GeoErase: Sparse Adversarial Perturbations for Geolocation Privacy Protection

Anonymous ACL submission

Abstract

Vision-language models (VLMs) exhibit strong geolocation inference capabilities, posing serious challenges to personal geolocation privacy. Existing adversarial protection methods typically perturb the entire image, often introducing visible artifacts such as text-like noise and reducing protection stealthiness. To address this problem, we propose **GeoErase**, a black-box geolocation privacy protection framework that erases geolocation-revealing cues from model perception, preserves visual content, and balances protection effectiveness with visual stealthiness. GeoErase first uses feature attribution to precisely localize geolocation-sensitive regions and suppresses geographical cue features through perturbation optimization. It further introduces pixel-wise adaptive bounds and group sparse constraints, encouraging more structured and visually natural perturbations. Experimental results on two representative geolocation benchmarks show that GeoErase achieves stronger protection than previous work. On the latest GPT-5.4, GeoErase reduces the 1 km localization accuracy from 4.6% with previous work to 2.1%, indicating stronger suppression of location inference. Meanwhile, GeoErase improves perceptual quality, reducing LPIPS from 0.4056 to 0.3790. These results demonstrate that attribution-guided feature disentanglement and adaptive sparse optimization provide effective and visually natural geolocation privacy protection. Code is available at `https://anonymous.4open.science/r/GeoErase-8DB9`.

1 Introduction

Recent studies show that pre-trained vision-language models (VLMs) can infer geographic locations with high accuracy from a single image, even without explicit geolocation fine-tuning, by exploiting fine-grained visual cues such as architectural styles, vegetation patterns, and environmental structures (Huang et al., 2025). However, this strong reasoning capability also raises serious geolocation privacy risks. VLMs can extract geographic coordinates from ordinary images shared by users in daily scenarios, which may lead to the leakage of sensitive personal location information (Wang et al., 2025; Grainge et al., 2025). As the visual understanding ability of these models continues to improve, protecting user geolocation privacy while preserving the utility of multimodal technologies becomes a practical and urgent research challenge.

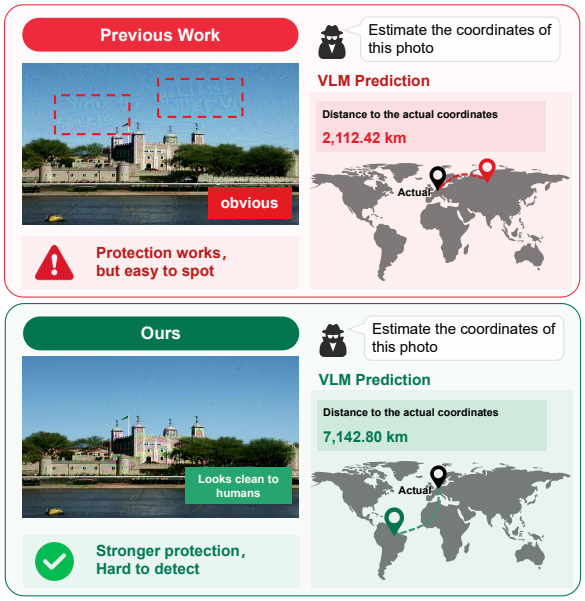


Figure 1: Comparison of geolocation privacy protection between previous work and our approach. Previous work results in obvious visual artifacts and limited prediction error, whereas our approach achieves higher visual fidelity and superior protection strength

Existing adversarial protection methods have shown promise in disrupting VLM-based geolocation inference (Liu et al., 2026; Zhang et al., 2025a), but two key limitations remain. First, many methods impose globally uniform perturbation constraints, ignoring regional variations in texture com-

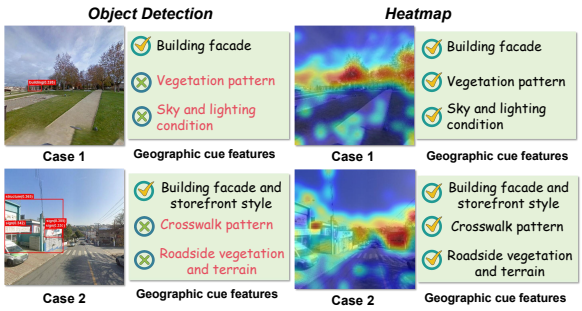


Figure 2: Qualitative comparison between object detection and heatmap. Object detection results are generated by GroundingDINO, while heatmaps are generated by GeoCLIP to highlight geolocation-sensitive regions.

plexity and perceptual sensitivity. As a result, perturbations may become visible in smooth regions while failing to concentrate on geolocation-relevant cues, which weakens the effectiveness of privacy protection (see Fig. 1). Second, previous geoprivacy protection methods typically rely on object detection to extract geographic cue features from images. However, object detection primarily focuses on explicit objects and may overlook implicit cues, such as vegetation, road layouts, and lighting conditions, which are also critical for geolocation inference (see Fig. 2). Consequently, the localized features may fail to accurately capture the information that is truly relevant to geolocation.

Motivated by these observations, we propose GeoErase. Unlike methods that use object detection to locate geolocation-related objects, GeoErase uses heatmaps to find geolocation-sensitive cues more accurately. It further introduces an adaptive perturbation bound and a structured group sparsity constraint to place perturbations mainly in important regions and reduce visible artifacts. This design improves the balance between geolocation protection and visual quality. The main contributions of this paper are summarized as follows:

- We identify the limitation of object detection in localizing geolocation-sensitive cues and propose a feature attribution strategy to capture both explicit and implicit geographic information, thereby extracting more precise geographic features.
- We introduce a pixel-wise adaptive perturbation bound based on local texture complexity and a structured group sparsity constraint, addressing the problem of insufficient perturbation concentration and visible artifacts in smooth regions.
- We evaluate GeoErase on two geolocation benchmarks and multiple black-box VLMs, showing improved geolocation obfuscation and a better trade-off among protection effectiveness, semantic preservation, and perceptual quality compared with strong adversarial baselines.

2 Related Work

2.1 Geolocation with Vision-Language Models

Image geolocation aims to infer the geographic location of an image from its visual content. Early methods mainly rely on image retrieval and matching against geo-tagged databases, such as IM2GPS (Hays and Efros, 2008). Later approaches formulate geolocation as a classification problem over discretized geographic regions, as represented by PlaNet (Weyand et al., 2016). Recent methods further improve visual-location representation learning through image-GPS alignment or semantic geocell modeling, such as StreetCLIP (Haas et al., 2023) and PIGEON (Haas et al., 2024). With the development of VLMs, image geolocation increasingly becomes a multimodal reasoning task, where models infer locations from landmarks, architecture, vegetation, signs, and other visual cues (Wang et al., 2024; Jay et al., 2025). However, this capability raises serious geoprivacy concerns, since VLMs may reveal sensitive locations from ordinary shared images (Luo et al., 2025). Meanwhile, VLMs are vulnerable to adversarial perturbations (Zhao et al., 2023), suggesting that carefully designed perturbations can mislead geolocation inference and may serve as a potential privacy protection mechanism.

