Addressing Privacy Concerns of Image Geolocation Capabilities in Vision-Language Models

Neel Jay* Teja Nallagorla* Daniel Pitzele*

* Equal contribution, alphabetical order

1. Intro

In the Internet age, visual data has become ubiquitous and personal privacy has become increasingly scarce. Modern social media platforms encourage people to share images that give away environmental cues about their geographic location, whether it is intentional or not. Image geolocation refers to the inference of geographical information given an image or series of images, and it has become powerful but also potentially dangerous. Early on, image geolocation systems focused on using specific features in images and large retrieval datasets. Now, because of advancements in large vision-language models (VLMs), the possibilities of image geolocation have expanded significantly.

AI geolocation has been a long-studied field. State-of-the-art (SOTA) methods like PIGEON [3] show remarkable global geolocation performance using carefully customized training procedures and architectures. However, the authors of systems like this make them closed-source because of concerns about privacy, misuse, and spreading sensitive datasets. In recent years, another emerging trend has been using general-purpose large language models (LLMs) and VLMs for their wide array of reasoning capabilities [16]. This includes their strong performance on image geolocation tasks [5, 15]. VLMs are able to use key identifying features in images, such as vegetation, architecture, signage, and climate cues, to geolocate images with accuracy that rivals human experts. Their ability to chain reasoning, call external tools, and process multimodal information furthers the geolocation capabilities that are accessible to the general public.

The increasing accessibility to powerful VLMs expands both the intended utility and the potential for misuse of geolocation technology. For example, people can easily scrape images from social media accounts to find personal information like home locations, travel patterns, nationality, and socioeconomic status [14]. Such VLM agents could be misused for purposes like doxxing, targeted advertising, stalking, surveillance, etc. However, even with the rapid advancements in VLM-based geolocation, there is no systematic understanding of how these models behave, where and how they fall short, and how easy it is for someone to direct them to making privacy-sensitive inferences.

The aim of our project is to focus on privacy implications of VLM geolocation, and evaluate current systems we have to defend against these capabilities. Our contributions include an analysis of two different approaches to user privacy: (1) developing a model that can score an images ”risk of geolocation” and (2) exploring the use and limitations of adversarial filters to reduce VLM performance.

2. Related Works

2.1. Classical and Task-Specific Image Geolocation

Early geolocation systems looked at the task through the lens of large-scale visual matching. The influential im2gps system by Hays and Efros [4] used a dataset of more than 6 million GPS-tagged images for nearest-neighbor retrieval. This set the first global geolocation benchmark, but it also had its faults, such as coarse accuracy and biases specific to retrieval-based approaches. Aside from retrieval, Müller-Budack et al. [10] proposed using a hierarchical deep model that uses scene semantics such as sky, vegetation, and environment to make direct estimates about geographic regions. This model was one of the earliest non-retrieval-based deep learning systems used for geolocation.

More recent approaches use customized architectures and carefully curated datasets. PIGEON [3] reaches strong global geolocation performance by training on a large proprietary dataset (unreleased by the authors because of privacy concerns) which has optimizations at the architecture level to explicitly improve its geographic prediction capabilities. Other multimodal approaches also exist. Siński, Żychowski, and Mańdziuk [12] show that combining images with textual information such as region-specific cultural or regulatory facts can significantly improve a model’s localization accuracy.

These classical and task-specific methods depend on purpose-built datasets or architectures, whereas our work evaluates general-purpose, pretrained VLMs that were not explicitly designed for geolocation. Because modern VLMs are widely accessible for everyday use and have strong general reasoning capabilities beyond geolocation, their potential for misuse is substantially more concerning.

2.2. Geolocation Capabilities of Large Language and Vision-Language Models

Recent research shows that vision-language models are surprisingly accurate at global geolocation, even without explicit domain-specific training. Jay et al. [5] test VLMs' coordinate-level inference capabilities using chain-of-thought and tool-augmented reasoning, demonstrating strong performance across multiple datasets. Liu et al. [7] benchmark large vision-language models (LVLMs) on classic geolocation datasets such as im2gps and analyze their reasoning strategies to show that models rely on subtle cues like vegetation distributions, architectural styles, signage, and linguistic artifacts.

