NOWHERE: EVIDENCE-SPACE PROTECTION FOR GEOLOCATION PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large reasoning models can infer the locations of ordinary images from fine-grained visual clues, posing serious geoprivacy risks. Existing methods disrupt overall image features or use coarse text descriptions to define attack targets. However, they may overlook key location clues in individual scenes. Moreover, static attacks may keep targeting already weakened clues while leaving evidence that still supports localization. Therefore, effective protection requires targeting specific visual evidence and adjusting the optimization focus as this evidence changes. Although models use different internal representations, their location predictions rely on visual evidence from the same physical scene. This evidence provides a shared target for cross-model attacks. Building on this insight, we propose NOWHERE, a black-box geoprivacy protection framework based on Evidence-Space Transfer. Evidence Space Construction (ESC) locates image-specific geographic clues and builds multidimensional evidence subspaces in each surrogate model's feature space. It also uses a complementary scene reference to guide content preservation. Dynamic Evidence Tracking (DET) compares clue features before and after perturbation, shifting optimization toward clues that remain insufficiently suppressed. Extensive experiments show that NOWHERE provides stronger geoprivacy protection while preserving visual quality and non-geographic semantic consistency. On Im2GPS3k, NOWHERE offers stronger protection against GPT-5.6-Terra's location inference. The model locates only 4.2% of protected images within 1 km, compared with 6.8% under the strongest baseline. This helps users reduce location leakage while keeping their photos suitable for sharing.

1 INTRODUCTION

A photo can reveal where it was taken even without geotags. Its details such as text on signs, building exteriors, and roadside vegetation may appear insignificant, yet vision-language models (VLMs) can easily combine them to infer the actual location (Li et al., 2025; Han et al., 2025; Dou et al., 2024; Jin et al., 2026; Dou et al., 2026). Users sharing travel experiences or daily life may unintentionally reveal places they visit or even their homes. Limiting location inference while preserving image appearance and content has therefore become an important privacy challenge in image sharing.

It is generally impossible for the users to anticipate which models will later geolocate their shared images. In this black-box scenario, traditional transfer-based attacks tend to generate transferable image perturbations on accessible surrogate models, aiming to disrupt unseen models (Zhao et al., 2023; Zhang et al., 2025; Li et al., 2026; Jia et al., 2026). However, indiscriminate attacks do not explicitly distinguish geographic clues from scene content that should be preserved, which may also disrupt the original visual representation or aligned features of the image, and therefore affect non-geographic information or induce outputs unsupported by the original image. Even without considering this, these perturbations may still be insufficient to disrupt location inferences and result in weak geoprivacy protection (Liu et al., 2026; Zhang et al., 2026).

In general, the limitations of recent geoprivacy defense methods can be summarized in two categories. First, they do not effectively represent the geographic evidence, but approximate geographic information with coarse text descriptions or predefined concepts instead. These representations cannot fully capture scene-specific materials, textures, and structures. Some methods also aggre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Motivation of our proposed NOWHERE.** Existing coarse text descriptions and feature averaging can obscure distinct geographic clues, while static attacks may keep targeting already-weakened clues (left). NOWHERE, on the other hand, represents visual evidence in a multidimensional subspace and dynamically prioritizes remaining clues (right).

gate regional visual features into a single representation, obscuring differences between geographic clues. Second, they solely focus on clues identified in the original image, and lack an adjustment mechanism for changed clues during optimization, especially for clues that do not weaken at the same rate. This static approach may keep directing updates toward already weakened regions while leaving residual evidence insufficiently suppressed. Effective geoprivacy protection thus requires diverse evidence representations and dynamic attacks guided by residual evidence.

To address these issues, we propose **NOWHERE**, a black-box geoprivacy protection framework based on **Evidence-Space Transfer**. We find that although models may have different internal representations, they nevertheless rely on visual evidence from the same physical scene for localization. NOWHERE treats this shared evidence as a core attack target, and aims to suppress the same set of geographic clues for each surrogate model in its own feature space. Our framework contains two modules. Firstly, **Evidence Space Construction (ESC)** localizes image-specific geographic clues and organizes their visual features into multidimensional evidence subspaces, which retain distinct evidence directions. **ESC** also builds a complementary reference of the remaining scene to guide content preservation. Secondly, **Dynamic Evidence Tracking (DET)** repeatedly compares clue features before and after perturbation. As some clues weaken, our optimization shifts toward the clues with remaining significant responses, expanding the scope of evidence suppression. Experiments on Google Street View (Jay et al., 2025) and Im2GPS3k (Hays & Efros, 2008) show that NOWHERE provides stronger geoprivacy protection than existing state-of-the-art (SOTA) adversarial defenses. It preserves visual quality and maintains competitive non-geographic semantic consistency. On Im2GPS3k, NOWHERE significantly reduces GPT-5.6-Terra's localization accuracy within 1 km from 6.8% with the strongest baseline to 4.2%.

Our main contributions are as follows:

- We introduce **Evidence-Space Transfer**, which guides multi-surrogate optimization using shared geographic evidence to improve transferability to unseen models.
- We propose **NOWHERE**, combining multidimensional evidence subspaces with dynamic evidence tracking to suppress location evidence while preserving scene content.
- We demonstrate SOTA black-box geoprivacy protection performance on 2 benchmarks and 4 VLMs, supported by ablations and visual and semantic evaluations.