2.2 Sparse Adversarial Perturbations

Sparse adversarial perturbations aim to fool neural networks by modifying only a small number of pixels or spatial regions, rather than applying dense noise to the entire image. Early sparse attacks select highly influential input dimensions using saliency-based or search-based strategies, such as JSMA (Papernot et al., 2016) and One-Pixel Attack (Su et al., 2019). Later methods directly optimize sparse perturbations under ℓ_0-related constraints, including geometry-based sparse attacks such as SparseFool (Modas et al., 2019), black-box pixel selection methods such as CornerSearch (Croce and Hein, 2019),

and factorization-based sparse optimization methods such as SAPF (Fan et al., 2020). Beyond pixel-level sparsity, structured sparse attacks introduce group or spatial constraints to make perturbations more coherent and interpretable, as shown in StrAttack (Xu et al., 2018) and exact group sparsity-based transferable attacks (Ming et al., 2024). However, existing sparse perturbation methods are mainly designed for image classification or general adversarial transfer, and their potential for geolocation privacy protection remains underexplored.

3 Methodology

3.1 Preliminary

Problem Definition. Given an input image x with ground-truth coordinates y, we generate a protected example $x' = x + \delta$ by adding a perturbation δ. The target model $f(\cdot)$ predicts the coordinates as $\hat{y} = f(x')$. The objective is to maximize the Haversine distance $D(\cdot, \cdot)$ between the predicted coordinates and the ground truth, i.e. computing

$$\max_{\delta} D(f(x + \delta), y). \tag{1}$$

To preserve visual imperceptibility, we impose a sparsity constraint $\|\delta\|_0 \leq K$, where K denotes the maximum number of modified pixels. During optimization, we approximate this constraint with group sparsity, which retains perturbations only in critical spatial regions and induces a structured sparse pattern. After each update, the adversarial image is clipped to the valid pixel range.

Overview. Fig. 3 shows the overall framework of GeoErase. The upper panel presents the joint global-local optimization process, which suppresses geographic features while preserving non-geographic semantics. The lower panels summarize FAGD and PASSP: FAGD localizes geolocation-sensitive regions and extracts feature anchors, while PASSP applies adaptive bounds and structured sparsity. The main symbols used in this section 3 are summarized in Table 1.

Symbol	Description
x, x'	Original and protected images
δ	Adversarial perturbation
$y, \hat{y}$	True and predicted coordinates
$D(\cdot, \cdot)$	Haversine distance between two coordinates
$f_{\theta_i}, g_{\theta_i}$	The i-th visual and textual encoders
N	Number of surrogate encoders
$S(\cdot, \cdot)$	Cosine similarity
z_{geo}	Estimated geographic feature
$z_{\text{non-geo}}$	Non-geographic semantic feature
$T_{\text{non-geo}}$	Location-agnostic image description
z_{feature}	Attribution-guided geographic feature
H	Geolocation-sensitive attribution heatmap
$\mathcal{M}$	structured sparsity mask
$\mathcal{L}_{\text{base}}$	Base multi-scale feature loss
$\mathcal{L}_{\text{total}}$	Final optimization objective
β	Weight for semantic preservation
γ	Weight for attribution-guided feature repulsion
σ_{ij}	Local texture complexity at pixel (i, j)
ϵ_{ij}	Pixel-wise adaptive perturbation bound
κ	Adaptive bound coefficient
$G_{p,q}$	Spatial perturbation group
$I_{p,q}$	Importance score of group $G_{p,q}$
k_{ratio}	Ratio of retained groups

Table 1: Summary of the main symbols used in GeoErase.

3.2 Feature Attribution-Based Geographic Information Disentanglement (FAGD)

In constructing the proposed geolocation privacy protection framework, we retain the semantic preservation and multi-scale robustness mechanisms of GeoShield (Liu et al., 2026). We first query an auxiliary VLM with a location-agnostic prompt to generate $T_{\text{non-geo}}$, which describes only generic visual content while avoiding geographic identifiers, including place names, coordinates, landmarks, and region-specific cues. The detailed prompt template and filtering protocol are provided in Appendix B.1. We then encode $T_{\text{non-geo}}$ as $z_{\text{non-geo}} \approx g_{\theta_i}(T_{\text{non-geo}})$, which serves as an approximate reference for the non-geographic semantic content of the image. Based on this reference, we compute $z_{\text{geo}} \approx f_{\theta_i}(x) - z_{\text{non-geo}}$ as a coarse residual feature related to geographic cues. Here, z_{geo} is not intended to represent an exact or uniquely disentangled geographic component in the CLIP embedding space. Instead, it provides a global optimization direction for reducing geolocation-relevant information while preserving general visual semantics.

For the protected image $x' = x + \delta$, each optimization iteration extracts the global feature $f_{\theta_i}^{\text{global}}(x')$ and computes the average feature over randomly sampled local patches as $f_{\theta_i}^{\text{local}}(x')$. To simplify notation, we define the multi-scale similarity with respect to a target feature $z \in \{z_{\text{geo}}, z_{\text{non-geo}}\}$ as,

$$\mathcal{S}_i^{\text{total}}(x', z) = \mathcal{S}(f_{\theta_i}^{\text{global}}(x'), z) + \mathcal{S}(f_{\theta_i}^{\text{local}}(x'), z), \tag{2}$$

where $\mathcal{S}(\cdot, \cdot)$ denotes the cosine similarity between

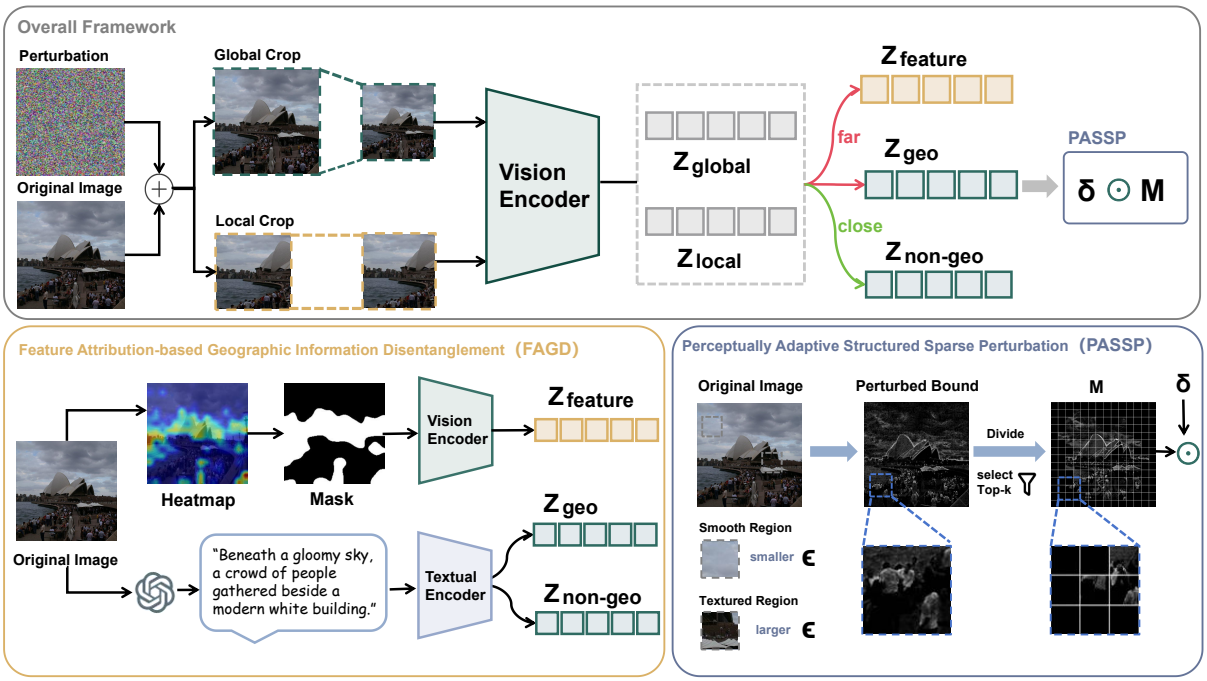


Figure 3: Overview of the proposed geolocation privacy protection framework. The framework first generates multi-scale features through global and local crops, and then optimizes the protected perturbation $\delta \odot \mathcal{M}$ by suppressing geo-related features while preserving non-geographical semantics. FAGD localizes geo-sensitive regions with attribution heatmaps and masks, while PASSP imposes perceptually adaptive bounds and structured sparsity for imperceptible perturbation.

