable moves accuracy more than the choice of model does.

V2, the harm radius. The radius at which a benchmark reports accuracy is a convention. The radius at which a location harms someone depends on what the location is: a home address is harmful at tens of metres, a protest at the neighbourhood, a clinic at the building. Section 6.

V3, the adversary configuration. A is usually a single query to one model with one prompt. Real adversaries can look around, search, aggregate photographs, combine models, jailbreak refusals or fine-tune open weights. Section 7.

V4, the counterfactual baseline. Uplift requires a baseline. Published comparisons use non-expert crowd workers, expert game players, traditional search engines or nothing. Section 8.

V5, the unit of exposure. Benchmarks score photographs. People are exposed as accounts, with many photographs and captions that share a location. Section 9.

Contamination, the question whether benchmark images were seen in training, is not a sixth variable. It is a way in which D can fail to stand for P even when the two look alike, and it gets its own section because it is the deflation most often offered (Section 5).

4. The image population decides more than the model does

The clearest evidence that V1 dominates comes from studies that evaluate the same models on several image sets.

Grainge et al. (2025) ran 25 generative models on four datasets with all metadata removed. On GPTGeoChat, GPT-4.1 placed 61.0 percent of images within a kilometre, Claude-3.7-Sonnet 60.0 percent and GPT-4-Turbo 60.0 percent. On the IM2GPS test set the same GPT-4.1 placed 23.2 percent within a kilometre, and on IM2GPS3k 19.1 percent. On a 500-image sample of OpenStreetView-5M, street-level imagery from Mapillary spanning 225 countries, the paper reports accuracy at 100 kilometres rather than one, and GPT-4.1, the best model there too, reached 11.6 percent. The authors' own summary is that the models do poorly on structured street-level imagery and strikingly well on images resembling social-media content, and they suggest that the asymmetry reflects what the models were trained on. The anchor's GPT-4V result, 24 percent within a kilometre on IM2GPS, sits beside GPT-4.1's 23.2 percent on the same 237 images, which is some reassurance that the benchmark result is stable across model generations; what moves the number is the photo set.

The spread survives inside a single benchmark. EarthWhere (Qian et al., 2025) draws its street-level tasks from two video platforms. With web search enabled, Gemini-2.5-Pro placed 65.6 percent of the YouTube-sourced items within a kilometre (of 96 scored) and 6.4 percent of the Bilibili-sourced items (of 47 scored, about three hits), and with web search disabled models scored on average 42.7 percent higher on the YouTube items than on the Bilibili ones. The authors read this as regional bias. Whatever its cause, it is a factor of ten in street-level accuracy for one model, one protocol and one benchmark, carried by which platform the frames came from.

Street View is not one population either. Zhang and Cheng (2025) sampled points within 10 kilometres of the 1,000 most populous cities in the world, stitched four directional views into panoramas, and evaluated 300 of them; Gemini-2.5-Pro placed 49 percent within a kilometre. Jay et al. (2025) sampled single 90-degree views at random points around up to 200 cities per continent and population class, 1,602 images from 88 countries; o1, the best model, named the right country for 84.5 percent and the right city for 14.2 percent, with a mean error of 681 kilometres. The gap between these two Street View studies is a gap in what was sampled (big cities against a stratified spread, four views against one) more than in which model was asked.

Why does the population matter so much? The diagnostic studies agree on the answer: the models lean on cues that can be read or named. IMAGEO-Bench (Li et al., 2025) regressed each model's success on the scene attributes the model itself reported, and found landmark presence and street-level scenes among the strongest positive predictors and scenic or indoor images among the negative ones; its authors summarise that performance peaks on urban, outdoor, landmark-rich scenes and degrades fast under distribution shift. KoreaGEO (Wang et al., 2025) found that indoor scenes were located more accurately, which the authors tentatively attribute to richer text, that rural and high-greenery scenes did worse, that transport signage helped and decorative text did not, and that Latin-script text helped more than Korean text. When the authors manually occluded key text in 120 photographs of commercial areas, accuracy fell for every model at the 20 and 100 kilometre thresholds, and for most models at 1 kilometre. Bharadwaj et al. (2026), testing open models of up to eight billion parameters on country prediction, found an urban-rural gap of about 15 points for their best model (83.8 against 68.5 percent on a GeoGuessr dataset) and noted that 82.8 percent of their OpenStreetView-5M evaluation sample was rural. The OpenStreetView-5M authors themselves judged by inspection that 35 percent of IM2GPS3k images are not localisable at all (Astruc et al., 2024).

The pattern has a familiar precedent outside geolocation. DeVries et al. (2019) found that commercial object-recognition systems performed worse on household items

from low-income countries, partly because the items looked different and partly because they appeared in different contexts. Geolocation benchmarks inherit the same skew in where photographs come from: IMAGEO-Bench reports better accuracy in North America and Western Europe and notes that its own sampling mirrors the availability of street-level imagery; KoreaGEO reports predictions biased towards core cities; GeoAgent reports a developed-developing gap across frontier models (Mukherjee et al., 2026).

For privacy this has an uncomfortable implication that cuts both ways. Exposure is concentrated. A photograph with a legible shop sign in a dense, well-photographed city is close to locatable to the street by today's models; a photograph of a lawn is not. A benchmark average is a mixture over these, with mixture weights set by whoever assembled the benchmark. GPTGeoChat's weights were set deliberately towards clue-rich images, because its purpose was to test moderation of conversations that do reach fine granularity. IM2GPS's weights were set by 2008-era Flickr. OpenStreetView-5M's were set by where Mapillary contributors drive. None of them was set by what a particular population of users posts, and the paper finds no published estimate of that mixture for any platform.

5. Contamination: the deflation that does not survive

The most common sceptical response to the capability numbers is that they are inflated by memorisation. The concern is legitimate for the older benchmarks. IM2GPS, IM2GPS3k, YFCC4k and the MP-16 training set are all drawn from Flickr (Thomee et al., 2016; Vo et al., 2017), and Flickr photographs are plausible pretraining data for any large model trained on web images. The geolocation literature has taken the concern seriously in specific ways. OpenStreetView-5M enforces a one-kilometre separation between every training and test image precisely because, without it, geolocation reduces to place recognition, and its authors show that removing the separation makes the task markedly easier (Astruc et al., 2024). GeoSearch (Le-Duc et al., 2026) rebuilt its retrieval database from OpenStreetView-5M rather than MP-16 because all three standard benchmarks share MP-16's Flickr source, and reports that most retrieval-based methods lose accuracy on IM2GPS3k when the leaked database is removed. GEO-Detective (Zhang, X. et al., 2025) states that its main Flickr-derived evaluation may be contaminated and adds DoxBench as a cleaner test. VLM-GeoPrivacy (Yang et al., 2026), which reuses images from YFCC subsets, IM2GPS3k and GPTGeoChat, argues that this is not contamination for its purpose because it measures disclosure judgement, not geolocation capability, which concedes the point for the latter.

The deflation predicts something testable: that on photographs no model could have seen, accuracy should collapse. Three studies test it, and none finds a collapse.