2 Related Work

2.1 Image Geolocation

Image geolocation predicts where an image is captured (Zhang et al., 2021). Classification methods assign images to geographic cells (Weyand et al., 2016; Muller-Budack et al., 2018; Seo et al., 2018; Clark et al., 2023), while retrieval methods match image features against reference images or location embeddings (Shi et al., 2019; 2020; Vivanco Cepeda et al., 2023; Choi et al., 2023). Classification accuracy depends partly on how geographic space is divided. Retrieval performance

relies on reference coverage and matching quality. Recent work uses VLMs to reason about visual clues and geographic knowledge (Li et al., 2025; Dou et al., 2024). Other approaches introduce tool-assisted localization and multi-agent collaboration (Jin et al., 2026; Han et al., 2025), or use multimodal features to guide coordinate generation (Dou et al., 2026). These advances expand geolocation beyond classification and visual matching. Models can combine scattered clues to infer locations from ordinary images. On DoxBench, multimodal reasoning models outperform non-expert humans in location inference (Luo et al., 2025). This capability creates privacy risks for everyday image sharing (An et al., 2026; Shao & Li, 2025) and motivates protection that suppresses geographic evidence while preserving useful scene content.

2.2 Adversarial Perturbation

Transfer-based attacks target black-box VLMs through accessible surrogate models (Zhao et al., 2023; Xu et al., 2025; Cai et al., 2026a; He et al., 2025; Jiang et al., 2025). AnyAttack (Zhang et al., 2025) and AdvDiffVLM (Guo et al., 2024) generate perturbations through self-supervised learning and diffusion, respectively. Optimization-based methods use crop-based alignment (Li et al., 2026), global–local alignment with adaptive ensemble weighting (Jia et al., 2026), and surrogate selection (Huang et al., 2025). Recent work explores cross-prompt transfer (Luo et al., 2024) and cross-model jailbreak limitations (Schaeffer et al., 2025). UltraBreak (Cui et al., 2026) combines visual regularization with semantic supervision, while SAGA (Kwak et al., 2026) schedules attention-guided regional updates. These attack strategies focus on transferability rather than explicitly separating geographic evidence from other content. Privacy-oriented protection instead seeks to prevent sensitive inference while preserving useful content (Liu et al., 2026; Zhang et al., 2026; Cai et al., 2026b; Yang et al., 2025; Wang et al., 2025). GeoShield (Liu et al., 2026) combines text-assisted feature decomposition with local exposure-region optimization. ReasonBreak (Zhang et al., 2026) guides perturbation generation with region-grounded concepts and hard-negative embedding priors. However, textual proxies can overlook instance-specific visual details, while feature pooling can obscure clue differences. NOWHERE constructs a multidimensional evidence subspace from grounded visual clues. It dynamically prioritizes residual evidence during multi-view optimization and uses a complementary scene reference to preserve content.

3 Methodology

3.1 Preliminary

Problem Definition. We study visual geoprivacy protection for images shared online. The models that may later infer image locations are unknown, so we adopt a black-box setting, *i.e.* the protector has no access to the architecture, parameters, training data, or gradients of the target VLM f_t. Given an image I_n with ground-truth coordinates $G_n = (\phi_n, \lambda_n)$, we add a perturbation δ_n before publication to obtain the protected image $I_n^{\mathrm{p}} = I_n + \delta_n$. The target VLM then predicts $\widehat{G}_n = f_t(I_n^{\mathrm{p}})$. Our goal is to increase the geographic distance between $\widehat{G}_n$ and G_n while preserving the image's visual content:

$$\max_{\delta_n} \; D_{\mathrm{geo}}(f_t(I_n + \delta_n), G_n) \quad \text{s.t.} \quad \|\delta_n\|_\infty \leq \epsilon, \tag{1}$$

Here, D_{geo} denotes the Haversine distance in kilometers. A larger distance indicates stronger protection against accurate localization. The budget ϵ bounds the change to each pixel channel. NOWHERE pursues this objective by optimizing evidence-based losses on accessible surrogate encoders, without querying the target model during perturbation optimization.

Overview. We introduce NOWHERE, a black-box framework for geoprivacy protection. It protects high-resolution images under a small perturbation budget while preserving visual quality and semantic content. NOWHERE follows the principle of evidence-space transfer. It organizes attacks around scene evidence shared across heterogeneous models. As shown in Fig. 2, the framework consists of two modules. Evidence Space Construction (ESC) builds multidimensional evidence subspaces from grounded visual clues and establishes a complementary reference for scene preservation. Dynamic Evidence Tracking (DET) tracks residual evidence and adjusts clue priorities during multi-view optimization. These priorities updates toward clues that remain insufficiently suppressed.