two feature vectors. The base loss $\mathcal{L}_{\text{base}}$ moves the multi-scale representations away from the coarse geographic residual feature while keeping them close to the non-geographic semantic reference,

$$\mathcal{L}_{\text{base}} = \sum_{i=1}^{N} \left[\mathcal{S}_i^{\text{total}}(x', z_{\text{geo}}) - \beta \mathcal{S}_i^{\text{total}}(x', z_{\text{non-geo}}) \right], \tag{3}$$

where N denotes the number of vision encoders in the ensemble, and β is the semantic preservation weight.

Since the residual feature provides only a coarse global cue, we further introduce a feature heatmap-based strategy to localize regions that are more directly sensitive to geolocation inference.

Attribution-Guided Geographic Feature Anchoring. To obtain an explicit anchor for geolocation-sensitive cues, we first leverage the geolocation prior encoded in GeoCLIP (Vivanco Cepeda et al., 2023). Since GeoCLIP aligns visual patterns with global geographic coordinates during pre-training, its deep features are sensitive to location-relevant cues. Inspired by Grad-CAM (Selvaraju et al., 2016), we register forward and backward hooks at the last LayerNorm layer of the encoder. The forward hook records the spatial patch feature sequence A, while the backward hook captures the gradients of the similarity score Y with respect to these features. We globally average the gradients over all spatial patches to obtain the importance weight α_k for each feature channel,

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y}{\partial A_{i,j}^k}, \tag{4}$$

where Z denotes the total number of spatial patches. These channel weights are used to linearly combine the spatial feature maps, followed by ReLU and upsampling, producing the raw geolocation-sensitive heatmap H. Since H may contain scattered activations and blurred boundaries, we further refine it with normalization, exponential enhancement, Gaussian smoothing, Otsu thresholding, and median filtering to obtain a binary attribution mask $Mask$. The masked image region is then extracted as $x_{\text{mask}} = x \odot Mask$ and fed into each vision encoder to obtain the attribution-guided geographic feature,

$$z_{\text{feature}} = f_{\theta_i}(x_{\text{mask}}). \tag{5}$$

This feature serves as an explicit geographic anchor. During optimization, we push the global and local features of the perturbed image away from

this anchor, resulting in the final attribution-guided objective,

$$\min_{\delta} \Big\{ \mathcal{L}_{\text{base}} + \gamma \sum_{i=1}^{N} \big[\mathcal{S}(f_{\theta_i}^{\text{global}}(x'), z_{\text{feature}}) + \mathcal{S}(f_{\theta_i}^{\text{local}}(x'), z_{\text{feature}}) \big] \Big\}, \tag{6}$$

where γ controls the strength of attribution-guided feature repulsion. The objective is optimized iteratively using I-FGSM (Kurakin et al., 2018).

3.3 Perceptually Adaptive Structured Sparse Perturbation (PASSP)

Existing geolocation privacy protection methods typically impose globally uniform norm constraints, such as fixed ℓ_∞ or ℓ_2 bounds, assigning the same perturbation budget to all pixels. This allocation is suboptimal for high-resolution natural images, where geolocation-relevant cues often appear in high-texture regions, such as architectural structures, street signs, and vegetation, while smooth regions, such as sky, water, and walls, contain fewer geographic signals and are more sensitive to visible noise. As illustrated in Fig. 4(a), perturbations under global constraints tend to spread across the image and produce noisy responses in smooth backgrounds. In contrast, our adaptive perturbations are more spatially structured and concentrated around visually informative regions. This motivates a perception-adaptive perturbation bound that assigns larger budgets to textured, geo-informative regions while suppressing unnecessary perturbations in smooth areas.

Perception-Adaptive Bound. To address the limitation of globally uniform constraints, we introduce a pixel-wise adaptive perturbation bound based on local texture complexity. For each pixel (i, j) in the input image x, we compute local statistics on the normalized image. Specifically, we calculate the standard deviations along the horizontal and vertical directions, denoted as std_x^{ij} and std_y^{ij}, respectively, and define the local texture complexity as

$$\sigma_{ij} = \sqrt{\min(\text{std}_x^{ij}, \text{std}_y^{ij})}. \tag{7}$$

Based on this measure, the pixel-wise perturbation bound is defined as

$$\epsilon_{ij} = \kappa \cdot \sigma_{ij} \cdot x_{ij}, \tag{8}$$

where κ controls the overall perturbation strength and x_{ij} denotes the pixel intensity. The texture term σ_{ij} assigns smaller budgets to smooth regions and larger budgets to textured regions, while the intensity term x_{ij} further suppresses perturbations in low-intensity areas to avoid visually salient relative changes. Together, they reduce artifacts in sensitive regions while preserving sufficient perturbation capacity around texture-rich geolocation cues. A detailed discussion of this design is provided in Appendix D.

During optimization, the adaptive bound only constrains the perturbation update:

$$\delta_{ij} \leftarrow \text{Clip}(\delta_{ij}, -\epsilon_{ij}, \epsilon_{ij}), \quad \forall (i, j). \tag{9}$$

The protected image is further clipped to the valid pixel range $[0, 255]$, ensuring that the final output remains within the valid image domain.

Structured Group Sparsity. Although the perception-adaptive bound limits the perturbation magnitude, it does not explicitly control the spatial density of perturbations. We therefore introduce a structured group sparsity constraint to concentrate perturbations in high-utility regions. Specifically, we divide the perturbation δ into non-overlapping spatial groups $\{G_{p,q}\}$ with stride s, and compute the importance of each group using its ℓ_2 norm,

$$I_{p,q} = \|\delta_{p,q}\|_2. \tag{10}$$

Let $k_{\text{ratio}} \in (0, 1]$ denote the sparsity ratio, and let k be the corresponding number of retained groups determined by this ratio. We then retain only the top-k groups with the largest importance scores and set the remaining groups to zero,