IMAGEO-Bench (Li et al., 2025) includes a held-out set of 220 photo-address pairs collected by the authors and not drawn from any public source, spanning natural, rural and urban scenes, weather and times of day. On it, o3 named the correct city 60.8 percent of the time, GPT-4.1 60.0 percent, and Gemini-2.5-Pro reached a median error of 0.7 kilometres. Open models trailed (40.8 and 41.9 percent city accuracy for two Llama variants), but not by a margin that suggests the closed models' results were memorised.

KoreaGEO (Wang et al., 2025) ran the cleanest version of the test. To check for leakage from pretraining, the authors took 270 new photographs in Seoul with the same place-type sampling as their Street View subset for that city, extracted ground truth from EXIF data, stripped the metadata and evaluated closed models on both sets. Within a kilometre, o3 located 23.3 percent of the Street View images and 28.1 percent of the new photographs; GPT-4o 20.7 and 20.4 percent; Gemini 28.5 and 23.7 percent. At 20 kilometres every model did better on the new photographs, but the authors warn against reading too much into that radius: smaller models that did well there show "a manifestation of severe geographic bias, since these models tend to overpredict Seoul", and every new photograph was taken in Seoul. At one kilometre the result is mixed across the seven models, three up, three down and one roughly flat; at 100 metres six of the seven did better on the new photographs. The authors suggest that Gemini's drop at one kilometre may reflect prior exposure to Street View, and that the gains on new photographs reflect better image quality and clearer text. Either way, the photographs that a person would actually take did not defeat the models; at the strictest radius they slightly helped.

DoxBench (Luo et al., 2025) was built from 500 photographs that the researchers took on iPhones in six California cities, including selfies and pictures inside private residences, with ground-truth coordinates from EXIF. It was collected in 2025, and later work uses it, in one case explicitly to avoid contamination (Zhang, X. et al., 2025), and in another as the test set for a defence (Wang et al., 2026). Under the California Consumer Privacy Act, location data that places a person within 1,850 feet (563.88 metres) is "precise geolocation" and counts as sensitive personal information; the authors count a prediction as a violation at that radius. Across the reasoning models that produced valid answers, the top-1 prediction fell within that radius for 11.61 percent of images on average, and within the top three guesses for 14.95 percent; GPT-5's top-three figure was 22.03 percent. The authors report that this is about twice the rate achieved by 268 non-expert crowd workers.

The three studies differ in population and radius, so their numbers do not line up into one estimate, and that is the point of Section 4. What they share is that each was built to remove the contamination explanation, and in each the capability remained. The honest reading is that contamination inflates the old landmark-heavy

benchmarks, probably substantially for retrieval systems, and that it does not explain why current models locate fresh, clue-bearing photographs. The sceptic who dismisses the alarm on contamination grounds is wrong on the evidence. The alarm still needs the other four variables before it becomes a measured risk.

The figures that come closest to measuring exposure for private photographs, DoxBench's 11.61 and 22.03 percent at the statutory radius, are also notably lower than the benchmark figures that circulate in the alarm. The strongest models place six in ten curated stock photographs within a kilometre; the reasoning models evaluated on DoxBench place, on average, roughly one in nine fresh private-setting photographs within the statutory radius on their first guess. Both numbers are alarming in absolute terms for anyone who posts photographs from home. They are different numbers, and the difference is V1 and V2.

6. The harm radius is not the benchmark radius

The one-kilometre threshold is called "street level" across this literature, and the next ones up are city (25 kilometres), region (200), country (750) and continent (2,500), a ladder inherited from the retrieval benchmarks (Vo et al., 2017) and reused by the privacy papers (Mendes et al., 2024; Yang et al., 2026). The names are conventions for comparing systems. They are not statements about harm, and the privacy literature has started to replace them with radii that are.

DoxBench is the clearest case. It scores predictions at 563.88 metres because that is where Californian law draws the line for precise geolocation (Luo et al., 2025). A result at one kilometre and a result at 564 metres are not interchangeable: the former counts as a hit a guess that lands in the right district of a dense city, the latter roughly a guess that lands within a few blocks. For a home address even 564 metres is coarse; a residential street is what makes a photograph a threat to the person who took it, and the hit rate at 100 metres is the one that matters. KoreaGEO is one of the few benchmarks to report a 100-metre threshold at all, and at that radius no model in its Seoul comparison exceeded 11.1 percent on either the Street View or the newly taken photographs (Wang et al., 2025). In the band from 100 metres to a kilometre its authors find tourist landmarks most exposed, with commercial, civic, educational, religious and transport sites at moderate risk.

The harm radius also depends on what is being revealed, which is the insight behind contextual integrity: whether an information flow is a violation depends on the norms of the context in which it was shared, not on the information alone (Barth et al., 2006). VLM-GeoPrivacy (Yang et al., 2026) operationalises this for geolocation. Annotators labelled 1,200 images with whether the sharer appeared to intend to reveal the location and with the finest granularity that would be appropriate to disclose, ranging from abstention to exact coordinates; a photograph of a famous landmark warrants an exact answer, a photograph of a political protest warrants much less. On this benchmark the best of 14 models matched the human disclosure judgement in 49.7 percent of free-form cases, and GPT-5 disclosed more than the annotated level in 47.6 percent of cases when simply asked where a photograph was taken. An earlier taxonomy grounded in legal frameworks found models' judgements of contextual privacy inconsistent (Tsaprazlis et al., 2025), and Tsaprazlis et al. (2026) make the related point that privacy severity is compositional: attributes harmless alone combine into a serious exposure, as in their example of "a patient's wristband, a clinic room number, and a distinctive treatment device in the background", and frontier models in their evaluation systematically underestimated that composition.

Coarse radii are not harmless either, once they are combined. De Montjoye et al. (2013) showed, for mobile-phone traces located only to the nearest antenna and the nearest hour, that four spatio-temporal points were enough to single out 95 percent of 1.5 million people, and that uniqueness decays only as about the tenth power of resolution, so coarsening buys little anonymity. A photograph that is located only to the city is still a spatio-temporal point when it carries a timestamp, and a person who posts many such photographs supplies many such points. The relevant radius for an account is therefore not the radius of any one photograph; Section 9 returns to this.

The consequence for reading results is direct. A claim of risk should name its harm radius and justify it by what the location would let an adversary do. A claim that a model is "street-level accurate" a quarter of the time on a benchmark is a statement at one kilometre; it can be consistent with a far lower rate at the radius where a home is found, and with a far higher rate at the radius where a person's city, combined with three other posts, identifies them.

7. The adversary is a configuration, not a model

Most published accuracies come from a single query to one model with one prompt. That is the weakest adversary that the privacy argument has to consider, and the published evidence on stronger ones shows that capability is not monotone in effort: some tools help, some hurt, and which is which depends on the image.

More views. Jay et al. (2025) let models control the Street View camera for five successive views. Claude 3.5 Sonnet cut its mean error by 30.6 percent over the five guesses and GPT-4o by 28.1 percent, while GPT-4o mini cut its by only 14.3 percent; after three views the agent's median error was below that of GeoGuessr players in the Champion Division, as reported by the PIGEON authors, although still above PIGEON's. GeoAgent (Mukherjee et al., 2026) reports that navigation through Street View improves accuracy over static images but that models still struggle with regional distinctions. GeoGuesser Bench (Mao et al., 2026) lets agents rotate, zoom and

move between real panoramas; on 42 rounds from the GeoGuessr World Championship the best agent averaged 4,352.4 points against 4,770.3 for the better of the two human experts in each round.