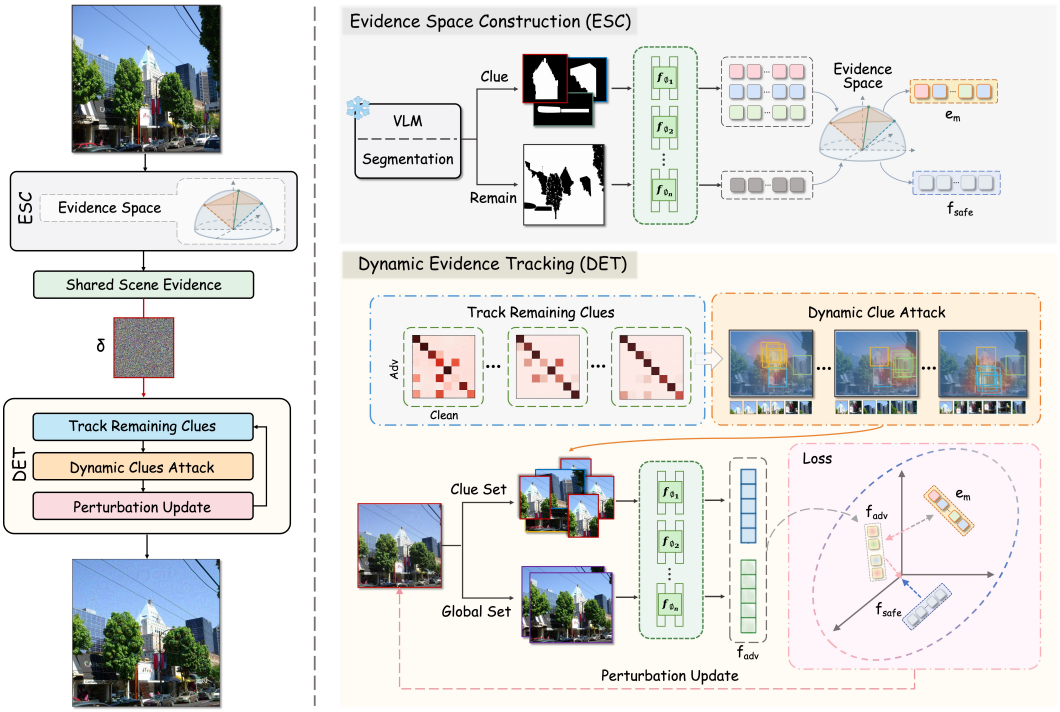


Figure 2: **An overview of NOWHERE.** ESC builds a multidimensional evidence space from localized visual clues and provides a reference for scene preservation, and DET tracks remaining evidence and reallocates local views to clues that remain strong.

3.2 Evidence Space Construction

Street-view images often contain geographic clues that reveal where the images were taken. Representing these clues through coarse text descriptions can miss scene-specific textures and structures. Averaging their visual features also merges distinct evidence into a single direction. ESC instead builds multidimensional evidence subspaces directly from visual clues in the image. This retains distinct evidence directions and gives different surrogate encoders the same scene clues to target.

Grounding instance-level visual evidence. ESC first locates geographic evidence within the image. Given an image I, we combine a VLM's geographic clue identification with segmentation method (Wang et al., 2024; Carion et al., 2026). This produces a set of evidence masks $\mathcal{M}(I) = \{M_k\}_{k=1}^{K}$, where $M_k \in \{0,1\}^{H\times W}$ marks the k-th clue region. For each clue, we keep the pixels inside its mask and replace the surrounding pixels with the image's mean color B. This reduces interference from unrelated content. The local clue view is defined as:

$$V_k = \mathcal{T}_k\left(M_k \odot I + (1 - M_k) \odot B\right), \tag{2}$$

Here, $\mathcal{T}_k$ extracts and resizes the square context crop around clue k. This operation preserves the clue's appearance in a focused local view, allowing the encoder to extract its material, texture, and structural features directly. The resulting views $\{V_k\}_{k=1}^{K}$ serve as separate visual inputs for constructing the geographic evidence subspaces.

Multidimensional geographic evidence subspaces. Different clues provide distinct localization evidence. Averaging their features into one vector can obscure these differences. We therefore construct a multidimensional subspace to retain multiple evidence directions. Let $\{f_m\}_{m=1}^{L}$ denote the surrogate encoders and $\bar{f}_m(X) = f_m(X)/\|f_m(X)\|_2$ their normalized visual features. For each encoder, we extract clue features $c_{m,k} = \bar{f}_m(V_k)$ and form the evidence matrix:

$$C_m = [c_{m,1}, \ldots, c_{m,K}] \in \mathbb{R}^{d_m \times K}, \tag{3}$$

where d_m is the feature dimension. The columns of C_m keep the clue features separate. To form a compact attack target, we seek a subspace that retains as much of their combined feature energy as possible. For a chosen dimension r_m, we select a basis $U \in \mathbb{R}^{d_m \times r_m}$ by solving:

$$U_m^\star \in \underset{U^\top U = \mathbf{I}_{r_m}}{\arg\max} \sum_{k=1}^{K} \|U^\top c_{m,k}\|_2^2. \tag{4}$$

The projection norm measures how much of each clue feature is retained in the subspace. The constraint keeps the basis vectors orthonormal, so they do not repeat the same direction.

This objective is solved by the uncentered SVD $C_m = A_m \Sigma_m W_m^\top$. The r_m left singular vectors associated with the largest singular values form the optimal basis $U_m^\star$. We choose r_m using relative singular values and the rank limit $r_{\max}$, with $1 \le r_m \le \operatorname{rank}(C_m)$. The resulting evidence subspace and its projector are:

$$\mathcal{S}_m(I) = \operatorname{span}(U_m^\star), \qquad P_m = U_m^\star (U_m^\star)^\top. \tag{5}$$

During optimization, P_m extracts image feature components along these evidence directions, providing the geographic evidence targets for suppression.

The evidence subspace provides the geographic reference direction used during optimization. To guide preservation of the remaining scene, we construct a reference image I^{rem} by replacing the selected clue regions with the image's mean color B. We then remove its feature component along the evidence directions, so the preservation reference lies outside the subspace targeted by the attack:

$$q_m = (\mathbf{I}_{d_m} - P_m)\bar{f}_m(I^{\mathrm{rem}}). \tag{6}$$

Here, $\mathbf{I}_{d_m}$ is the identity matrix. The complementary reference q_m guides preservation of the remaining scene content during optimization. Details are described in Appendix A.2.

3.3 Dynamic Evidence Tracking

ESC identifies the geographic evidence in the original image, but perturbations do not weaken all clues equally. For example, one clue's features may change substantially, while those of nearby clues remain insufficiently perturbed. A static attack may continue focusing on the weakened clue while overlooking the others. DET tracks this uneven progress and adjusts clue priorities using both remaining evidence and the effects of recent updates. These priorities determine how local views are allocated, shifting optimization toward clues that still need suppression.