There has also been a growing body of work exploring robustness, prompt sensitivity, explanation quality, and scaling behavior of LVLM geolocation, collectively supporting the idea that geolocation is an emergent capability that becomes accessible once models actually grow to the needed scale and multimodal grounding is achieved. Importantly, these studies show that VLMs do not just memorize locations, but instead actually synthesize world knowledge with fine-grained visual evidence.

Our work extends this research by reframing geolocation performance as a privacy risk metric and using it to evaluate image-level defense mechanisms.

2.3. Privacy Risks, Attribute Inference, and Defense Mechanisms

A closely related area of research looks at the privacy risks that multimodal foundation models pose. Tömekçe et al. [14] demonstrate that VLMs can infer sensitive private information such as ethnicity or socioeconomic status from seemingly ”safe” or ”normal” everyday images. This raises concerns about profiling and surveillance. Mendes et al. [9] introduce GPTGeoChat, which is a dataset of annotated human-model geolocation conversations with varying granularity levels, and show that tool-augmented reasoning can bypass refusal-trained and instruction-based guards, which can lead to unintended coordinate-level information leakage.

Among defense mechanisms, GeoShield [6] suggests using a targeted adversarial perturbation framework specifically designed to mask any visual cues that could help with geolocation while also preserving overall image quality. GeoShield finds and suppresses features like architectural textures, skyline geometry, vegetation patterns, and atmospheric cues that VLMs implicitly depend on for accurately performing localization tasks. Unlike generic adversarial noise, GeoShield explicitly optimizes against geolocation inference. This shows that privacy-preserving transformations must be task-aware to have a chance at being effective.

While adversarial defenses are an intentional and optimized approach to privacy protection, they also raise an important question: to what extent does geolocation accuracy decrease due to simpler, non-adversarial image transformations?

To effectively contextualize the usefulness of adversarial methods like GeoShield, we also look at image transformations caused by aggressive compression and representation bottlenecks. Recent work on Nano Banana-style extreme image compression (designed to significantly reduce the amount of available visual information while still keeping the coarse semantic content) gives a natural baseline for evaluating geolocatability reduction without adversarial optimization [2]. Methods like these significantly distort high-frequency and fine-grained spatial details (which are generally needed for accurate geolocation) while still preserving object-level recognizability for standard downstream tasks. Past studies on compression and low-bit visual representations suggest that, while these transformations degrade performance on fine-grained recognition tasks, they do not reliably block task-specific inference unless explicitly optimized for that purpose [1].

By comparing Nano Banana-style compression effects with GeoShield's targeted adversarial changes, our work empirically evaluates whether intentional privacy defenses provide any meaningful advantages over minor information loss. This comparison helps to clarify whether the robustness of geolocation in VLMs comes from high-frequency visual cues or from deeper semantic and contextual reasoning that still exists in images even after significant visual degradation.

These works collectively show that geolocation alone is a powerful means of compromising privacy, since it can enable downstream personal attribute inference. Our study builds on this insight by directly measuring how different image-level transformations affect this information leakage.

2.4. Emergent Abilities in Large-Scale Foundation Models

Wei et al. [16] show that large language models develop qualitatively new behaviors (such as multi-step reasoning and abstraction) once they surpass certain scale thresholds. VLMs show similar emergent capabilities, which allows them to combine visual perception with linguistic and world knowledge in ways that go beyond their explicit objectives.

In geolocation settings, these emergent abilities let VLMs infer location by combining multiple weak visual cues like architectural styles, vegetation patterns, signage, and environmental context, rather than depending on explicit coordinate supervision. This explains why even general-purpose models that were trained on broad image–text data have strong geolocation performance.

Emergent reasoning also adds to the robustness of geolocation inference. Because localization depends on redundant and complementary cues, naive strategies like blurring, downsampling, or aggressive compression often fail to hinder the geolocatability of an image entirely. Models can re-route their reasoning using the remaining cues, making unintended inference difficult to suppress.

Although emergent abilities help explain why VLMs can geolocate so effectively, our work focuses on the privacy implications of these capabilities. By evaluating both adversarial defenses and non-adversarial transformations, we empirically analyze how resilient emergent geolocation is to different image-level privacy-preserving strategies.