Search and maps. The effect of retrieval is the least predictable. In Jay et al. (2025), giving the agent the top ten Google Lens results made it worse, with an 85.3 percent higher mean error and the loss of all its correct city predictions, which the authors attribute to noise in the results. EarthWhere found web search no help for country identification, with an average drop of 1.72 percent and a 13.2 percent drop for GPT-4o, but an average 6.5 percent relative gain on street-level tasks where the image contained richer clues (Qian et al., 2025). GEO-Detective (Zhang, X. et al., 2025), which adds image-based reverse search to an agent rather than converting cues into text queries, reports gains of 11.1 percent at country level and about 5.2 percent at finer levels over its baselines, mostly on images without visible geographic features, and a more than 50.6 percent reduction in "unknown" answers when external clues are supplied. Thinking with Map (Ji et al., 2026) trains an agent to query maps and reports accuracy within 500 metres rising from 8.0 percent for Gemini-3-Pro with Google Search and Maps grounding to 22.1 percent for its own agent. Retrieval does not help uniformly: GEO-Detective's reverse search helped most on images without visible geographic features, whereas EarthWhere found that web search helped only where the image already carried rich clues, and the adversary who uses retrieval badly loses accuracy.

Captions and context. A posted photograph rarely travels alone. KoreaGEO (Wang et al., 2025) attached two kinds of caption to each image: functional captions resembling ordinary posts with no place names, and high-risk captions resembling check-ins. Functional captions had inconsistent effects. High-risk captions raised every model: o3 and Gemini passed 45 percent within a kilometre, GPT-4.1-nano rose from 3.06 to 21.85 percent within a kilometre, and LLaMA from 28.08 to 82.73 percent within 20 kilometres. The authors conclude that the input modality outweighs model scale. This is the variable most clearly missing from image-only benchmarks, and it is the one whose effect is largest.

Many photographs. PlaNet already showed in 2016 that pooling the photographs of an album with a recurrent model improved accuracy by 50 percent over single images (Weyand et al., 2016). In the vision-language setting, Wang et al. (2026) aggregated the per-image answers of GPT-5 and of Claude Opus 4.5 across the three images in each of 50 sets they assembled to simulate a social-media post, using GPT-5.4 as aggregator. Accuracy within a kilometre reached 33.33 percent for GPT-5 and 25.00 percent for Claude Opus 4.5 on unprotected image sets, and Claude's mean error fell to 5.013 kilometres. Luo et al. (2025) show an example in which o3, given two keyframes from one video recorded at a fixed position, placed the location at the exact address, where single frames had not; that is a single example, offered by its authors as a demonstration and not as a rate.

Several models. Wang et al. (2026) also pooled the reasoning of five commercial models through an aggregator and report that the ensemble raised geolocation accuracy, particularly at the city and region levels.

Refusals and how to get past them. Grainge et al. (2025) counted failures to give coordinates as part of the result and found them large for some models: 61.4 percent for o4-mini, largely refusals, and 75.5 percent for IDEFICS. A single-query benchmark therefore partly measures refusal policy. Refusal policy is not a barrier to a motivated adversary: Wang et al. (2026) report that jailbreak prompts raised the response rate of refusing reasoning models to 100 percent, and they themselves prepended jailbreak templates to GPT-5 and Claude Opus 4.5 in order to evaluate them, raising response rates above 70 percent.

Open weights and fine-tuning. The adversary who owns the weights needs neither prompts nor jailbreaks. Fine-tuning is cheap in this domain: GeoLocSFT reports competitive results after supervised fine-tuning of Gemma 3 on 2,700 selected image-coordinate pairs (Yi and Shan, 2025); Bharadwaj et al. (2026) report that a LoRA adaptation of a one-billion-parameter model raised top-1 country accuracy on a 20-country subset from 40.40 to 79.44 percent, with the largest gains on rural scenes; Geo-R (Wu, B. et al., 2026) and GeoReasoner (Li et al., 2024) train open models with reinforcement or reasoning supervision, and retrieval-augmented pipelines such as Img2Loc (Zhou et al., 2024), G3 (Jia et al., 2024) and the open-weight system of Bicakci et al. (2025) reach strong benchmark results, the first and last without training any model. LocationAgent (Li et al., 2026) and GaGA (Dou et al., 2024) package the same capability as an agent or an interactive assistant. A capability that can be added to a small open model with a few thousand labelled examples is not governed by the refusal policy of any frontier deployment.

Taken together, the configuration evidence says that single-query accuracy is a lower bound on what a capable adversary achieves, but not a tight one and not a uniform one. Some of the stronger rungs gain most on images that single-query evaluation finds hard: GEO-Detective's reverse-search agent helps most on images without visible geographic features, and PlaNet's album model recovers photographs its single-image model misses. The evidence is not uniform: EarthWhere found web search helped only where the image already carried rich clues, and the caption effect in KoreaGEO was reported by model, not by photograph. A privacy claim that reports only the single-query number can therefore understate the exposure of people whose individual photographs look safe.

8. The counterfactual: uplift over whom?

A model that locates a photograph creates a privacy harm only to the extent that the location was not already obtainable. The counterfactual matters for policy because it decides what the model adds: a new capability, or an old capability at a new price.

The baselines in the literature are of three kinds, and they support different conclusions.

Non-experts. Jay et al. (2025) collected a human baseline from eight volunteers, including the authors, with little or no GeoGuessr experience; their mean error was 5,633 kilometres. DoxBench used 268 crowd workers and found the models' precise-location rate about twice theirs (Luo et al., 2025). OpenStreetView-5M asked 80 annotators to locate 50 test images and found them well above chance but far below the trained models (Astruc et al., 2024). Against non-experts working alone the uplift is large and not in dispute.

Experts. Against the best human players the picture is closer. PIGEON beat a top professional in six of six matches (Haas et al., 2023). GeoGuesser Bench's best agent averaged 4,352.4 points to the better expert's 4,770.3 on championship rounds (Mao et al., 2026). GeoRC reports that large closed models rival champion-tier experts at predicting locations but trail them by about 12 points in the quality of their reasoning chains, which often cite cues that are not in the image (Talreja et al., 2026). Against an expert, then, the models are comparable rather than superior in accuracy, and inferior at explaining what they used.

Ordinary tools. A person with a search engine is a third baseline, and the one most relevant to the realistic adversary. The only controlled comparison the author found is Wazzan et al. (2024), in which 60 participants geolocated images with either a traditional search engine (Bing) or an early LLM-based search assistant (Bing Chat); the traditional-search group scored higher (mean 3,189 against 2,725, $p = 0.041$). That study tested a text-only assistant, with the human doing the looking, and does not measure a vision-language model; it shows only that the tool-assisted human baseline is not weak. For any photograph that has already been posted publicly, reverse image search is a further baseline: it does not infer a location, it finds the page where the photograph was posted, which may state it. Jay et al. (2025) found that results of this kind confused a model more than they helped.

The pattern is the one that Staab et al. (2023) documented for text. Their models inferred personal attributes, location among them, from real Reddit profiles at up to 85 percent top-1 accuracy, at about a hundredth of the cost and a two-hundred-and-fortieth of the time required by human labellers. The privacy change is not that a new attribute became inferable; skilled people could already infer it. It is that inference became cheap enough to run on everyone. Tomekce et al. (2024) extended the finding to images, inferring personal attributes from benign photographs at up to 77.6 percent accuracy and finding that accuracy rose with general model capability.