Residual evidence tracking. To assess what remains, we compare each clue with its representation in the clean image. Let $I_t = I + \delta_t$ denote the protected image at iteration t, and let $\mathcal{R}_k$ extract the same fixed region around clue k from both images. For encoder m, the current feature is $v_{m,k}^t = \bar{f}_m(\mathcal{R}_k(I_t))$, and its clean reference is $v_{m,k}^0 = \bar{f}_m(\mathcal{R}_k(I))$. We compare these features along the evidence directions selected by ESC. We also include full-space similarity to account for the clue's broader visual representation. The resulting retention score is:

$$s_{m,k}^t = \lambda_e \operatorname{sim}_{\cos}\left(P_m v_{m,k}^t, P_m v_{m,k}^0\right) + (1-\lambda_e) \operatorname{sim}_{\cos}\left(v_{m,k}^t, v_{m,k}^0\right). \tag{7}$$

Here, $\operatorname{sim}_{\cos}$ denotes cosine similarity, and $\lambda_e \in [0,1]$ balances the two comparisons. A higher score indicates greater similarity to the clean reference. We use this as a signal that the clue needs further suppression. The same clue may retain its original features in some encoders but change more in others. For cross-model protection, we therefore consider both its typical retention and the disagreement among encoders. We measure these using the median and median absolute deviation:

$$\mu_k^t = \operatorname{median}\{s_{m,k}^t\}_{m=1}^L, \qquad d_k^t = \operatorname{median}\{|s_{m,k}^t - \mu_k^t|\}_{m=1}^L. \tag{8}$$

We penalize disagreement to favor clues with high retention and more consistent responses across encoders:

$$R_k^t = \max\left\{0, \frac{1 + \mu_k^t - \beta d_k^t}{2}\right\}. \tag{9}$$

Here, $\beta \ge 0$ controls the disagreement penalty. DET periodically updates R_k^t to redirect optimization toward clues that still need suppression.

Multi-view sustained suppression. The updated priorities indicate which clues need further suppression. We use local views to focus on these clues and global views to include the broader scene layout. Following the global–local multi-crop design in DINO (Caron et al., 2021), DET combines both types of views during optimization. Each local crop is sampled around a selected clue and contains at least part of it. As priorities change, DET reallocates the fixed local-view budget, assigning more crops to higher-priority clues. Let $\mathcal{V}_g^t$ and $\mathcal{V}_l^t$ denote the global and local view sets at iteration t. For encoder m, a view transformation T gives the feature:

$$z_{m,T}(\delta) = \bar{f}_m\left(T(I+\delta)\right). \tag{10}$$

Following the multi-crop aggregation in M-Attack-V2 (Zhao et al., 2026), we average per-view losses over views and encoders within each branch. We then combine the global and local objectives, guiding updates with scene context and DET-prioritized clues. See Appendix A.3 for details.

Optimization target. We use the geographic evidence subspace to define an image-specific reference direction for perturbation optimization. Let $z_m^0 = \bar{f}_m(I)$ denote the clean-image feature. Its normalized projection $e_m = P_m z_m^0 / \|P_m z_m^0\|_2$ defines the geographic reference direction. With the complementary scene reference q_m from ESC, the loss for each view is $\mathrm{sim}_{\cos}(z_{m,T}(\delta), z_m^0) + \mathrm{sim}_{\cos}(z_{m,T}(\delta), e_m) - \lambda_t \langle z_{m,T}(\delta), q_m \rangle$. Here, $\mathrm{sim}_{\cos}$ denotes cosine similarity, and $\lambda_t \geq 0$ controls scene preservation.

We average this loss over all encoders and views within each branch, giving the global loss $\mathcal{L}_g^t(\delta)$ and local loss $\mathcal{L}_l^t(\delta)$. We use Adam-style projected updates (Kingma & Ba, 2014) to optimize their combined objective:

$$\min_{\delta} \ \mathcal{L}_g^t(\delta) + \mathcal{L}_l^t(\delta) \quad \text{s.t.} \quad \|\delta\|_\infty \leq \epsilon. \tag{11}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate NOWHERE on Google Street View (Jay et al., 2025) and Im2GPS3k (Hays & Efros, 2008). Google Street View contains 1,602 images from 1,563 cities across 88 countries and regions. Im2GPS3k contains 2,997 geotagged images from Flickr and other sources. All images have GPS coordinates and are processed to 640×640 pixels.

Baselines. We compare NOWHERE with five transferable attacks: AnyAttack, SSA-CWA, M-Attack, ReasonBreak and GeoShield. We follow each method's original attack settings. Targeted attack baselines use target images randomly sampled from MSCOCO (Lin et al., 2014).

Implementation Details. We use four CLIP surrogates: ViT-B/16, ViT-B/32, ViT-g-14-laion2B-s12B-b42K, and ViT-B-32-laion2B-s34B-b79K. Perturbations are optimized for 300 steps under an ℓ_∞ budget of $\epsilon = 8/255$, using Adam-style projected updates with a step size of $1.275/255$. We set the projection similarity weight $\lambda_e = 0.75$ and disagreement penalty $\beta = 0.25$. DET starts dynamic allocation after a 20-step warm-up and refreshes clue priorities every 20 steps. We evaluate four black-box targets: GPT-5.6-Terra, GPT-5.4, Claude-Opus-4.8, and Gemini-3.1-Pro. Unless otherwise specified, GPT-5.5 identifies geographic clues as the auxiliary VLM. Optimization requires no target-model queries and runs on four NVIDIA H100 GPUs. Hyperparameter sensitivity analyses are provided in Appendix B.