3. Methodology

We first aim to create a model that scores the ”geolocatability” of an image. That is, given an image input, the model outputs the ease with which the image could be tracked to its real-life location on Earth by a VLM.

We then explore the efficacy of common adversarial filters, such as GeoShield, on reducing VLM performance. The robustness of these approaches to real-life threat scenarios are also stressed.

3.1. Data

Our first step towards this was to assemble a dataset of the best attempts of current VLMs at geolocating various images. We queried many (at the time) SOTA VLMs and asked them to predict, in latitude and logitutde coordinates, where in the world an image was taken. We then recorded their responses to be used as a ground truth for training our model. Note that these responses need not be perfectly accurate to the true location of the photograph. Since we aim to estimate the geolocatability of an image, the distance between the true location and the VLM’s guess is the metric we are most interested in.

For this milestone, we use VLM prediction data from 1602 images taken from the Google Street View Static API. These images are taken from urban areas around the world. For each VLM model, the image is labeled with a predicted coordinates, distance error, and model Chain-of-Thought. For the GPT-4o model, the Chain-of-Thoughts are tagged with several ”prediction categories” (e.g. Road and infrastructure, Urban layout and elements, etc.) that represent elements in the image the model used to make its prediction.

In the future, we hope to integrate more image datasets to have more training data and greater location diversity. Some image datasets of interest that contains coordinate labels include im2gps [4] and DoxBench [8].

3.2. Geolocatability Classifier

Our main model was a multi-way classifier meant to predict the accuracy at which a VLM can geolocate an image. We chose to fine-tune an existing model to leverage the patterns of general image data encoded in the weights of the pretrained model. Specifically, we chose to try both EfficientNet-B0 [13] and CLIP [11] as our base models. We chose EfficientNet for its lightweight and simple nature as a visual feature extractor. We chose CLIP for its large pretraining dataset and semantic associations with visual features.

To predict ”geolocatability” from our dataset, we explored three related deep learning model setups.

First, we treat VLM distance error a multi-class classification problem, stratifying images into buckets:

$$\begin{aligned} Neighborhood: &\quad [0, 10)\text{ km},\\ City: &\quad [10, 25)\text{ km},\\ Region: &\quad [25, 200)\text{ km},\\ Country: &\quad [200, 750)\text{ km},\\ Other/Global: &\quad [750, \infty)\text{ km}. \end{aligned}$$

We fine-tune both an EfficientNet-B0 classifier and a CLIP-based classifier on these distance error buckets using the dataset previously mentioned. We utilize a 70/15/15 train/validation/test split.

Second, we simplified the problem to a binary classification task. The model should predict whether an image is geolocatable within 200 km or not. The same EfficientNet-B0 is fine-tuned.

Finally, we use the collected Chain-of-Thought and prediction category labels to inform the model. There are 12 total categories, and every image is labeled with any amount of these. We trained a EfficientNet-B0 model to output a 12-dimensional logit vector:

$$P(x) = \big(p_1(x), \ldots, p_{12}(x)\big), \quad p_k(x) \in [0, 1],$$

where p_k is the probability that category k is present and useful in the image. The training loop uses BCEWithLogits loss weighted such that images with fewer active categories (e.g. less visually cluttered) are weighted higher. After the model comes up with category probabilities, these are fused via a weighted logistic fusion model, providing binary predictions for the same 200 km threshold.

3.3. Adversarial Filters

We implement the GeoShield adversarial filter, as well as two baseline filters: Random Noise and CLIP-PGD. In addition, the use of prompt-based diffusion model editing has been unexplored for the prupose of geolocation privacy. Thus, we evaluate two additional image edit methods: NanoBana and NanoBana w/ CoT. Description of all methods implemented are found below:

- **Random Noise:** Additive random noise is applied to each image under a fixed perturbation budget. The noise is sampled independently per pixel and doesn't target any specific semantic features or geolocation-related features.
- **CLIP-PGD:** Starting from the original image, pixels are updated via gradient descent (PGD [17]) to minimize cosine similarity between the CLIP embedding of the perturbed image and the embedding of the original image. This method aims to disrupt high-level semantic representations that a stronger VLM may use in an image while attempting to preserve the visual similarity.
- **GeoShield [6]:** GeoShield is a geolocation-specific filter that is optimized to suppress representations in the image that are useful for location inference. The method using PGD against a series of CLIP-like vision encoders. Randomized cropping and transformations are also applied to ensure the filter is robust.
- **NanoBanana:** NanoBanana is the name of Google's SoTA text/image-to-image model. As of writing it leads various popular benchmarks, like GenAI-Bench, in many categories. The model is prompted to generate a subtly edited version of the original image, preserving the overall scene while weakening geolocation-relevant cues such as architecture, signage, and environmental details. The edits are photorealistic and aim to confuse a potential geolocation model.
- **NanoBanana w/ CoT:** In this variant, NanoBanana is also provided with the Chain-of-Thought provided by a separate VLM (Gemini 1.5 Pro) when geolocating the image. The reasoning explicitly highlights what visual elements were used by the VLM to infer the image's location, guiding NanoBanana to make more targeted and effective semantic edits.

All methods are applied to a sample of 20 images from our dataset, 5 from 4 different buckets (see Section 3.2). The strength of the three adversarial filters are thresholded to 0.36 LPIPS comparing the clean and filtered images. LPIPS (Learned Perceptual Image Patch Similarity) is a perceptual similarity metric based on a learned model. My matching LPIPS across filters, we control for perceptual distortions that human might observe when seeing the image.

4. Results

To validate our approach, we will evaluate the geolocatability model on a test set of images labeled with coordinates and VLM guesses. Predicted "geolocatability scores" will then be compared with the actual localization errors of the SOTA VLM to measure correlation and predictive confidence. We will also validate model generalization across different environments (e.g., urban, rural, coastal) and image sources to determine robustness. Performance will be assessed using metrics such as mean absolute localization error and R^2 correlation with ground truth performance to determine the accuracy and reliability of the model.

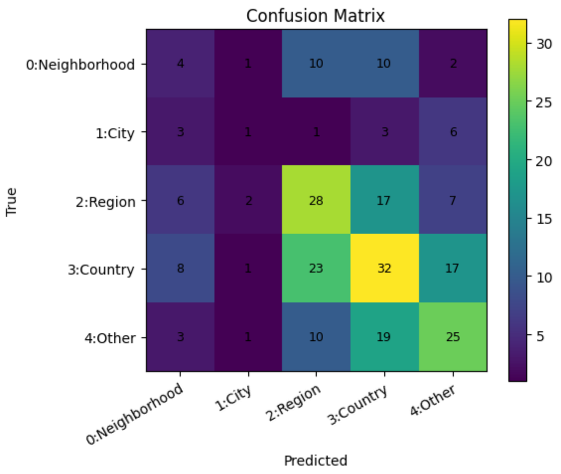

Figure 1. Confusion matrix for the five-way EfficientNet-B0 distance classification model.

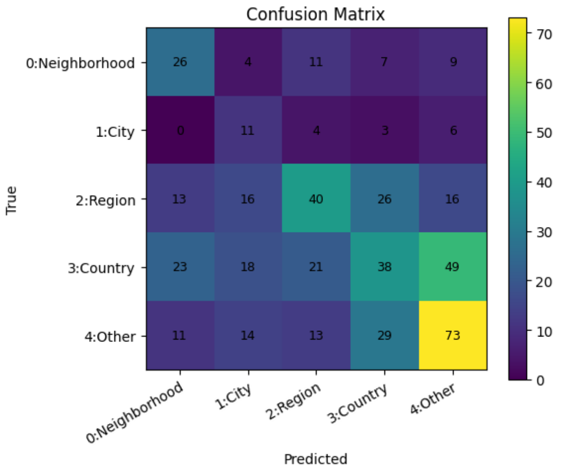

Figure 2. Confusion matrix for the five-way CLIP-based distance classification model.