Applied to geolocation, this says that the uplift worth measuring is in cost per located photograph and in the number of photographs an adversary can process, more than in the accuracy ceiling. No geolocation paper the author found reports cost per located photograph at a stated harm radius, against a stated human or tool baseline. DoxBench's framing, that the models lower the barrier for non-experts, is the closest; it is a statement about who can now do it, which is the right kind of statement, without the cost term that would make it a measurement of scale.

9. Images, posts and people

Benchmarks score photographs one at a time. People are exposed as accounts.

The arithmetic of the difference is elementary and is worth stating because it changes the sign of the usual reading. Suppose a person posts k photographs taken near the same sensitive place, and an adversary attempts each separately. If each attempt independently succeeds at the harm radius with probability p, the probability that at least one succeeds is $1 - (1 - p)^k$. This is a derivation, not a measurement, and the independence assumption is wrong in a known direction: photographs of the same place share their cues, so their outcomes are positively associated, and under positive association the chance of at least one success is lower than the independent formula gives. The formula is therefore an upper bound for separate attempts on homogeneous photographs.

Two facts push the other way. First, the photographs are not homogeneous. Section 4 showed that p is close to zero for a photograph of a lawn and far higher for one that shows a shop sign. For an account, what matters is roughly the most locatable photograph the person has posted near the place, not the average one, and a benchmark mean is a statement about the average one. Second, an adversary need not attempt photographs separately. Joint reasoning over several photographs can succeed where every single attempt fails, as PlaNet's album model, the multi-image aggregation of Wang et al. (2026) and the two-frame example of Luo et al. (2025) each illustrate.

Linkage then multiplies the effect. A located photograph is one spatio-temporal point; de Montjoye et al. (2013) showed how few such points single a person out. Location also composes with other inferred attributes: the same models that place a photograph can infer, from the same photograph, attributes of the person who posted it (Tomekce et al., 2024), and compositions of individually benign attributes are where severity lies (Tsaprazlis et al., 2026). Users interviewed by Zhang, S. et al. (2025) named exactly these concerns, identification, misuse and pervasive surveillance, alongside harm from inferences that are wrong, and

doubted that their own mitigations would work.

The privacy-relevant unit is therefore the account over a period, and the relevant statistic is the probability that an adversary with a given configuration reaches the harm radius for at least one sensitive place of that person. No published benchmark estimates it. The closest are the multi-image aggregation result of Wang et al. (2026), which aggregates the images of one post rather than of one person, and DoxBench's categories, which separate photographs by the kind of place they show but not by whose they are.

10. Defences: who does each one bind?

The defence literature is now as large as the attack literature, and it is harder to read for the same reason: a defence is evaluated against an adversary, and the adversary is usually one point on the ladder of Section 7. This section sorts the published defences by the class of adversary they can bind at all, which is a property of where the defence acts, before asking how well they work.

Deployment-side policies act inside a model or the service that hosts it: moderation of conversations (Mendes et al., 2024), refusal, context-sensitive disclosure (Yang et al., 2026), activation steering towards an appropriate granularity (Chen et al., 2026), and privacy-aware fine-tuning (Samson et al., 2024). They bind only adversaries who use that deployment and do not circumvent it. On their own terms some are promising: Chen et al. (2026) find that the appropriate disclosure granularity can be read from hidden states before generation, even when the generated answer is misaligned, and use it to steer; Samson et al. (2024) report substantial gains in privacy sensitivity from a few hundred fine-tuning examples. But the measured rates of the current deployments are not reassuring. A system prompt instructing refusal in DoxBench cut o3's answer rate on the highest-risk images from 88.0 to 0.0 percent, at the cost of cutting its answer rate on benign images from 100.0 to 32.0 percent; GPT-4.1 kept 98.0 percent of benign answers but still answered 54.0 percent of the highest-risk images (Luo et al., 2025). Jailbreaks raised refusing models to full compliance (Wang et al., 2026). And none of these measures binds an adversary who runs another model, an open one, or a fine-tuned one (Section 7).

Image-side perturbation acts on the photograph before it is posted: adversarial noise optimised to mislead geolocation (Liu et al., 2025), perturbation in a diffusion model's latent space (Wang et al., 2026), perturbations against attribute inference more generally (Fan et al., 2025), and, for images sent to cloud services, perturbations meant to hide private content from the provider while keeping the service useful (Hui et al., 2026). It binds any adversary whose model the perturbation transfers to. The difficulty is the asymmetry that Radiya-Dixit et al. (2021) identified for facial-recognition cloaking: the user perturbs a photograph once, at posting, and it must then defeat every model that will ever be applied to it, including models trained later or trained specifically on perturbed images. Their analysis concerned Fawkes, whose authors had reported more than 95 percent protection in their experiments and complete success against the commercial facial-recognition services they tested (Shan et al., 2020); later models and adaptive training undid the protection for photographs already collected. Honig et al. (2024) found the same for protections against style mimicry, where off-the-shelf steps such as upscaling sufficed. The geolocation defences evaluate some adaptive steps and report partial robustness. Wang et al. (2026) find that after DiffPure purification, the mean error distance of protected images is about 2.5 times that of purified clean images on GPT-5 (49.9 against 20.3 kilometres) and about 1.8 times on Claude Opus 4.5 (43.2 against 24.0 kilometres); their own deviation metric, the shift of the top-1 prediction, is 4.75 and 1.83 times the unprotected baseline. These figures come from a subset of the 150 DoxBench images the authors describe as most fragile, and the authors leave "a full evaluation of purification-aware perturbations as future work". Luo et al. (2025) found Gaussian-noise protection unstable, plateauing or reversing at high noise and leaving the lowest-risk images unprotected at moderate noise. None of these evaluations can test models that do not yet exist, which is the adversary that matters for a photograph that stays online.

Information removal edits out the cue itself. KoreaGEO's manual text occlusion lowered every model's accuracy at the 20 and 100 kilometre thresholds, and most models' accuracy at 1 kilometre (Wang et al., 2025). GeoCloak (Wu, Y. et al., 2026) builds on an observation it calls the visual-geolocation asymmetry: decisive clues such as street signs tend to sit at the periphery of a photograph while the subject sits in the centre. Its agent identifies such clues, edits them out semantically and adds perturbation for what remains. City-level accuracy fell from 41.2 to 9.9 percent, against 19.7 percent for GeoShield and 24.5 percent for another perturbation baseline. Against an adaptive attacker told that the image might have been modified, accuracy fell from 41.6 to 15.9 percent; the adaptive prompt also slightly reduced accuracy on unprotected images, from 43.4 to 41.6 percent. Semantic editing alone was used in 39.9 percent of cases and combined with perturbation in 58.4 percent. Removal is the only family whose protection does not depend on outrunning future models, because a sign that has been edited out cannot be read by any model. It is paid for in the photograph, and it depends on finding the cues. That second dependence is fragile: GeoRC shows that vision-language models' accounts of which cues they used are often hallucinated (Talreja et al., 2026), so a defender who asks a model which clues to remove may be told about the wrong ones. The text-privacy literature faced the same problem and answered it by letting the inferring model guide the rewrite: adversarial anonymisation uses an attribute-inferring model to decide what to change and outperformed commercial anonymisers on both privacy

and utility in its authors' evaluation (Staab et al., 2024). GeoCloak follows that pattern for images, and its own adaptive mode supplies the ground-truth location to its clue-finding agent in the 12.1 percent of cases where it fails to identify the country unaided.