Evaluation Metrics. We report geolocation accuracy within 1, 25, 200, 750, and 2,500 km, using the Haversine distance between predicted and true coordinates. Lower accuracy indicates stronger privacy protection. To assess semantic preservation, we use the same prompt to generate location-agnostic VLM descriptions of clean and protected images. We compare these descriptions using BLEU, ROUGE, and BERTScore (BERT-S). Prompt details are described in Appendix D.

4.2 Main Results

Effectiveness. Table 1 shows that NOWHERE achieves the lowest geolocation accuracy across four black-box VLMs and two datasets. It outperforms both general attack baselines and existing

Table 1: Geolocation accuracy (%; lower is better) on Google Street View and Im2GPS3k at five distance thresholds. NOWHERE achieves the lowest accuracy across all settings, indicating stronger geoprivacy protection. Best results are **bold**.

Method	Google Street View					Im2GPS3k				
	1km	25km	200km	750km	2500km	1km	25km	200km	750km	2500km
GPT-5.6-Terra										
Clean	15.2	26.1	53.3	82.6	96.3	19.7	55.3	67.8	80.2	90.9
AnyAttack	11.3	19.8	39.6	47.2	71.5	13.4	32.3	45.6	53.4	66.8
SSA-CWA	10.1	15.5	33.2	51.5	70.9	6.8	25.2	29.7	36.0	57.8
M-Attack	14.0	20.6	37.3	53.9	73.4	16.1	37.5	48.8	58.3	70.2
ReasonBreak	7.9	13.2	35.7	60.9	79.4	14.4	27.9	42.3	60.0	70.8
GeoShield	5.9	9.4	25.4	36.6	51.1	8.8	15.5	27.1	32.5	45.2
NOWHERE	**1.4**	**2.7**	**6.3**	**19.1**	**30.5**	**4.2**	**9.0**	**10.8**	**17.1**	**30.5**
GPT-5.4										
Clean	10.2	22.4	50.1	75.6	94.8	17.6	48.2	58.9	73.6	88.1
AnyAttack	9.1	17.6	39.3	63.8	83.9	11.9	29.4	42.8	64.2	76.1
SSA-CWA	8.9	16.8	37.9	62.1	79.3	10.8	27.9	38.9	62.0	71.6
M-Attack	8.6	14.6	32.7	56.8	72.0	9.1	23.2	30.3	51.5	64.6
ReasonBreak	6.1	11.2	31.4	58.4	78.0	9.2	22.2	34.3	54.6	67.7
GeoShield	4.6	8.1	21.3	39.9	54.5	6.4	12.9	20.9	33.7	39.3
NOWHERE	**0.9**	**2.0**	**6.2**	**14.6**	**31.3**	**2.8**	**7.7**	**10.2**	**16.1**	**29.8**
Claude-Opus-4.8										
Clean	6.0	16.6	37.1	72.7	89.9	18.3	58.0	71.5	82.0	89.4
AnyAttack	4.7	15.4	37.8	65.2	81.9	12.5	40.1	51.4	62.5	76.4
SSA-CWA	4.8	8.2	24.6	50.1	67.5	9.9	23.7	38.8	51.1	60.2
M-Attack	6.4	9.1	27.7	54.3	76.2	12.9	36.7	45.3	52.7	62.9
ReasonBreak	4.5	11.3	25.3	43.9	59.7	11.7	23.4	36.8	46.5	57.8
GeoShield	3.3	7.4	17.2	36.3	49.5	10.6	21.9	27.0	32.3	42.4
NOWHERE	**1.6**	**3.2**	**7.4**	**10.8**	**22.0**	**5.5**	**9.8**	**11.3**	**13.6**	**20.5**
Gemini-3.1-Pro										
Clean	16.3	30.9	56.0	84.8	95.4	22.7	61.4	76.1	86.4	93.2
AnyAttack	15.6	28.3	58.7	80.6	91.1	18.7	47.6	56.2	66.1	80.7
SSA-CWA	13.8	18.5	39.2	67.4	88.3	9.0	21.6	25.8	31.5	39.6
M-Attack	14.3	27.0	49.4	78.5	92.2	21.5	49.8	61.6	70.4	79.5
ReasonBreak	12.1	22.5	38.2	75.3	89.1	13.9	38.3	49.7	69.3	81.4
GeoShield	9.8	15.6	31.9	57.3	73.2	14.4	35.4	41.2	51.9	61.2
NOWHERE	**7.2**	**10.1**	**18.8**	**33.6**	**53.0**	**7.8**	**18.7**	**21.1**	**25.9**	**37.8**

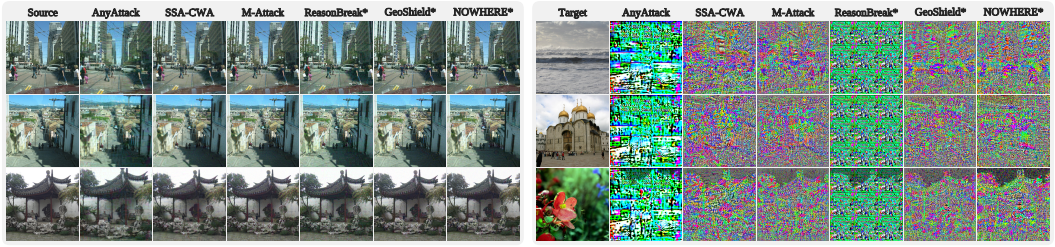


Figure 3: Adversarial images and perturbations. NOWHERE largely preserves the original scene appearance. * denotes untargeted attacks; the Target image (right) is used only for targeted attacks.

geoprivacy methods in these settings. On Im2GPS3k, NOWHERE reduces GPT-5.6-Terra's accuracy within 1 km from 6.8% to 4.2%. Accuracy within 25 km drops from 15.5% to 9.0%. Results on Google Street View support its effectiveness across datasets and black-box models.