We expect our results to demonstrate that specific visual and contextual features (e.g., vegetation type, architecture, signage) can dramatically impact an image's geolocatability. The objective is to develop an estimation system that can predict how geolocatable a specific image would be for VLMs without running the entire geolocation pipeline. The final deliverables will include (1) a dataset of image-geolocation error pairs, (2) an estimated geolocatability model learned from this dataset, and (3) a report identifying key factors that contribute to geolocation vulnerability. These results will provide valuable information about the risks to visual privacy during the age of multimodal AI systems.

4.1. Geolocatability

Across all three initial experiments, we observe that image geolocability prediction is possible but imperfect with the data we have so far.

The 5-bucket EfficientNet-B0 model achieved limited performance (validation accuracy of $\approx 37.5\%$). Figure 1 shows that confusions were common between adjacent buckets. However, the seperation between broader categories was more stable.

The CLIP-based classifier had a very similar performance with validation accuracy of $\approx 40\%$. There is a notable difference in the confusion matrices. Note that the CLIP-based classifier was trained on a merged dataset of responses from Gemini and Claude, doubling the size of the sample (but repeating each image twice in the training set). The CLIP-based model seems to confuse categories more equally than the EfficientNet-B0 model, which confuses closer categories more often. This may be an indicator of worse generalization from the CLIP model.

The binary classifier produced slightly stronger results, albeit not much better than random guessing. With a 200km threshold, the EfficientNet-B0 classifier achieved a test accuracy of 61.8% and ROC-AUC of 0.64.

Finally, category-informed pipeline provided a more interpretable method. However, we did not see much increase in performance beyond the previous experiment. The fusion model has a test accuracy of 54.4% and AUC of 0.59.

One advantage is that we are able to see the relationships between visual categories and model prediction. The learned coefficients reveal which categories the model associated with ”easy” images (¡200 km). Categories such as *Signage* (+0.958), *Architecture* (+1.100), *Urban layout* (+0.492), and *Road and infrastructure* (+0.326) had strong positive weights, indicating that the presence of these features substantially increases the likelihood that a VLM can geolocate the image accurately. Categories like *Environment and climate* (−1.409), *Lighting and shadows* (−1.286), *Other cultural elements* (−0.726), and *Other* (−0.635) had strong negative weights. These categories might not be useful for the VLM model or be generally misleading.

The main purpose of doing these two different trainings with so many variations was to show the limitations of our dataset. Despite varying many aspects of the dataset and training logic, our model had consistent metric across many experiments. This indicates that our issue was likely related to the size and quality of our dataset, and improving this could yield the greatest returns in terms of model performance.

4.2. Adversarial Filters

Figure 5 and Figure 6 show the median distance error in kilometers after prompting Gemini to geolocate each of the image variants. The figures show results for 2 different buckets. In the 10-25 km bucket, the CLIP-PGD filter manages to affect model performance the most. For the 200-500 km bucket, many filters resulted in similar performance to clean images; however, the NanoBanana-edited images had slightly higher error. In both cases, as well as the other 3 buckets, the GeoShield image variants elicited a significantly higher model distance error, often by orders of magnitude.

An example of the CLIP-PGD and GeoShield filters being applied at 0.36 LPIPS can be seen in Figure 7. It is notable that at this strength, the filters are relatively easy to notice through human observation. To test the robustness of the GeoShield filter in a more realistic use case, we lower the residual strength of the filter until it is no longer perceptible to a 3rd party viewer. Then, we compare distance error of GPT-5.1 with and without Agent Mode on these filtered images. Agent Mode allows the model to utilize tools such as a web browser and code editor for deeper reasoning. Figure 8 shows that while the weak filtered images still decreases VLM performance, allowing the VLM to use tools brings performance back to levels similar to the clean image.

4.3. Discussion

Our work has covered two main methods of enhancing user location privacy in the age of VLMs: geolocatability models and adversarial defense filters.

The geolocatability model results show promise but are limited by the amount of data and the noise that comes with using VLM distance error for loss. We hope to incorporate more datasets and SoTA model predictions in the future. We also hope to look into different visual category definitions and find ways to better incorporate them as features for the geolocatability models.