GEO-Detective's comparison of four simple defences (a misleading location caption, a visible warning watermark, a distracting visual trigger and forged EXIF data) found the watermark most effective at inducing refusal but visibly intrusive, and found its agent more robust to all four than single-query models (Zhang, X. et al., 2025). That is the general shape: defences evaluated against the weakest adversary look stronger than they are.

The sorting suggests a reading rule. A deployment-side result is evidence about a product. An image-side perturbation result is evidence about the models it was tested against, with an unknown expiry. An information-removal result is evidence about the photograph, valid against any adversary for the cues removed, and conditional on the cue-finder. Only the third kind speaks directly to exposure as Section 3 defines it, and it is the kind least often evaluated against the realistic population of photographs and the realistic utility cost.

11. The measured values

Table 1 consolidates the published measurements on which the argument rests, one row per result as its source reports it. The rows are not on a common scale. Image sets, radii, models and protocols differ from row to row, and that heterogeneity is the paper's first finding, not an inconvenience to be averaged away. Where a row gives two numbers separated by an arrow, both come from the same source under the two conditions named.

Figure 1 draws the street-level rows of Table 1 on one axis. Panel (a) places each published accuracy within a kilometre, or within the statutory radius for DoxBench, against a description of the photo population; the spread from 6.4 to 65.6 percent is far larger than the gap between frontier closed models on a single population, such as GPT-4.1, Claude-3.7-Sonnet and Gemini-1.5-Pro on GPTGeoChat at 61, 60 and 50 percent (Grainge et al., 2025). Panel (b) draws KoreaGEO's leakage check: for each of seven models, accuracy within a kilometre on 270 Street View images of Seoul and on 270 photographs the authors took there, which do not separate in the way a contamination explanation predicts.

Three readings of Table 1 and Figure 1 should be kept apart. The population rows (Grainge, Qian, Zhang and Cheng, Jay) show that V1 moves accuracy by an order of magnitude. The contamination rows (Li, Wang 2025, Luo) show that the capability survives on photographs no model could have seen, at levels that are lower than the curated benchmark but far from negligible. The adversary and defence rows (Jay, Ji, Wang 2026, Bharadwaj, Luo, Wu) show that both capability and protection depend on which point of the adversary ladder is evaluated, in directions that are not uniform.

12. An exposure protocol

The argument implies a protocol, and stating it as an algorithm makes the decision structure explicit. Algorithm 1 is what the author proposes a geolocation privacy claim, or a defence claim, should report. Every step is drawn from something at least one published study already does; what no study does is all of them on one population.

```
Algorithm 1. Exposure protocol for a geolocation
privacy or defence claim.

Input: photo population P with sampling frame;
  harm radii h(t) per place type t;
  adversary ladder A0..A6; human baseline B;
  optional defence d.

Step 0. Describe P: platform, sampling frame, and
  shares with text, landmark, urban, indoor, people.
  A benchmark may stand in for P only if these
  shares are reported for both.
Step 1. Fix h(t) for each sensitive place type t
  (home, workplace, clinic, school, protest) and
  justify each radius by what it enables.
Step 2. Clean set: photos taken after every tested
  model's cutoff, never posted; check that reverse
  image search finds no copy.
Step 3. For each adversary a in the ladder:
  A0 one query, one model, one prompt
  A1 several views of the same scene
  A2 photo plus the caption it was posted with
  A3 web search, reverse search and maps
  A4 all photos of one account near one place
  A5 jailbreak, or an ensemble of models
  A6 open weights fine-tuned on public data
  record acc(h; P, a), refusal rate, and
  cost per photo (money and minutes).
Step 4. Run B (non-expert with search; expert) on
  the same photos at the same h; record its cost.
Step 5. Report uplift(a) = acc(h; P, a) - acc(h; P, B)
  and the cost ratio cost(B) / cost(a).
Step 6. Account level: for each account, report
  whether any attempt reached h for a sensitive
  place; give the share of accounts, not photos.
Step 7. If d is given: classify it as deployment,
  perturbation or removal; rerun Steps 3-6 on the
  protected photos. For perturbation add A7: input
  purification, upscaling, and a model released
  after d was tuned. Report the utility cost.
Step 8. State exposure only for the (P, h, a)
  combinations actually run, and name the weakest
  rung that was not run.
```

Three steps carry most of the weight, and each corresponds to a gap identified above.

Step 0 is the V1 correction. The contrasts in Section 4 show that a benchmark's composition sets most of its number. A geolocation privacy claim that does not report the composition of its photo set cannot be compared with any other, and cannot be transferred to a population of users. The shares Step 0 asks for are cheap to compute: Mendes et al. (2024) report the share of images with text from an off-the-shelf OCR tool, and IMAGEO-Bench extracts scene attributes from the models' own structured outputs (Li et al., 2025). What is expensive is obtaining a sampling frame for P, which requires access to what a platform's users actually post; Section 14 lists this as unknown.

Step 3 is the V3 correction. The ladder is ordered roughly by the effort it costs an adversary, not by the accuracy it yields, because Section 7 showed that

Table 1. Published measurements bearing on vision-language image geolocation as a privacy risk. "1 km" and the other radii are the error thresholds at which the source counts a prediction correct; "CCPA" is 563.88 m; "L3" is DoxBench's highest-risk image level.

Source	Image set	System or adversary	Metric	Reported value
Mendes et al., 2024	IM2GPS, 237 Flickr images	GPT-4V, least-to-most prompt	1 km; median error	24%; 13 km
Grainge et al., 2025	GPTGeoChat, 1,000 curated stock photos	GPT-4.1	1 km	61.0%
Grainge et al., 2025	IM2GPS / IM2GPS3k	GPT-4.1	1 km	23.2% / 19.1%
Grainge et al., 2025	OpenStreetView-5M, 500 images	GPT-4.1	100 km	11.6%
Qian et al., 2025	EarthWhere street tasks, YouTube / Bilibili	Gemini-2.5-Pro with web	1 km	65.6% / 6.4%
Zhang and Cheng, 2025	300 panoramas near large cities	Gemini-2.5-Pro	1 km	49%
Jay et al., 2025	1,602 single Street View views	o1	country / city	84.5% / 14.2%
Jay et al., 2025	319-view subset	Claude 3.5 Sonnet, five views	mean error	30.6% lower
Jay et al., 2025	319-view subset	agent with Google Lens results	mean error	85.3% higher
Li et al., 2025	220 privately collected photos	o3 / GPT-4.1	city	60.8% / 60.0%
Wang et al., 2025	Seoul, Street View vs new photos	o3	1 km	23.3% vs 28.1%
Wang et al., 2025	Seoul, Street View vs new photos	Gemini	1 km	28.5% vs 23.7%
Wang et al., 2025	KoreaGEO, no caption → check-in caption	GPT-4.1-nano	1 km	3.06% → 21.85%
Luo et al., 2025	DoxBench, 500 private-setting photos	reasoning models, mean	CCPA, top-1 / top-3	11.61% / 14.95%
Luo et al., 2025	DoxBench	GPT-5	CCPA, top-3	22.03%
Wang et al., 2026	50 three-image sets, one post each	GPT-5, three images aggregated by GPT-5.4	1 km	33.33% (Claude Opus 4.5: 25.00%)
Ji et al., 2026	MAPBench	grounded Gemini-3-Pro → map agent	500 m	8.0% → 22.1%
Mao et al., 2026	42 championship rounds	best agent vs better expert	game points	4,352.4 vs 4,770.3
Bharadwaj et al., 2026	20-country GeoGuessr subset	1B model, before → after LoRA	country	40.40% → 79.44%
Yang et al., 2026	VLM-GeoPrivacy, 1,200 images	best of 14 / GPT-5	disclosure	49.7% match / 47.6% over
Luo et al., 2025	DoxBench	o3, refusal prompt added	answer rate L3 / benign	88.0->0.0% / 100.0->32.0%
Wu, Y. et al., 2026	four datasets, zero-shot	naive / adaptive attacker, GeoCloak	city	43.4->11.1% / 41.6->15.9%
Wang et al., 2026	150 DoxBench L3 images	perturbation after DiffPure, GPT-5	mean error vs purified clean	about 2.5 times