Semantic Preservation. Table 2 compares location-agnostic descriptions of clean and protected images. NOWHERE maintains semantic consistency comparable to existing geoprivacy methods. Although it does not achieve the highest score on every semantic metric, it combines competitive semantic preservation with lower localization accuracy in most settings. These results indicate that NOWHERE strengthens privacy protection while retaining the main content users wish to share.

Table 2: Semantic consistency at $\epsilon = 8/255$ and $16/255$ (higher is better). NOWHERE achieves scores comparable to GeoShield and ReasonBreak at both budgets.

Method	LLaVA-1.5						GPT-5.6-Terra					
	BLEU		ROUGE		BERT-S		BLEU		ROUGE		BERT-S	
	8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255
AnyAttack	0.48	0.43	0.59	0.55	0.94	0.94	0.14	0.13	0.32	0.31	0.90	0.91
SSA-CWA	0.45	0.32	0.57	0.47	0.93	0.92	0.13	0.11	0.31	0.28	0.91	0.90
M-Attack	0.22	0.15	0.39	0.34	0.91	0.89	0.09	0.08	0.26	0.25	0.90	0.88
ReasonBreak	0.27	0.24	0.45	0.40	0.92	0.91	0.12	0.10	0.30	0.27	0.91	0.90
GeoShield	0.25	0.20	0.41	0.37	0.92	0.91	0.11	0.10	0.29	0.27	0.90	0.89
NOWHERE	0.28	0.23	0.46	0.41	0.92	0.92	0.12	0.10	0.31	0.28	0.91	0.89

Table 3: Ablation studies of NOWHERE. Removing either ESC or DET weakens geoprivacy protection. Avg. Dis. is measured in km; Acc. denotes localization accuracy within 25 km (%).

Configuration	GPT-5.6-Terra		Claude-Opus-4.8		Semantic Consistency↑		
	Acc.↓	Avg. Dis.↑	Acc.↓	Avg. Dis.↑	BLEU	ROUGE	BERT-S
w/o ESC	14.3	5,157	16.0	5,511	0.1146	0.3004	0.9018
w/o DET (Global-only)	13.1	5,738	13.5	5,929	**0.1274**	0.3058	0.9031
NOWHERE	**9.0**	**6,961**	**9.8**	**7,020**	0.1181	**0.3090**	**0.9075**

Visualization. Fig. 3 compares protected images and perturbations from different methods. In these examples, NOWHERE preserves scene layout, subject appearance, and local details, with less visible noise than AnyAttack and M-Attack. NOWHERE reduces the introduction of unrelated target semantics and better preserves the authenticity of the image content.

4.3 Ablation Studies and Discussion

Impact of ESC & DET. We remove ESC and DET separately to examine the roles of visual evidence construction and dynamic multi-view optimization. We first replace ESC's visual evidence with a single text prototype. We form this prototype by averaging and normalizing text features from VLM-generated geographic descriptions. Subtracting it from the full-image feature gives the scene reference. As shown in Table 3, Acc@25 rises from 9.0% to 14.3% on GPT-5.6-Terra and from 9.8% to 16.0% on Claude-Opus-4.8, indicating weaker protection. With ESC retained, we next examine whether DET further improves how this evidence is used. We remove DET while keeping the evidence suppression and content preservation objectives unchanged. Each step uses only the full image, without evidence tracking or local-view allocation. Protection again weakens. Average geolocation error falls from 6,961 to 5,738 km on Terra and from 7,020 to 5,929 km on Claude. These results support both ESC's visual evidence construction and DET's dynamic multi-view optimization. The full method also maintains comparable semantic consistency and achieves the highest ROUGE and BERT-S scores in this comparison.

Effect of crop count. Fig. 4 examines global and clue-local view counts. With two global views, increasing local views from two to six lowers 25 km accuracy from 11.41% to 9.00%. Average distance stays near 7,000 km. More local views lower accuracy to at best 8.67%, only 0.33 percentage points further, with no consistent increase in average distance. With six local views, increasing global views from one to two raises average distance from 6,547 km to 6,961 km. Accuracy also falls from 11.02% to 9.00%. Additional global views provide no consistent gain. We therefore use two global and six clue-local views to balance protection and computation.

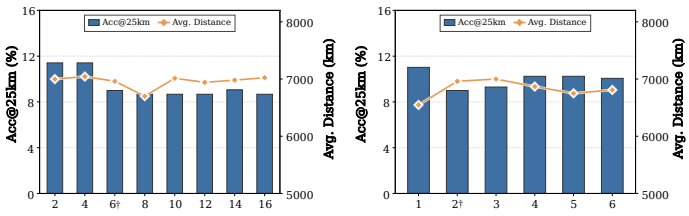


Figure 4: Ablation analysis of crop count.

Table 4: Ablations of evidence space construction and local-view allocation. Retaining multiple evidence directions and dynamically allocating clue-guided views improve protection.

(a) Evidence space construction.

Reference	GPT-5.6-Terra		Claude-Opus-4.8	
	Acc.↓	Avg. Dis.↑	Acc.↓	Avg. Dis.↑
Mean	14.8	5,933	14.6	6,201
$r_{\max} = 1$	12.9	6,116	13.1	6,359
$r_{\max} = 2$	12.0	6,269	11.7	6,631
$r_{\max} = 4$	11.0	6,307	10.3	6,794
$r_{\max} = 8$	10.8	6,270	10.6	6,881

(b) Clue guidance and allocation.