Our tests of adversarial filters show that location-feature aware filters like GeoShield also show promise. However, we also find that VLM agents may be able to override some of the defenses that these methods provide. Image-editing techniques using modern tools like NanoBanana weren’t as effective in our study. However, as visual understanding of these models increases, there is a lot of room for this approach to grow. Future work could refine these filter methods further with NanoBanana and/or human experts.

5. Contributions

- **Neel Jay:** Main contributor for adversarial filter evaluation, also contributed to geolocability model training. For writing, contributed to methodology, results, and discus-

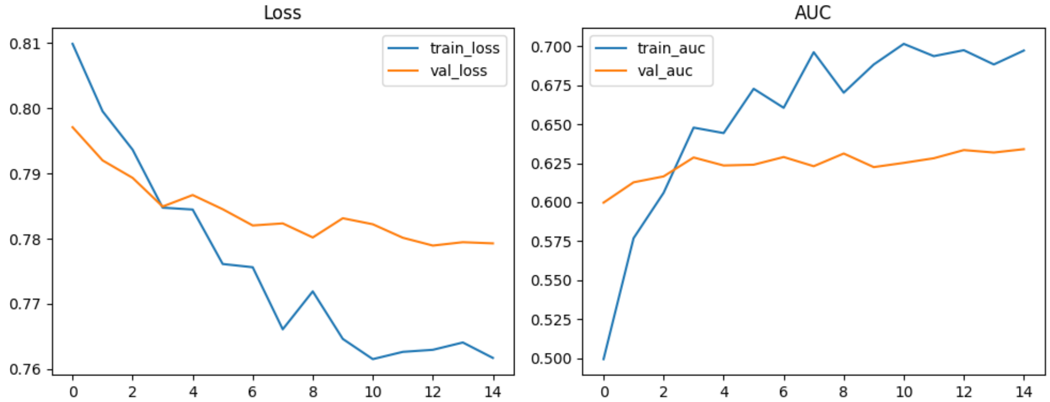

Figure 3. Training and validation AUC/loss curves for the binary classifier.

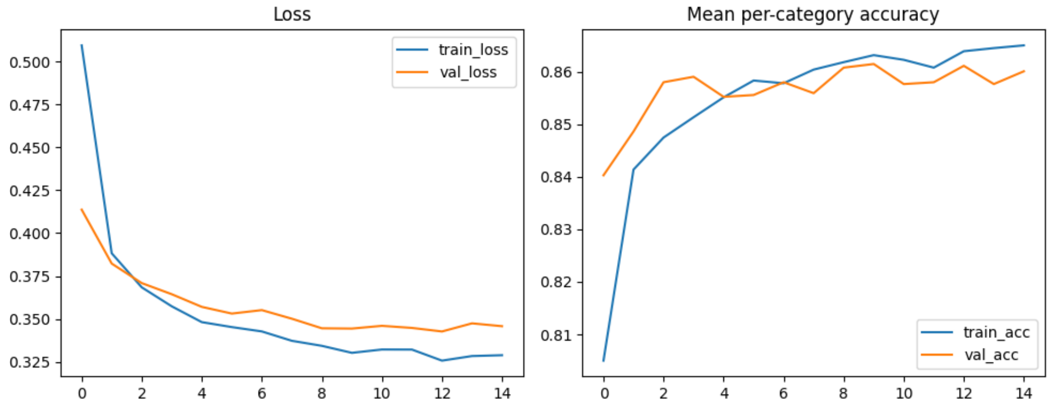

Figure 4. Training curve for category prediction model.

sion.

- **Teja Nallagorla:** Main contributor to the background research and writing for the Intro and Related Works sections, while assisting elsewhere as needed.
- **Daniel Pitzele:** Main contributor to the training and evaluation of the geolocatability model(s). For writing, contributed to methodology and results sections heavily and somewhat to the background section.