accuracy is not monotone along it: retrieval hurt in one study and helped in another, and reasoning depth helped little on country identification. Reporting each rung separately keeps those non-monotonicities visible instead of folding them into a single best number. A4, the account-level rung, is the one most often missing and, by the argument of Section 9, the one most likely to change the conclusion.

Step 7 is the defence correction. A perturbation defence evaluated only against the models available at tuning time has been evaluated against the weakest adversary its users will face, for the reason given by Radiya-Dixit et al. (2021). A7 does not solve that problem, because nobody can test a future model; it bounds the problem by testing the cheapest known removal steps and the newest available model, and by recording the date. A removal defence does not need A7 against the cues it removed, but it needs the clean set of Step 2 and an independent check that the removed cues were the ones the attacker used, since a model's account of its cues is not reliable evidence of them (Talreja et al., 2026).

The protocol makes some claims more expensive and some cheaper. A claim that a model is a privacy risk for users of a platform becomes expensive, because it needs P. A claim that a model raises exposure at a stated radius for a stated population over a stated baseline becomes cheap and comparable. That trade is intended. The published literature is rich in the second kind of measurement, under different names, and poor in the first kind of claim, made under its name anyway.

13. The strongest case against this paper

The decomposition is obvious, and benchmark authors know it. Many of the sources state one or more of the five variables themselves: Grainge et al. (2025) attribute the population effect to training data, DoxBench chooses a statutory radius, KoreaGEO tests captions and leakage, VLM-GeoPrivacy makes the harm context-dependent, and GeoCloak tests adaptive attackers. The paper's claim is narrower than novelty of any variable. It is that no published result fixes all five, so none measures exposure, and that the variables interact: some adversary rungs, reverse image search and pooling an album, raise accuracy most on the photographs a population-blind benchmark scores as safe, although web search in EarthWhere helped only on

clue-rich images. The reader who finds the decomposition obvious should find Table 1 more troubling, not less.

Precaution makes the measurement question secondary. If a model can place one in nine private photographs within a few hundred metres, that is enough to act on, whatever the exact exposure. The author agrees, as a normative matter: the evidence of Section 5 is sufficient to treat geolocation as a real privacy capability now. The argument of this paper is about which actions follow. Precautionary action chosen on the basis of the curated-benchmark number points towards deployment-side refusals; the same action chosen on the basis of the account-level, adversary-ladder picture points towards information removal at the point of sharing and towards limits on aggregation. Those are different interventions, and only one of them binds an adversary with open weights.

The populations contrasted in Section 4 differ in more than their photographs. Grainge et al.'s four datasets differ in radius as well as composition for the street-view row, and EarthWhere's two sources differ in region, language and culture as well as platform. This is correct, and the paper does not claim that composition alone explains the spread. It claims that population, however decomposed, moves the number more than the model does, which the within-study comparisons support without needing a single cause.

The deflation section rests on three studies, one of them small. IMAGEO-Bench's private set has 220 photographs, KoreaGEO's 270, DoxBench's 500, and each covers a narrow geography. A larger clean study could find contamination effects that these miss, and Gemini's drop on new Seoul photographs at one kilometre is consistent with some exposure to Street View, as KoreaGEO's authors suggest. What the three studies rule out is the strong form of the deflation, that the capability is mostly memorisation. They do not rule out a moderate contamination premium on the public benchmarks, and the GeoSearch result for retrieval systems suggests there is one.

The numbers are already out of date. Most measurements in Table 1 use models from 2024 and 2025. Later models are likely to be more capable, and Tomekce et al. (2024) found attribute-inference accuracy rising with general capability. If so, the absolute levels in Table 1 understate current exposure. The structural argument does not depend on the levels, and the protocol is designed to be rerun as models change.

Several sources are recent preprints. The anchor, DoxBench, VLM-GeoPrivacy, KoreaGEO, GeoRC and GeoSearch carry venue acceptances in their own metadata; many others, including some that carry specific numbers in Table 1, are preprints from small teams. The paper relies on each for what it measured and states the conditions of each measurement on its row.

14. What is not known

The composition of the photographs that users of any platform actually share is not known, in the terms that matter here: what share carry legible text, landmarks, dense urban structure, people, or private interiors. Without it no benchmark number can be transferred to a population, and every exposure claim in the literature, including this paper's, is conditional on a population nobody has described.

Account-level exposure is not known. No published study reports, for a set of real or simulated accounts, the probability that an adversary reaches a harm radius for at least one sensitive place. The single-post aggregation result of Wang et al. (2026) and PlaNet's album result are the closest evidence, and both concern collections that were assembled to be geolocated, not the posting histories of ordinary people.

The cost of a located photograph is not known. No geolocation study reports money or time per photograph located within a stated radius, for a model and for a human baseline, so the scale claim on which the privacy argument most depends, that location inference is now cheap enough to run on everyone, rests on analogy with the text-inference results of Staab et al. (2023).

The durability of image-side perturbations against models released after the perturbation was tuned is not known, and cannot be known at publication time. The facial-recognition precedent suggests it is short, and the geolocation defences have not yet been tested against a later model.

Whether the cues a model reports are the cues it uses is not known for geolocation in general. GeoRC's finding that reasoning chains often cite absent features is evidence that reports and use come apart, but no study has tested, by intervening on reported cues, which ones carry the prediction. Removal defences that rely on reported cues inherit that uncertainty.

The marginal harm of a correct geolocation is not known, as distinct from its frequency. The literature counts hits at radii; it does not measure what happened to anyone as a result, and for good ethical reasons it cannot do so directly.

15. Conclusion

General-purpose vision-language models can locate photographs, and the evidence that they can do so for fresh, privately collected photographs is now stronger than the evidence usually cited for the claim. What the evidence does not support is reading any single benchmark accuracy as the size of the risk. Published geolocation accuracy is a function of the photo population, the harm radius, the adversary configuration, the baseline and the unit counted, and the published numbers fix at most two of these at a time. The same model moves from 61 percent within a kilometre on curated stock photographs to under 12 percent within a hundred kilometres on sampled street imagery. Some adversary rungs, reverse image search

and pooling several photographs, help most on the photographs a single-query benchmark scores as safe, though not every rung does. Defences differ chiefly in which adversary they can bind: deployment policies bind a product, image perturbations bind today's models, and only removal of the information binds every model. The exposure protocol of Algorithm 1 asks a privacy claim to name its population, radius, adversary rungs, baseline and unit before it reports a number. Until claims do, the right summary of the field is the one this paper gives: the capability is real, its distribution is extremely uneven, and its size for any actual population of users has not been measured.