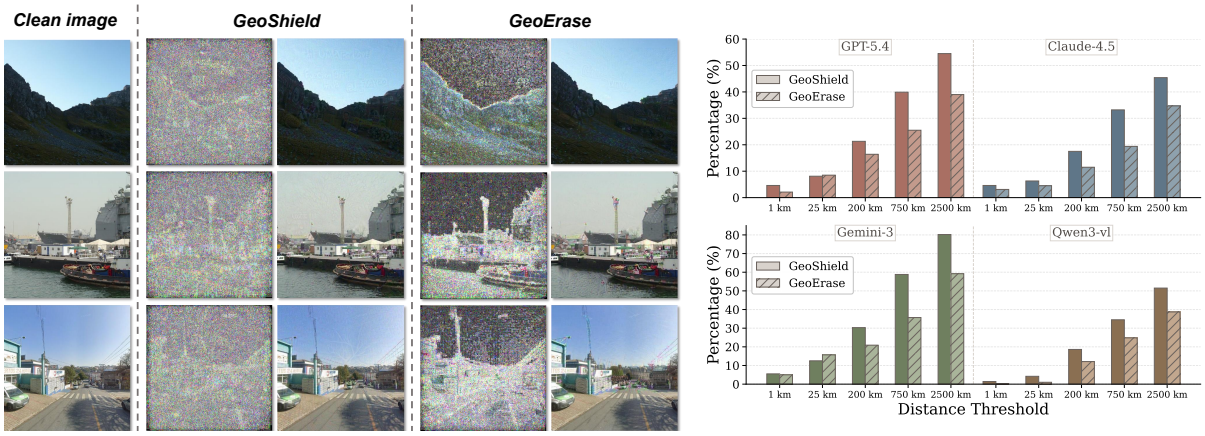
$$\delta \leftarrow \delta \odot \mathcal{M}, \quad \mathcal{M}_{p,q} = \begin{cases} 1, & \text{if } I_{p,q} \in \text{Top-}k, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This operation approximates a group-level ℓ_0 constraint and reduces unnecessary perturbation.

Optimization Target. We combine the above feature objective and spatial constraints into the final optimization problem,

$$\min_{\delta} \mathcal{L}_{\text{total}}(x + \delta \odot \mathcal{M}), \quad \text{s.t.} \quad |\delta_{ij}| \le \epsilon_{ij}, \quad \forall (i, j). \tag{12}$$

Here, $\mathcal{M}$ is the structured sparsity mask and $\epsilon_{ij} = \kappa \cdot \sigma_{ij} \cdot x_{ij}$ is the adaptive perturbation bound. The detailed iterative procedure is provided in Appendix A.



(a) Perturbation visualization (b) Protection Comparison

Figure 4: (a) GeoErase produces more localized and structured perturbations while preserving image appearance. (b) GeoErase achieves lower geolocation accuracy across distance thresholds on multiple VLMs, indicating stronger privacy protection.

4 Experiments

4.1 Experimental Setup

We evaluate GeoErase on two representative benchmarks, Google Street View (GSV) and Im2GPS3k, which cover street-level and global-scale geolocation tasks. The evaluation uses four state-of-the-art VLMs as black-box targets: GPT-5.4, Claude-4.5, Gemini-3, and Qwen3-vl. A detailed description of dataset partitions, model configurations, and hyperparameter settings is provided in Appendix B.1.

4.2 Evaluation Metrics

We evaluate protection effectiveness using the Haversine distance and geolocation accuracy at five thresholds: 1, 25, 200, 750, and 2,500 km, covering street-level to continent-level localization errors. In ablation studies, we report the average geolocation error (Avg. Dis.), the mean distance between predicted and ground-truth coordinates, to measure protection effectiveness. Visual quality is measured by PSNR, SSIM, and LPIPS to quantify pixel-level distortion, structural fidelity, and perceptual similarity, while semantic preservation is assessed by comparing location-agnostic descriptions of clean and protected images using BLEU, ROUGE, and BERTScore. Detailed definitions are provided in Appendix B.2.

5 Results

5.1 Effectiveness

As shown in Table 2, we compare GeoErase with representative adversarial protection methods, including AnyAttack (Zhang et al., 2025b), SSA-CWA (Dong et al., 2023), M-Attack (Li et al., 2026), ReasonBreak (Zhang et al., 2025a), and GeoShield (Liu et al., 2026). GeoErase consistently achieves the lowest geolocation accuracy across most models and datasets. On Google Street View, it reduces the street-level accuracy of GPT-5.4 from 10.2% to 2.1%, outperforming both ReasonBreak and GeoShield. Similar gains are observed on Im2GPS3k, indicating that the proposed attribution-guided and adaptive sparse optimization transfers well across datasets and black-box VLMs.Additional comparisons with the prior method GeoShield on earlier-generation large models are provided in Appendix C.1.

5.2 Imperceptibility

Imperceptibility is essential for practical geolocation privacy protection, since visible artifacts may reveal defensive perturbations. As shown in Table 3, GeoErase achieves better SSIM and LPIPS than strong geoprivacy baselines, despite a slightly lower PSNR. Since PSNR is sensitive to localized pixel deviations, it may not fully reflect perceptual quality, whereas SSIM and LPIPS better capture structural consistency and perceptual similarity. The visual comparison in Appendix C.2 further shows that GeoErase produces more natural-looking protected images. Fig. 5 explains this advantage through perturbation distribution: GeoErase has lower mean and P50 magnitudes but a higher P90 value than GeoShield and ReasonBreak, indicating that it reduces perturbations for most pixels while concentrating stronger ones in a small subset of high-texture regions.

Dataset	Model	GPT-5.4 (Distance in km)					Claude-4.5 (Distance in km)					Gemini-3 (Distance in km)					Qwen3-vl (Distance in km)				
		1	**25**	**200**	**750**	**2500**	**1**	**25**	**200**	**750**	**2500**	**1**	**25**	**200**	**750**	**2500**	**1**	**25**	**200**	**750**	**2500**
Google Street View	Clean	10.2	22.4	50.1	75.6	94.8	5.8	11.6	32.5	59.7	78.1	12.1	24.2	54.6	81.8	96.9	7.9	18.4	37.0	63.7	89.2
	AnyAttack	9.1	17.6	39.3	63.8	83.9	5.1	10.2	24.0	47.9	64.2	12.0	23.5	54 .3	78.1	93.3	4.5	16.1	36.1	54.7	70.9
	SSA-CWA	8.9	16.8	37.9	62.1	79.3	5.1	9.2	23.0	45.8	60.2	11.9	22.5	53.8	77.9	91.6	4.3	14.9	32.6	53.9	69.4
	M-Attack	8.6	14.6	32.7	56.8	72.0	4.9	7.1	21.4	42.2	58.2	11.8	22.1	53.7	77.8	90.5	0.9	5.8	24.9	53.2	68.3
	ReasonBreak	6.1	11.2	31.4	58.4	78.0	**2.6**	5.2	20.5	37.9	53.8	12.8	21.8	36.7	76.7	88.5	0.7	6.3	22.1	50.5	74.2
	GeoShield	4.6	**8.1**	21.3	39.9	54.5	4.6	6.3	17.5	33.2	45.4	5.5	12.5	30.3	58.8	80.2	1.4	4.2	18.6	34.5	51.5
	GeoErase	**2.1**	8.5	**16.4**	**25.5**	**39.0**	3.1	**4.5**	**11.5**	**19.4**	**34.7**	**5.0**	**15.7**	**20.9**	**35.7**	**59.2**	**0.2**	**1.0**	**12.1**	**24.8**	**38.7**
Im2GPS3k	Clean	17.6	48.2	58.9	73.6	88.1	10.2	34.1	48.9	67.4	78.3	18.7	44.5	66.2	84.2	88.3	8.4	20.7	43.5	69.9	78.4
	AnyAttack	11.9	29.4	42.8	64.2	76.1	9.2	21.2	37.6	50.8	67.1	16.2	41.0	64.7	82.1	87.4	5.0	12.9	38.1	62.9	88.1
	SSA-CWA	10.8	27.9	38.9	62.0	71.6	9.1	20.5	34.7	51.7	65.8	15.9	40.1	62.5	80.9	86.8	4.8	12.3	37.9	59.2	83.5
	M-Attack	9.1	23.2	30.3	51.5	64.6	8.8	19.3	30.2	51.0	63.5	10.2	36.0	52.1	65.6	80.2	4.5	11.1	36.1	50.2	65.4
	ReasonBreak	9.2	22.2	34.3	54.6	67.7	6.4	14.6	27.1	45.8	60.4	13.2	39.4	50.5	69.7	80.8	1.7	**7.8**	24.8	51.2	71.9
	GeoShield	6.4	12.9	20.9	33.7	39.3	7.0	16.7	25.6	33.9	46.9	6.2	20.3	37.9	47.1	58.0	3.9	10.8	19.3	24.8	34.4
	GeoErase	**5.6**	**10.2**	**14.8**	**22.4**	**34.6**	**5.2**	**10.7**	**14.5**	**19.8**	**33.2**	**1.1**	**10.7**	**19.6**	**24.7**	**34.5**	**0.9**	10.4	**13.3**	**21.2**	**28.8**