References

[1] Saeed Ranjbar Alvar and Ivan V. Bajić. Scalable privacy in multi-task image compression. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2021. 2

[2] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression, 2017. 2

[3] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations, 2024. 1

[4] James Hays and Alexei A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 3

[5] Neel Jay, Hieu Minh Nguyen, Trung Dung Hoang, and Jacob Haimes. Evaluating precise geolocation inference capabilities of vision language models, 2025. 1, 2

[6] Xinwei Liu, Xiaojun Jia, Yuan Xun, Simeng Qin, and Xiaochun Cao. Geoshield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations, 2025. 2, 4

[7] Yi Liu, Junchen Ding, Gelei Deng, Yuekang Li, Tianwei Zhang, Weisong Sun, Yaowen Zheng, Jingquan Ge, and Yang Liu. Image-based geolocation using large vision-language models, 2024. 2

[8] Weidi Luo, Tianyu Lu, Qiming Zhang, Xiaogeng Liu, Bin Hu, Yue Zhao, Jieyu Zhao, Song Gao, Patrick McDaniel, Zhen Xiang, and Chaowei Xiao. Doxing via the lens: Revealing location-related privacy leakage on multi-modal large reasoning models, 2025. 3

[9] Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models, 2024. 2

[10] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model

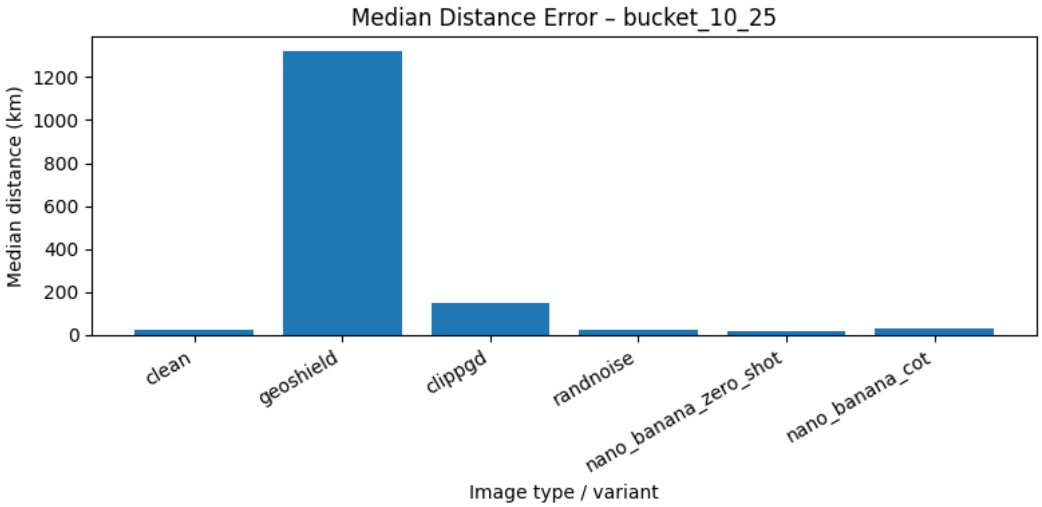

Figure 5. Median kilometer distance error when prompting Gemini 3 Pro to geolocate all image variants. These image are taken from the 10-25 km bucket, meaning clean images are typically geolocated within that range.