References

Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V. N., et al. (2024). OpenStreetView-5M: The Many Roads to Global Visual Geolocation. CVPR 2024. arXiv:2404.18873. doi:10.48550/arXiv.2404.18873.

Barth, A., Datta, A., Mitchell, J. C., & Nissenbaum, H. (2006). Privacy and contextual integrity: framework and applications. 2006 IEEE Symposium on Security and Privacy. doi:10.1109/SP.2006.32.

Bharadwaj, S., Vashist, A., Aleem, F., & Vyas, S. (2026). Where Do Vision-Language Models Fail? World Scale Analysis for Image Geolocalization. CVPR EarthVision Workshop 2026. arXiv:2604.16248. doi:10.48550/arXiv.2604.16248.

Bicakci, Y. S., Shingleton, J., & Basiri, A. (2025). Street-Level Geolocalization Using Multimodal Large Language Models and Retrieval-Augmented Generation. arXiv:2509.01341. doi:10.48550/arXiv.2509.01341.

Chen, W., Cheng, Y., Yang, R., Mendes, E., Ritter, A., Sun, J., Li, H., & Chen, Y. (2026). Steering Vision-Language Models for Visual-Contextual Privacy in Location Disclosure. arXiv:2610.27323. doi:10.48550/arXiv.2610.27323.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, 3, 1376. doi:10.1038/srep01376.

DeVries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does Object Recognition Work for Everyone? arXiv:1906.02659. doi:10.48550/arXiv.1906.02659.

Dou, Z., Wang, Z., Han, X., Li, G., Huang, Z., & Han, Z. (2024). Towards Interactive Global Geolocation Assistant. arXiv:2412.08907. doi:10.48550/arXiv.2412.08907.

Fan, Y., Chen, J., Tian, Y., & Yin, Z. (2025). Who Can See Through You? Adversarial Shielding Against VLM-Based Attribute Inference Attacks. arXiv:2512.18264. doi:10.48550/arXiv.2512.18264.

Grainge, O., Waheed, S., Stilgoe, J., Milford, M., & Ehsan, S. (2025). Assessing the Geolocation Capabilities, Limitations and Societal Risks of Generative Vision-Language Models. AAAI Fall Symposium 2025. arXiv:2508.19967. doi:10.48550/arXiv.2508.19967.

Haas, L., Skreta, M., Alberti, S., & Finn, C. (2023). PIGEON: Predicting Image Geolocations. CVPR 2024. arXiv:2307.05845. doi:10.48550/arXiv.2307.05845.

Hays, J., & Efros, A. A. (2008). IM2GPS: estimating geographic information from a single image. 2008 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/CVPR.2008.4587784.

Honig, R., Rando, J., Carlini, N., & Tramer, F. (2024). Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI. arXiv:2406.12027. doi:10.48550/arXiv.2406.12027.

Hui, X., Wu, Q., Qu, H., Mirmehdi, M., Rahmani, H., & Liu, J. (2026). When Visual Privacy Protection Meets Multimodal Large Language Models. International Journal of Computer Vision, 134, 167. arXiv:2603.13978. doi:10.48550/arXiv.2603.13978.

Jay, N., Nguyen, H. M., Hoang, T. D., & Haimes, J. (2025). Evaluating Precise Geolocation Inference Capabilities of Vision Language Models. AAAI 2025 Workshop DATASAFE. arXiv:2502.14412. doi:10.48550/arXiv.2502.14412.

Ji, Y., Wang, Y., Ma, Z., Hu, Y., Huang, H., Hu, X., Chen, G., Wu, L., et al. (2026). Thinking with Map: Reinforced Parallel Map-Augmented Agent for Geolocalization. arXiv:2601.05432. doi:10.48550/arXiv.2601.05432.

Jia, P., Liu, Y., Li, X., Wang, Y., Du, Y., Han, X., Wei, X., Wang, S., et al. (2024). G3: An Effective and Adaptive Framework for Worldwide Geolocalization Using Large Multi-Modality Models. NeurIPS 2024. arXiv:2405.14702. doi:10.48550/arXiv.2405.14702.

Le-Duc, T.-D., Nguyen-Son, H.-Q., & Dao, M.-S. (2026). GeoSearch: Augmenting Worldwide Geolocalization with Web-Scale Reverse Image Search and Image Matching. SIGIR 2026. arXiv:2604.25390. doi:10.48550/arXiv.2604.25390.

Li, L., Ye, Y., Zhou, Y., Jiang, B., & Zeng, W. (2024). GeoReasoner: Geo-localization with Reasoning in Street Views using a Large Vision-Language Model. arXiv:2406.18572. doi:10.48550/arXiv.2406.18572.

Li, L., Yu, R., Hu, Q., Li, B., Deng, M., Zhou, Y., & Jia, X. (2025). From Pixels to Places: A Systematic Benchmark for Evaluating Image Geolocalization Ability in Large Language Models. arXiv:2508.01608. doi:10.48550/arXiv.2508.01608.

Li, Q., Xiao, Z., Wang, X., Ma, Z., Yang, C., & Li, H. (2026). LocationAgent: A Hierarchical Agent for Image Geolocation via Decoupling Strategy and Evidence from Parametric Knowledge. arXiv:2601.19155. doi:10.48550/arXiv.2601.19155.

Liu, X., Jia, X., Xun, Y., Qin, S., & Cao, X. (2025). GeoShield: Safeguarding Geolocation Privacy from Vision-Language Models via Adversarial Perturbations. AAAI 2026. arXiv:2508.03209. doi:10.48550/arXiv.2508.03209.

Luo, W., Lu, T., Zhang, Q., Liu, X., Hu, B., Zhao, Y., Zhao, J., Gao, S., et al. (2025). Doxing via the Lens: Revealing Location-related Privacy Leakage on Multi-modal Large Reasoning Models. ICLR 2026. arXiv:2504.19373. doi:10.48550/arXiv.2504.19373.

Mao, H., Wang, Y., Koo, M., Huang, W., Doan, H., Xia, J., Wang, Z., & Gonzalez, J. E. (2026). GeoGuesser Bench: Benchmarking Vision-Language Agents on Interactive Geolocation. arXiv:2611.02698. doi:10.48550/arXiv.2611.02698.

Mendes, E., Chen, Y., Hays, J., Das, S., Xu, W., & Ritter, A. (2024). Granular Privacy Control for Geolocation with Vision Language Models. EMNLP 2024. arXiv:2407.04952. doi:10.48550/arXiv.2407.04952.

Mukherjee, A., Roy, S., Trivedi, K., & Ghosh, S. (2026). GeoAgent: Evaluating VLM Geolocalization Through Embodied Navigation. EMNLP 2026 Findings. arXiv:2608.29483. doi:10.48550/arXiv.2608.29483.