Table 2: Geolocation accuracy (%) of different methods under varying distance thresholds (1, 25, 200, 750, and 2500 km) on Google Street View and Im2GPS3k datasets across multiple vision-language models. Lower accuracy indicates stronger privacy protection.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
AnyAttack	28.38	0.6318	0.2809
SSA-CWA	25.95	0.6744	0.3568
M-Attack	18.62	0.4361	0.4005
ReasonBreak	21.82	0.5733	0.3785
GeoShield	18.63	0.4385	0.4056
GeoErase	18.57	0.4527	0.3790

Table 3: Imperceptibility comparison of different protection methods. Higher PSNR and SSIM indicate better visual fidelity, while lower LPIPS indicates better perceptual similarity.

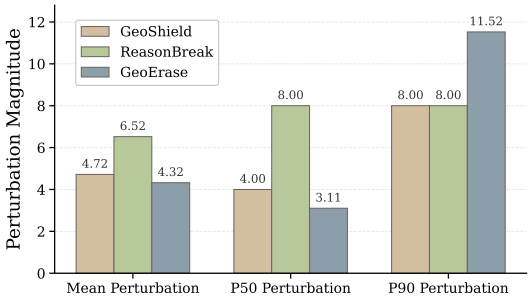


Figure 5: Perturbation magnitude statistics. Mean is the average perturbation magnitude, while P50 and P90 denote the median and 90th-percentile perturbation magnitudes.

Constraint	Method	LLaVA-1.5			GPT-5.4		
		BLEU	ROUGE	BERT-S	BLEU	ROUGE	BERT-S
$\epsilon : 8/255$	AnyAttack	0.48	0.59	0.94	0.14	0.31	0.90
	SSA-CWA	0.45	0.57	0.93	0.13	0.32	0.91
	M-Attack	0.22	0.39	0.91	0.08	0.26	0.90
	ReasonBreak	0.27	0.45	0.92	0.12	0.29	0.91
	GeoShield	0.25	0.41	0.92	0.11	0.28	0.89
$\epsilon : 16/255$	AnyAttack	0.43	0.55	0.94	0.13	0.30	0.91
	SSA-CWA	0.32	0.47	0.92	0.11	0.28	0.90
	M-Attack	0.15	0.34	0.89	0.08	0.25	0.89
	ReasonBreak	0.24	0.40	0.91	0.10	0.27	0.90
	GeoShield	0.20	0.37	0.91	0.09	0.27	0.89
Adaptive ϵ	**GeoErase**	0.22	0.38	0.91	0.10	0.27	0.90

Table 4: Semantic similarity between original and protected images under different perturbation constraints.

Method	Effectiveness (Avg. Dis) ↑		Imperceptibility		
	GPT-5.4	Claude-4.5	PSNR ↑	SSIM ↑	LPIPS ↓
w/o FAGD	4764.24	5936.01	18.56	0.4526	0.3712
w/o PASSP	2398.86	4135.36	18.62	0.4295	0.4084
Full Model	**7608.11**	**7483.92**	**18.57**	**0.4527**	**0.3790**

Table 5: Ablation study. Removing either FAGD or PASSP consistently degrades performance across different VLMs, confirming the necessity of each module.

5.3 Semantic Preservation

As shown in Table 4, GeoErase achieves slightly lower semantic similarity than high-fidelity but weaker baselines such as AnyAttack and SSA-CWA. Compared with stronger geoprivacy methods such as GeoShield, GeoErase maintains comparable semantic similarity while achieving stronger geolocation obfuscation. This indicates that GeoErase suppresses location-relevant cues without substantially altering the core semantic content of the image.

5.4 Ablation Study

Table 5 shows that both FAGD and PASSP are necessary for GeoErase. Removing FAGD substantially reduces protect effectiveness, confirming the importance of explicit geographic feature modeling. Removing PASSP degrades perceptual quality, mainly reflected by lower SSIM and higher LPIPS. The full model achieves the best trade-off between protection effectiveness and visual fidelity.

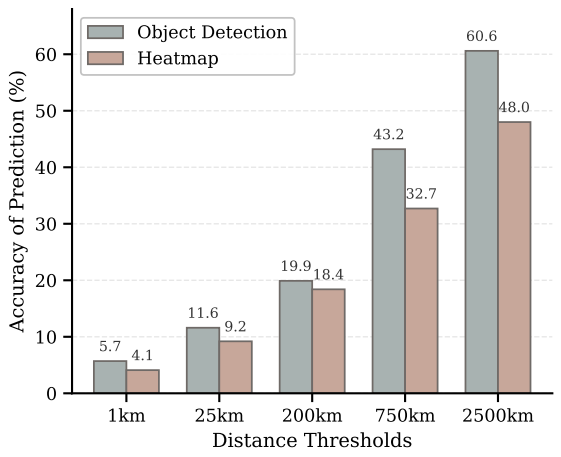


Figure 6: Comparison of protection effectiveness under different geographic feature extraction strategies.

Effectiveness of Feature Attribution-based Extraction. We compare feature heatmap-based extraction with object detection-based extraction under the same optimization setting. As shown in Fig. 6, heatmap-based extraction yields lower geolocation accuracy across most distance thresholds, indicating stronger protection against VLM-based geolocation inference. The difference mainly stems from their localization granularity. Object detection focuses on a few explicit objects and may miss broader geographic regions, whereas feature heatmaps capture both object-level and scene-level cues to localize geolocation-sensitive regions more accurately.

Analysis of Adaptive Bound and Sparsity. As shown in Fig. 4(a), the full model produces structured perturbations mainly concentrated in high-texture regions and suppressed in smooth areas, confirming the adaptive perturbation bound.

We further analyze the effect of the sparsity ratio k_{ratio}. As shown in Fig. 7, increasing k_{ratio} improves protection effectiveness but degrades visual quality, indicating a trade-off between privacy protection and imperceptibility. Smaller values restrict perturbations to fewer regions, whereas larger values allow broader coverage and stronger disruption of geographic cues. Additional PSNR and SSIM results under different k_{ratio} values are provided in Appendix C.3. Overall, an intermediate k_{ratio} better balances effectiveness and visual quality.