Crop	Method	GPT-5.6-Terra	
		Avg. Dis.↑	Acc.↓
Random	Static	5,948	13.2
Random	Dynamic	6,115	12.0
Clue	Static	6,307	11.0
Clue	Shuffled	6,083	12.7
Clue	**Dynamic**	**6,961**	**9.0**

We examine whether preserving differences between geographic clues improves geoprivacy protection. Using the same visual clue features, we compare normalized averaging with subspace-based geographic references at $r_{\max} \in \{1, 2, 4, 8\}$. The scene-preservation reference remains fixed, and all attacks are static. Table 4(a) shows that $r_{\max} = 4$ lowers localization accuracy within 25 km from 14.8% to 11.0% on GPT and from 14.6% to 10.3% on Claude compared with averaging. Average distance also increases on both models. Rank-1 improves on averaging but remains weaker than multi-direction constructions, supporting the benefit of retaining additional directions. Increasing the rank limit to 8 yields mixed results, with no consistent further improvement.

Clue guidance and dynamic attacks. We examine whether protection benefits from focusing on geographic clues and adjusting their priorities during optimization. We compare cropping and allocation strategies on GPT-5.6-Terra with a fixed view count. As shown in Table 4(b), replacing random crops with clue-guided crops under static allocation raises average distance from 5,948 to 6,307 km and lowers localization accuracy within 25 km from 13.2% to 11.0%. Dynamic allocation further improves these results to 6,961 km and 9.0%, respectively. To examine whether this gain depends on which clues receive priority, we shuffle the priorities among clues. Average distance then drops to 6,083 km, while accuracy rises to 12.7%. This drop supports using clue-specific feedback to guide local-view allocation, rather than simply changing the allocation during optimization.

Robustness to Transformation. Shared images often undergo further processing before they are viewed. We test NOWHERE under JPEG compression and Gaussian blur with $\epsilon = 8/255$. As shown in Fig. 5, average geolocation errors remain high across the tested settings. JPEG compression causes fluctuations across quality factors, but the errors stay well above the clean-image baseline. Under Gaussian blur, this gap also persists as blur strength increases. These results show that NOWHERE continues to limit location inference after both transformations, supporting its use for privacy-preserving image sharing.

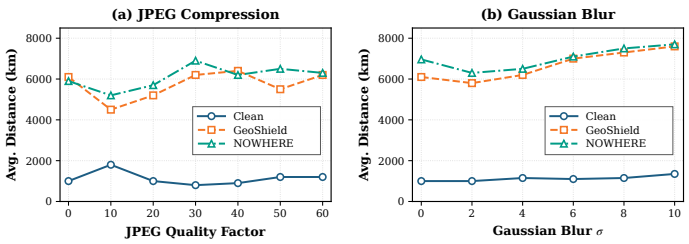


Figure 5: Robustness to image transformations.

5 Conclusion

We propose NOWHERE, a black-box geoprivacy protection framework based on the idea of evidence-space transfer. By representing distinct geographic clues and adapting the attack as they weaken, NOWHERE focuses perturbations on shared evidence across models. Experiments on two benchmarks show stronger protection against black-box geolocation while preserving visual appearance and non-geographic content. This helps users reduce location exposure before sharing photos, without access to the models that may later analyze them. Together, these findings suggest that shared visual evidence provides a useful basis for privacy protection across different models. Future work may explore this approach for other forms of sensitive visual inference.

AI USE STATEMENT

During the preparation of this work, the authors used ChatGPT for language editing and grammar correction. Afterward, the authors reviewed and revised the content as needed and take full responsibility for this manuscript.

ETHICS STATEMENT

NOWHERE is designed to protect users' location privacy in shared images. We evaluate the method on public benchmarks. We recognize that its perturbations may also interfere with legitimate geolocation applications. We therefore encourage responsible use of the method for privacy protection and authorized research.

REPRODUCIBILITY STATEMENT

We describe NOWHERE's components and optimization procedure in the methodology section. The experimental setup reports the public datasets, image preprocessing, surrogate and target models, hyperparameters, hardware, and evaluation metrics used in our experiments.

REFERENCES

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. Vicinagearth, 3(1):1, 2026.

Wei Cai, Shujuan Liu, Jian Zhao, Ziyan Shi, Yusheng Zhao, Yuchen Yuan, Tianle Zhang, Chi Zhang, and Xuelong Li. When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 35322–35330, 2026a.

Wei Cai, Jian Zhao, Yuchen Yuan, Tianle Zhang, Ming Zhu, Haichuan Tang, and Xuelong Li. Visual attention reasoning via hierarchical search and self-verification. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8986–8997, 2026b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In International conference on learning representations, volume 2026, pp. 138846–138923, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In 2021 IEEE/CVF international conference on computer vision (ICCV), pp. 9630–9640. IEEE, 2021.

Seung-Min Choi, Seung-Ik Lee, Jae-Yeong Lee, and In So Kweon. Semantic-guided de-attention with sharpened triplet marginal loss for visual place recognition. Pattern Recognition, 141: 109645, 2023.

Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23182–23190, 2023.

Kaiyuan Cui, Yige Li, Yutao Wu, Xingjun Ma, Sarah Erfani, Christopher Leckie, and Hanxun Huang. Toward universal and transferable jailbreak attacks on vision-language models. arXiv preprint arXiv:2602.01025, 2026.

Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhipei Huang, and Zhenjun Han. Gaga: Towards interactive global geolocation assistant. arXiv preprint arXiv:2412.08907, 2024.