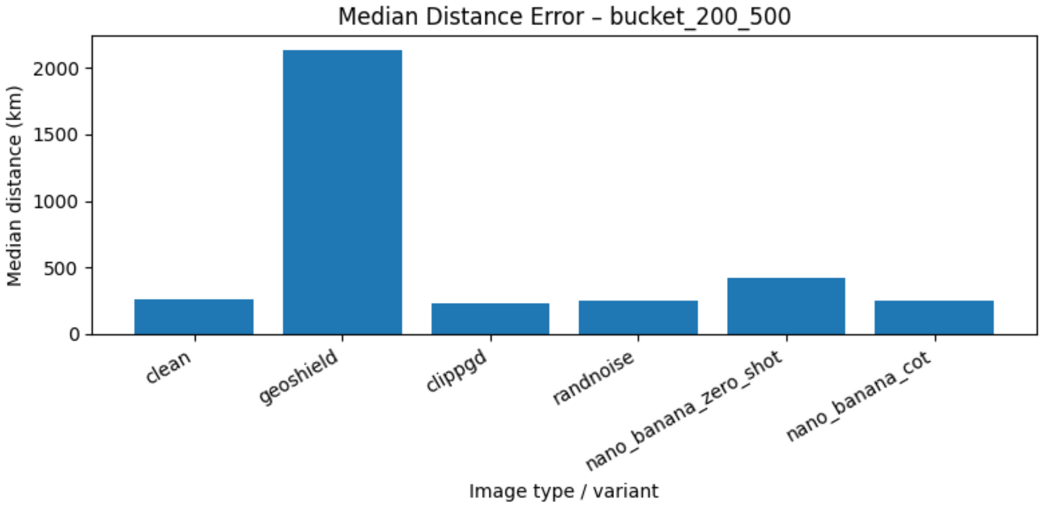

Figure 6. Median kilometer distance error when prompting Gemini 3 Pro to geolocate all image variants. These image are taken from the 200-500 km bucket, meaning clean images are typically geolocated within that range.

and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 3

[12] Bartosz Siński, Adam Żychowski, and Jacek Mańdziuk. Improving image geolocation with multimodal deep learning. In *Neural Information Processing*, pages 30–45, Singapore, 2025. Springer Nature Singapore. 1

[13] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 3

[14] Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models, 2024. 1, 2

[15] Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. Llmgeo: Benchmarking large language models on image geolocation in-the-wild, 2024. 1

[16] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean,

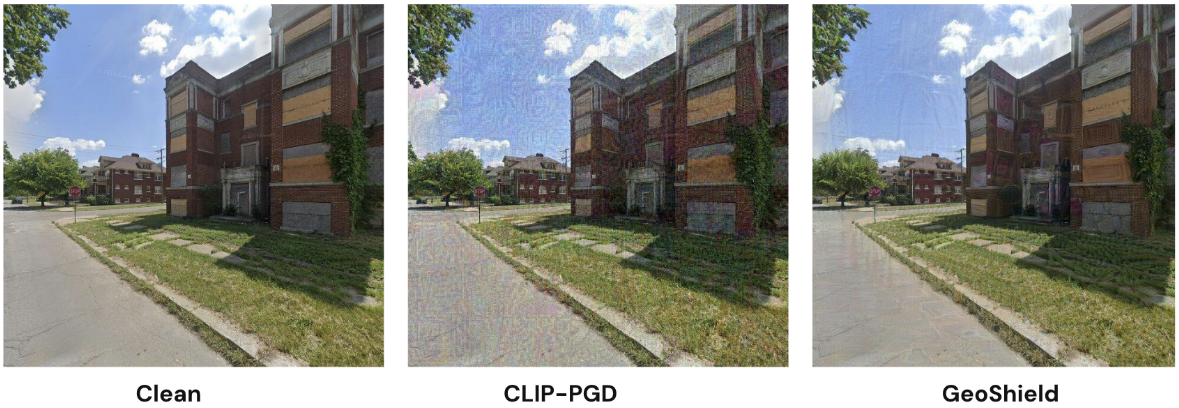

Figure 7. Clean sample simage vs CLIP-PGD vs GeoShield. Filters are thresholded to 0.36 LPIPS.

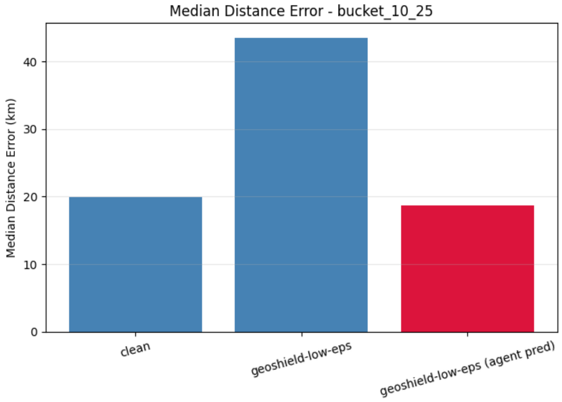

Figure 8. Median distance error comparison for imperceptible GeoShield filter (0.26 LPIPS). The blue bars (2 left bars) shows error from GPT-5.1 on clean vs GeoShield images. The red bar (right bar) shows error from GPT-5.1 w/ Agent Mode on GeoShield images.

and William Fedus. Emergent abilities of large language models, 2022. 1, 2

[17] Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack, 2018. 4