5.5 Robustness to Image Transformations

We evaluate robustness under JPEG compression and Gaussian blur, two common online image transformations. As shown in Fig. 8, GeoErase consistently causes larger geolocation errors than clean images and ReasonBreak, outperforms GeoShield under most JPEG settings, and remains competitive under Gaussian blur. This indicates that GeoErase retains protection after common degradations rather than relying on fragile high-frequency noise.

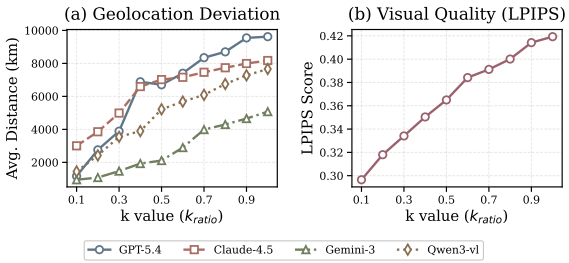


Figure 7: Trade-off Analysis between Geolocation Privacy and Perceptual Quality.

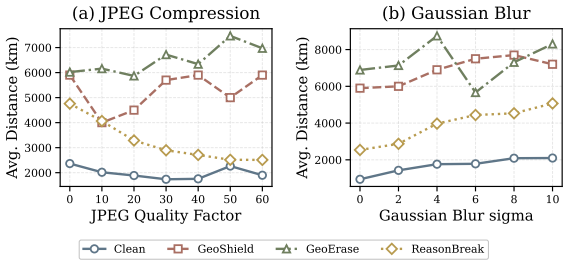


Figure 8: Robustness Analysis against Common Image Transformations.

6 Conclusion

We presented GeoErase, a black-box geolocation privacy protection framework that erases geolocation-revealing cues from model perception via optimized perturbations. GeoErase uses feature heatmaps to localize cues important for geolocation inference, thereby addressing the limitation that object detection may yield imprecise features unsuitable as reliable anchors for feature repulsion during perturbation optimization. It further applies a perception-adaptive bound and structured sparsity constraint to reduce visible artifacts and concentrate perturbations in selected regions. Extensive experiments show that GeoErase effectively reduces VLM localization accuracy while maintaining semantic consistency and perceptual quality, making it practical for privacy-preserving image sharing in real-world scenarios.

7 Limitations

GeoErase has one main limitation. Its current objective focuses on maximizing the geographic distance between the predicted and ground-truth coordinates. While this reduces geolocation accuracy, it does not explicitly optimize for model abstention or calibrated uncertainty. During our experiments, we observed that VLMs sometimes refused to provide precise coordinates or stated that the location could not be reliably determined from the protected image. This suggests that inducing uncertainty may be a promising direction for geolocation privacy protection. In real-world scenarios, such abstention may be preferable to forcing an incorrect coordinate prediction, since an alternative prediction may still reveal coarse geographic information. Future work could therefore incorporate uncertainty-aware or abstention-oriented objectives to encourage refusal, calibrated uncertainty, or coarse-grained responses, rather than relying solely on coordinate displacement.

References

Francesco Croce and Matthias Hein. 2019. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4724–4732.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. 2023. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*.

Yanbo Fan, Baoyuan Wu, Tuanhui Li, Yong Zhang, Mingyang Li, Zhifeng Li, and Yujiu Yang. 2020. Sparse adversarial attack via perturbation factorization. In *European conference on computer vision*, pages 35–50. Springer.

Oliver Grainge, Sania Waheed, Jack Stilgoe, Michael Milford, and Shoaib Ehsan. 2025. Assessing the geolocation capabilities, limitations and societal risks of generative vision-language models. In *Proceedings of the AAAI Symposium Series*, volume 7, pages 161–168.

Lukas Haas, Silas Alberti, and Michal Skreta. 2023. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275*.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. 2024. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902.

James Hays and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. 2025. Vlms as geoguessr masters: Exceptional performance, hidden biases, and privacy risks. *arXiv e-prints*, pages arXiv–2502.

Neel Jay, Hieu Minh Nguyen, Trung Dung Hoang, and Jacob Haimes. 2025. Evaluating precise geolocation inference capabilities of vision language models. *arXiv preprint arXiv:2502.14412*.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. 2026. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *Advances in Neural Information Processing Systems*, 38:45018–45051.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Simeng Qin, and Xiaochun Cao. 2026. Geoshield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35653–35661.

Weidi Luo, Tianyu Lu, Qiming Zhang, Xiaogeng Liu, Bin Hu, Yue Zhao, Jieyu Zhao, Song Gao, Patrick McDaniel, Zhen Xiang, and 1 others. 2025. Doxing via the lens: Revealing location-related privacy leakage on multi-modal large reasoning models. *arXiv preprint arXiv:2504.19373*.

Di Ming, Peng Ren, Yunlong Wang, and Xin Feng. 2024. Transferable structural sparse adversarial attack via exact group sparsity training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24696–24705.

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2019. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9096.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701.

Xiaonan Wang, Bo Shao, and Hansaem Kim. 2025. Ai knows where you are: Exposure, bias, and inference in multimodal geolocation with koreageo. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9897–9914.

Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. 2024. Llmgeo: Benchmarking large language models on image geolocation in-the-wild. *arXiv preprint arXiv:2405.20363*.

Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, pages 37–55. Springer.

Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. 2018. Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*.

Jiaming Zhang, Che Wang, Yang Cao, Longtao Huang, and Wei Yang Bryan Lim. 2025a. Disrupting hierarchical reasoning: Adversarial protection for geographic privacy in multimodal reasoning models. *arXiv preprint arXiv:2512.08503*.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. 2025b. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138.

Appendix Ovewview

This appendix supplements the main paper with implementation details,evaluation protocols, and additional analyses. We provide the full GeoErase optimization procedure, describe the datasets and model configurations, define the evaluation metrics, and report extended results on effectiveness, imperceptibility, sparsity-ratio sensitivity, and the perception-adaptive perturbation bound. The appendix is organized as follows:

- §A Algorithmic Details of GeoErase
- §B.1 Experimental Setup
- §B.2 Evaluation Metrics
- §C.1 Effectiveness
- §C.2 Qualitative Case Study
- §C.3 Effect of Sparsity Ratio on Visual Quality
- §D Rationale for the Perception-Adaptive Bound
- §E Threat Model and Deployment Assumptions

A Algorithmic Details of GeoErase

We summarize the iterative optimization procedure of GeoErase in Algorithm 1.The optimization procedure of GeoErase first localizes geo-sensitive regions through heatmap-based feature extraction and computes a perception-adaptive perturbation bound from local image statistics. At each iteration, global and local crops of the perturbed image are encoded to optimize a multi-scale loss that suppresses geo-related features while preserving non-geographical semantics. The perturbation is then updated, clipped by the adaptive bound, and projected under the group sparsity constraint to obtain a visually subtle protected image.

B Implementation Details and Experimental Setup

B.1 Experimental Setup

Dataset Selection. We conduct experiments on two representative geolocation benchmarks, Google Street View (GSV) and Im2GPS3k. To ensure a fair and computationally feasible evaluation across multiple black-box VLMs, we randomly select 1,000 images from each dataset for testing. The selected images cover diverse geographic scenes and are used consistently across all compared methods and target models. All images are resized to the same input resolution before perturbation generation and evaluation.

Parameter	Value
attack steps T	100
step size α	1.0
input resolution	640
crop resolution	224
crop scale	$[0.5, 0.9]$
κ	1.4
k_{ratio}	0.6
group stride	8

Table 6: Main hyperparameter settings used in our experiments.