Muller-Budack, E., Pustu-Iren, K., & Ewerth, R. (2018). Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. ECCV 2018, Lecture Notes in Computer Science. doi:10.1007/978-3-030-01258-8_35.

Qian, Z., Chen, H., Wang, Z., Zhang, L., Wang, Z., Huang, X., Liu, H., Tang, X., et al. (2025). Where on Earth? A Vision-Language Benchmark for Probing Model Geolocation Skills Across Scales. arXiv:2510.10880. doi:10.48550/arXiv.2510.10880.

Radiya-Dixit, E., Hong, S., Carlini, N., & Tramer, F. (2021). Data Poisoning Won't Save You From Facial Recognition. ICLR 2022. arXiv:2106.14851. doi:10.48550/arXiv.2106.14851.

Samson, L., Barazani, N., Ghebreab, S., & Asano, Y. M. (2024). Privacy-Aware Visual Language Models. Transactions on Machine Learning Research. arXiv:2405.17423. doi:10.48550/arXiv.2405.17423.

Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. USENIX Security Symposium 2020. arXiv:2002.08327. doi:10.48550/arXiv.2002.08327.

Song, Z., Yang, J., Huang, Y., Tonglet, J., Zhang, Z., Cheng, T., Fang, M., Gurevych, I., et al. (2025). Geolocation with Real Human Gameplay Data: A Large-Scale Dataset and Human-Like Reasoning Framework. arXiv:2502.13759. doi:10.48550/arXiv.2502.13759.

Staab, R., Vero, M., Balunovic, M., & Vechev, M. (2023). Beyond Memorization: Violating Privacy Via Inference with Large Language Models. arXiv:2310.07298. doi:10.48550/arXiv.2310.07298.

Staab, R., Vero, M., Balunovic, M., & Vechev, M. (2024). Large Language Models are Advanced Anonymizers. arXiv:2402.13846. doi:10.48550/arXiv.2402.13846.

Talreja, M., Diao, J., James, J. T., Casapu, R., Santanam, T., Mendes, E., Ritter, A., Xu, W., et al. (2026). GeoRC: A Benchmark for Geolocation Reasoning Chains. ACL 2026. arXiv:2601.21278. doi:10.48550/arXiv.2601.21278.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). YFCC100M: The New Data in Multimedia Research. Communications of the ACM, 59(2). doi:10.1145/2812802.

Tomekce, B., Vero, M., Staab, R., & Vechev, M. (2024). Private Attribute Inference from Images with Vision-Language Models. arXiv:2404.10618. doi:10.48550/arXiv.2404.10618.

Tsaprazlis, E., Feng, T., Ramakrishna, A., Gupta, R., & Narayanan, S. (2025). Assessing Visual Privacy Risks in Multimodal AI: A Novel Taxonomy-Grounded Evaluation of Vision-Language Models. arXiv:2509.23827. doi:10.48550/arXiv.2509.23827.

Tsaprazlis, E., Feng, T., Ramakrishna, A., Karimireddy, S. P., Gupta, R., & Narayanan, S. (2026). Rethinking Visual Privacy: A Compositional Privacy Risk Framework for Severity Assessment with VLMs. arXiv:2603.21573. doi:10.48550/arXiv.2603.21573.

Vivanco Cepeda, V., Nayak, G. K., & Shah, M. (2023). GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. NeurIPS 2023. arXiv:2309.16020. doi:10.48550/arXiv.2309.16020.

Vo, N., Jacobs, N., & Hays, J. (2017). Revisiting IM2GPS in the Deep Learning Era. arXiv:1705.04838. doi:10.48550/arXiv.1705.04838.

Wang, X., Shao, B., & Kim, H. (2025). AI Knows Where You Are: Exposure, Bias, and Inference in Multimodal Geolocation with KoreaGEO. EMNLP 2025. arXiv:2506.03371. doi:10.48550/arXiv.2506.03371.

Wang, Y., Li, X., Zhang, M., Zhang, X., You, X., Qian, Z., & Wen, M. (2026). Hiding in Plain Sight: A Diffusion-based Mitigation of Geolocation Privacy Leakage in Vision-Language Models. NDSS 2027. arXiv:2609.21363. doi:10.48550/arXiv.2609.21363.

Wang, Z., Xu, D., Khan, R. M. S., Lin, Y., Fan, Z., & Zhu, X. (2024). LLMGeo: Benchmarking Large Language Models on Image Geolocation In-the-wild. CVPR 2024 Workshop on Computer Vision in the Wild. arXiv:2405.20363. doi:10.48550/arXiv.2405.20363.

Wazzan, A., MacNeil, S., & Souvenir, R. (2024). Comparing Traditional and LLM-based Search for Image Geolocation. arXiv:2401.10184. doi:10.48550/arXiv.2401.10184.

Weyand, T., Kostrikov, I., & Philbin, J. (2016). PlaNet - Photo Geolocation with Convolutional Neural Networks. arXiv:1602.05314. doi:10.48550/arXiv.1602.05314.

Wu, B., Fang, M., Chen, L., Xu, K., Cheng, T., & Wang, J. (2026). Vision-Language Reasoning for Geolocalization: A Reinforcement Learning Approach. AAAI 2026. arXiv:2601.00388. doi:10.48550/arXiv.2601.00388.

Wu, Y., Li, M., Shen, Y., Backes, M., & Zhang, Y. (2026). GeoCloak: An Agentic Framework for Semantics-Preserving Geolocation Privacy via Peripheral Clue Manipulation. arXiv:2610.20055. doi:10.48550/arXiv.2610.20055.

Yang, R., Mendes, E., Wang, A., Hays, J., Das, S., Xu, W., & Ritter, A. (2026). Do Vision-Language Models Respect Contextual Integrity in Location Disclosure? ICLR 2026. arXiv:2602.05023. doi:10.48550/arXiv.2602.05023.

Yang, Y., Wang, S., Li, D., Zhang, Y., Sun, S., & He, J. (2023). GeoLocator: a location-integrated large multimodal model for inferring geo-privacy. Applied Sciences, 14(16), 7091. arXiv:2311.13018. doi:10.48550/arXiv.2311.13018.

Yi, Q., & Shan, L. (2025). GeoLocSFT: Efficient Visual Geolocation via Supervised Fine-Tuning of Multimodal Foundation Models. arXiv:2506.01277. doi:10.48550/arXiv.2506.01277.

Zhang, S., Zhang, G., Meng, Y., Zhang, Z., Zhao, H., Yi, X., & Li, H. (2025). Through Their Eyes: User Perceptions on Sensitive Attribute Inference of Social Media Videos by Visual Language Models. arXiv:2508.07658. doi:10.48550/arXiv.2508.07658.

Zhang, X., & Cheng, X. (2025). Evaluation of Geolocation Capabilities of Multimodal Large Language Models and Analysis of Associated Privacy Risks. arXiv:2506.23481. doi:10.48550/arXiv.2506.23481.

Zhang, X., Wu, Y., Zhang, B., Lin, C., Shen, C., Backes, M., & Zhang, Y. (2025). GEO-Detective: Unveiling Location Privacy Risks in Images with LLM Agents. arXiv:2511.22441. doi:10.48550/arXiv.2511.22441.

Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., & Mai, G. (2024). Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation. arXiv:2403.19584. doi:10.48550/arXiv.2403.19584.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24382475