Hyperparameter Settings. Table 6 summarizes the main hyperparameter settings used in our experiments.

Surrogate and Target Model Configurations To facilitate the generation of transferable perturbations in a black-box setting, we employ an ensemble of three CLIP variants as surrogate models: ViT-B/16, ViT-B/32, and ViT-g-14-laion2B-s12B-b42K. The diversity in the architectures of these surrogate models ensures that the generated adversarial features are not overfitted to a specific network structure. For the evaluation of protect effectiveness, we target four state-of-the-art VLMs: GPT-5.4, Claude-4.5, Gemini-3 , and Qwen3-vl. These target models are accessed via official APIs, and no internal gradients or architectural parameters are utilized during the optimization process.

Geolocation Query Prompt To evaluate the geolocation inference ability of VLMs, we query each model with the following prompt and require it to directly output geographic coordinates:

Geolocation Query Prompt

Act as a professional geolocation expert (OSINT). Examine the uploaded photo and estimate its GPS coordinates, including latitude and longitude.
You must output only two floating-point numbers separated by a comma. Do not include any text, labels, or explanations.
If the location is ambiguous, output the coordinates of the most probable candidate.
Example format: `48.8584, 2.2945`

Location-Agnostic Description Prompt To obtain $T_{\text{non-geo}}$, we query an auxiliary VLM with a location-agnostic prompt. The prompt is designed

to preserve general visual semantics while suppressing explicit geographic information, such as place names, coordinates, landmarks, signs, and region-specific cues. The generated description is directly used as $T_{\text{non-geo}}$ as a semantic reference for feature disentanglement. The prompt used in our implementation is shown below:

Location-Agnostic Description Prompt

You are an advanced image analysis AI. Describe the uploaded image using only general visual content.
Focus on common objects, scene type, colors, lighting, and general actions.
Do not mention any geographic location or location-revealing information, including country names, city names, region names, street names, GPS coordinates, landmarks, language on signs, license plates, flags, phone codes, or region-specific architectural and cultural cues.
Avoid any description that can directly identify where the image was taken.
Keep the description concise and location-agnostic.
Example output: `An urban street with buildings, vehicles, trees, and pedestrians.`

The prompt constrains the auxiliary VLM to focus on generic visual content, including objects, scene categories, colors, lighting conditions, and general activities, while excluding explicit location-revealing identifiers. In this way, $T_{\text{non-geo}}$ provides a location-agnostic semantic reference that supports feature disentanglement and semantic preservation without directly encoding identifiable geographic information.

B.2 Evaluation Metrics

Geographic Localization Accuracy The Haversine distance D is utilized to calculate the great-circle distance between the predicted and ground-truth coordinates. To ensure legibility, the formula is expressed using an intermediate variable h:

$$D = 2R \arcsin(\sqrt{h}) \tag{13}$$

$$h = \sin^2 \frac{\phi_2 - \phi_1}{2} + \cos\phi_1 \cos\phi_2 \sin^2 \frac{\lambda_2 - \lambda_1}{2} \tag{14}$$

where R denotes the radius of the Earth. Localization accuracy is reported across five geographic thresholds:

- **Street level (1 km):** Precise identification of specific locations or intersections.
- **City level (25 km):** Successful recognition of urban and metropolitan areas.
- **Region level (200 km):** Identification of broad administrative divisions or provinces.
- **Country level (750 km):** Localization within correct national boundaries.
- **Continent level (2,500 km):** Differentiation between global landmasses.

Imperceptibility Metrics. We utilize PSNR, SSIM, and LPIPS to evaluate image fidelity across pixel-wise, structural, and perceptual dimensions. PSNR measures pixel-level reconstruction accuracy, while SSIM assesses the preservation of structural integrity, luminance, and contrast. To better reflect human visual judgment, LPIPS leverages deep features to quantify the perceptual distance between the original and perturbed images. Together, these metrics provide a comprehensive assessment of the visual naturalness of the generated adversarial examples.

Semantic Preservation Metrics. To assess semantic preservation under geolocation privacy protection, we use LLaVA-1.5 and GPT-5.4 to generate location-agnostic descriptions for both clean and protected images, and compare them using BLEU, ROUGE, and BERTScore. BLEU and ROUGE measure lexical overlap from precision- and recall-oriented perspectives, respectively, while BERTScore captures semantic similarity with contextual embeddings. Higher scores indicate better preservation of the original visual content after protection.

C Additional Experimental Results

C.1 Effectiveness

As shown in the table 7, GeoErase consistently achieves the lowest geolocation accuracy across all target VLMs and distance thresholds, outperforming existing state-of-the-art methods like GeoShield. Specifically, it significantly reduces the localization success rate at the most challenging 1 km scale, demonstrating its superior ability to

Figure 9: Comparison of Geolocation Privacy Protection Effectiveness Across Different Methods

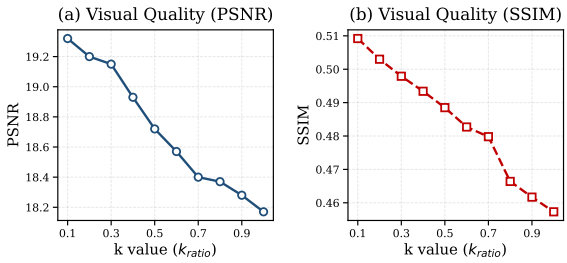


Figure 10: Effect of Sparsity Ratio on PSNR and SSIM

suppress critical geographic features while maintaining robust protection across diverse datasets.

C.2 Qualitative Case Study

Fig. 9 shows the performance of various geolocation privacy protection methods in misleading location estimation. Among the evaluated techniques, GeoErase achieves the highest protection efficacy by inducing the largest estimation deviation of 7,308 km, effectively shifting the perceived location from Ukraine to Singapore while maintaining superior visual naturalness.

C.3 Effect of Sparsity Ratio on Visual Quality

As shown in Fig. 10, visual quality consistently decreases as k_{ratio} increases. Both PSNR and SSIM exhibit a downward trend, indicating that allowing perturbations to cover more spatial groups introduces stronger image distortion.

D Rationale for the Perception-Adaptive Bound

We define the pixel-wise perturbation bound as $\epsilon_{ij} = \kappa \cdot \sigma_{ij} \cdot x_{ij}$, where σ_{ij} denotes local texture complexity and x_{ij} denotes pixel intensity. This design is based on two considerations.

First, textured regions usually tolerate larger perturbations. Due to contrast masking and spatial masking in human vision, local distortions are less visible in regions with complex textures, strong edges, or dense structures than in smooth regions. Therefore, σ_{ij} assigns larger perturbation budgets to textured regions and smaller budgets to smooth regions.

Second, perturbations in low-intensity regions can cause stronger relative visual changes. According to Weber's law and the Weber–Fechner law, the visibility of an intensity change is related to its ratio to the background luminance,

$$\frac{\Delta I}{I} = k, \tag{15}$$

where I is the background luminance, ΔI is the just-noticeable luminance difference, and k is the

Algorithm 1: GeoErase

Input: Image x; encoders $\{f_{\theta_i}, g_{\theta_i}\}_{i=1}^{N}$; steps T; step size η; adaptive coefficient κ; weights β, γ; sparsity ratio k_{ratio}; group stride s.

Output: Protected image x'.

Initialize perturbation $\delta \leftarrow 0$;

Obtain location-agnostic description $T_{\text{non-geo}}$ for x;

for $i = 1$ **to** N **do**

  Compute non-geographic feature $z^i_{\text{non-geo}} \leftarrow g_{\theta_i}(T_{\text{non-geo}})$;

  Estimate geographic feature $z^i_{\text{geo}} \leftarrow f_{\theta_i}(x) - z^i_{\text{non-geo}}$;

Generate attribution mask $M_{\text{attr}} \leftarrow \texttt{Binarize}(\texttt{Refine}(\texttt{GeoHeatmap}(x)))$;

for $i = 1$ **to** N **do**

  Extract attribution-guided geographic feature $z^i_{\text{feature}} \leftarrow f_{\theta_i}(x \odot M_{\text{attr}})$;

Compute texture complexity $\sigma_{ij} \leftarrow \sqrt{\min(\text{std}^{ij}_x, \text{std}^{ij}_y)}$;

Set adaptive perturbation bound $\epsilon_{ij} \leftarrow \kappa \cdot \sigma_{ij} \cdot x_{ij}$;

for $t = 1$ **to** T **do**

  $x' \leftarrow x + \delta$;

  Sample global and local crops $x_g \leftarrow \texttt{RandomResizedCrop}(x')$, $x_l \leftarrow \texttt{RandomCrop}(x')$;

  for $i = 1$ **to** N **do**

    Extract multi-scale features $f^i_g \leftarrow f_{\theta_i}(x_g)$, $f^i_l \leftarrow f_{\theta_i}(x_l)$;

  Compute loss:

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{N} \Big[& S(f^i_g, z^i_{\text{geo}}) + S(f^i_l, z^i_{\text{geo}}) \\ & + \gamma\big(S(f^i_g, z^i_{\text{feature}}) + S(f^i_l, z^i_{\text{feature}})\big) \\ & - \beta\big(S(f^i_g, z^i_{\text{non-geo}}) + S(f^i_l, z^i_{\text{non-geo}})\big)\Big].\end{aligned}$$

  Compute gradient $g \leftarrow \nabla_\delta \mathcal{L}$;

  Update perturbation $\delta \leftarrow \delta + \eta \cdot \text{sign}(g)$;

  Project perturbation: $\delta_{ij} \leftarrow \text{Clip}(\delta_{ij}, -\epsilon_{ij}, \epsilon_{ij}),\ \forall (i, j)$;

  Generate structured sparsity mask $M_{\text{sp}} \leftarrow \texttt{GroupTopK}(\delta, s, k_{\text{ratio}})$;

  Apply sparse perturbation $\delta \leftarrow \delta \odot M_{\text{sp}}$;

return $x' \leftarrow x + \delta$;

Weber fraction. For example, changing a pixel from 10 to 20 and from 200 to 210 produces the same absolute change, but the relative changes are 100% and 5%, respectively. Thus, the former is more visually salient. Multiplying the bound by x_{ij} suppresses perturbations in low-intensity regions and avoids disproportionately strong relative changes.

In summary, σ_{ij} adapts the budget to local texture complexity, while x_{ij} reduces salient changes in dark regions. Together, they improve visual stealthiness while preserving sufficient perturbation capacity for geolocation privacy protection.

E Threat Model and Deployment Assumptions

GeoErase is designed for a pre-sharing protection scenario, where a user applies a perturbation to an image before uploading or sharing it online. We assume that the potential adversary can query powerful black-box VLMs to infer the geographic location of the shared image, but has no access to the internal parameters, training data, gradients, or decoding processes of these models. This setting matches practical commercial VLM services, where users and attackers interact with the models only through public APIs.

The goal of GeoErase is to reduce the ability of such models to infer accurate geographic coordi-

Dataset	Model	GPT-4o (Distance in km)					GPT-4.1 (Distance in km)					Claude-3.5 (Distance in km)					Gemini-2.5 (Distance in km)				
		1	**25**	**200**	**750**	**2500**	**1**	**25**	**200**	**750**	**2500**	**1**	**25**	**200**	**750**	**2500**	**1**	**25**	**200**	**750**	**2500**
Google Street View	Clean	7.3	17.7	41.6	73.4	90.6	9.1	23.2	48.5	78.0	92.8	4.9	9.0	27.5	57.8	78.1	8.7	21.9	47.6	79.2	93.5
	AnyAttack	6.2	15.9	37.7	68.7	87.3	7.9	20.4	44.4	73.1	89.5	2.8	8.4	24.1	51.4	74.2	8.0	19.8	44.2	77.3	91.2
	SSA-CWA	4.7	12.1	31.8	62.9	83.4	7.2	18.4	40.8	70.0	88.3	1.4	6.1	18.3	44.9	62.1	7.2	18.8	40.6	74.5	90.5
	M-Attack	3.3	9.1	24.5	48.7	71.1	4.9	12.6	30.3	54.9	76.1	1.4	5.3	15.3	35.8	56.3	6.1	16.7	37.6	66.8	86.3
	GeoShield	1.1	2.9	7.6	17.5	33.8	1.4	3.6	9.1	20.9	37.9	0.1	1.1	5.2	12.5	27.4	4.7	13.0	32.4	59.1	80.9
	GeoErase	**0.5**	**1.0**	**2.8**	**8.0**	**24.5**	**0.0**	**1.3**	**7.5**	**12.8**	**29.8**	**0.2**	**0.5**	**2.8**	**8.0**	**20.5**	**4.1**	**7.2**	**14.7**	**38.9**	**57.1**
Im2GPS3k	Clean	14.4	38.9	55.8	71.4	84.6	18.2	46.3	59.4	73.9	87.4	9.1	30.0	43.4	61.9	77.1	18.2	45.5	59.7	74.9	86.4
	AnyAttack	14.0	36.9	52.1	67.1	81.4	17.4	43.3	56.6	71.5	84.3	8.2	26.5	40.0	57.8	74.3	17.8	40.8	56.5	69.4	81.6
	SSA-CWA	13.5	33.1	47.2	64.7	76.8	15.9	39.8	53.0	67.2	81.3	7.4	24.3	35.1	52.0	69.3	16.1	39.6	53.6	68.5	78.4
	M-Attack	9.2	23.0	32.9	46.2	61.1	13.2	30.1	40.0	50.5	65.3	5.6	16.9	24.1	36.4	52.8	15.2	36.6	49.3	61.6	77.3
	GeoShield	4.0	9.7	12.3	20.0	35.8	9.2	13.2	16.2	23.2	38.3	2.4	6.7	8.6	14.6	30.1	10.4	25.4	31.8	41.7	55.3
	GeoErase	**3.0**	**8.4**	**10.1**	**17.2**	**24.3**	**5.2**	**10.7**	**14.5**	**19.8**	**31.2**	**1.1**	**4.4**	**5.4**	**11.8**	**21.3**	**9.1**	**20.2**	**26.3**	**35.8**	**44.4**

Table 7: This table presents the comparison results with the VLMs reported in the original GeoShield paper, including GPT-4o, GPT-4.1, Claude-3.5, and Gemini-2.5. GeoErase consistently achieves lower geolocation accuracy across different models, datasets, and distance thresholds, indicating stronger privacy protection.

nates from visual content while preserving the general semantic utility of the image. GeoErase does not aim to remove metadata such as EXIF GPS tags, prevent manual recognition by human experts, or provide formal privacy guarantees against all possible geolocation systems. Therefore, metadata removal and standard privacy practices should be used together with GeoErase in real-world deployment.