Zhiyang Dou, Xumeng Han, Fengde Peng, Zipeng Wang, Moxuan Zhao, Zhipei Huang, and Zhenjun Han. Geobridge: Decoupled semantic conditioning for generative image geolocalization. arXiv preprint arXiv:2608.11838, 2026.

Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. IEEE Transactions on Information Forensics and Security, 20:1333–1348, 2024.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 814–825, 2025.

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

Yuxiang He, Jian Zhao, Yuchen Yuan, Tianle Zhang, Wei Cai, Haojie Cheng, Ziyan Shi, Ming Zhu, Haichuan Tang, Chi Zhang, et al. Aetheria: A multimodal interpretable content safety framework based on multi-agent debate and collaboration. arXiv preprint arXiv:2512.02530, 2025.

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on clip. arXiv preprint arXiv:2505.05528, 2025.

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

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment. Advances in Neural Information Processing Systems, 38:57993–58028, 2026.

Yuchu Jiang, Jian Zhao, Yuchen Yuan, Tianle Zhang, Yao Huang, Yanghao Zhang, Yan Wang, Yanshu Li, Xizhong Guo, Yusheng Zhao, et al. Never compromise to vulnerabilities: A comprehensive survey on ai governance. arXiv preprint arXiv:2508.08789, 2025.

Modi Jin, Yiming Zhang, Boyuan Sun, Dingwen Zhang, Ming-Ming Cheng, and Qibin Hou. Geoagent: Learning to geolocate everywhere with reinforced geographic characteristics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 41352–41364, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Jaehyun Kwak, Nam Cao, Boryeong Cho, Segyu Lee, Sumyeong Ahn, and Se-Young Yun. When and where to attack? stage-wise attention-guided adversarial attack on large vision language models. arXiv preprint arXiv:2602.04356, 2026.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. arXiv preprint arXiv:2506.14674, 2025.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. 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, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740–755. Springer, 2014.

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

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In International Conference on Learning Representations, volume 2024, pp. 23701–23722, 2024.

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

Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In Proceedings of the European conference on computer vision (ECCV), pp. 563–579, 2018.

Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, Tony Wang, et al. Failures to find transferable image jailbreaks between vision-language models. In International Conference on Learning Representations, volume 2025, pp. 44669–44704, 2025.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 536–551, 2018.

Jiawei Shao and Xuelong Li. Ai flow at the network edge. IEEE Network, 40(1):330–336, 2025.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems, 32, 2019.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4064–4072, 2020.

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

Leyan Wang, Jian Zhao, Xin Zhang, Xizhong Guo, Yuchen Yuan, Tianle Zhang, Jiaming Chu, Yuchu Jiang, Xu Yang, Lei Jin, et al. Erf-ba-tfd+: a multimodal model for audio-visual deepfake detection. Vicinagearth, 2(1):10, 2025.

Zipeng Wang, Xuehui Yu, Xumeng Han, Wenwen Yu, Zhixun Huang, Jianbin Jiao, and Zhenjun Han. P2seg: Pointly-supervised segmentation via mutual distillation. In International Conference on Learning Representations, volume 2024, pp. 45351–45362, 2024.

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

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. Cross-modal safety mechanism transfer in large vision-language models. In International Conference on Learning Representations, volume 2025, pp. 58315–58336, 2025.

Anwen Yang, Jian Zhao, Yuchen Yuan, Tianle Zhang, Yuchu Jiang, Jiaming Chu, Xin Zhang, Xu Yang, Lei Jin, Chi Zhang, et al. Aperture: a patch-aware framework for joint forgery detection and localization. Vicinagearth, 2(1):19, 2025.

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

Jiaming Zhang, Che Wang, Yang Cao, Longtao Huang, and Wei Yang Bryan Lim. Disrupting hierarchical reasoning: Adversarial protection for geographic privacy in multimodal reasoning models. In International Conference on Learning Representations, volume 2026, pp. 67992–68012, 2026.

Xiwu Zhang, Lei Wang, and Yan Su. Visual place recognition: A survey from deep learning perspective. Pattern Recognition, 113:107760, 2021.

Xiaohan Zhao, Zhaoyi Li, Yaxin Luo, Jiacheng Cui, and Zhiqiang Shen. Pushing the frontier of black-box lvlm attacks via fine-grained detail targeting. arXiv preprint arXiv:2602.17645, 2026.

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

A Appendix

A.1 Geographic Clue Extraction and Segmentation

ESC needs to know which scene details may reveal location and where they appear in the image. We first ask an auxiliary VLM to identify candidate clues and locate them with bounding boxes. SAM 3 then uses the clue descriptions and boxes to produce segmentation masks.

Identifying candidate clues. To connect clue selection with location inference, we ask GPT-5.5 to estimate where the clean image was taken and identify the visual evidence supporting its judgment. The prompts shown in Fig. 7 and Fig. 8 request a concise location estimate, a brief evidence summary, and at most five non-redundant clues in a single response. Each clue includes a description and a tight bounding box. No ground-truth coordinates are provided. The estimated location is not used as an adversarial target; only the clue descriptions and boxes are passed to SAM 3.

Images without usable clues. Some images have no usable clues after segmentation and cropping. In this case, we cannot construct the visual evidence subspace, so we use a previously generated and cached image description to define an alternative reference. For encoder m, let g_m be the normalized clean-image feature and a_m the normalized text feature of this description. Their difference gives a reference direction, while its orthogonal complement provides a space for retaining other image features:

$$u_m = \frac{g_m - a_m}{\|g_m - a_m\|_2}, \qquad Q_m^{\text{text}} = \mathbf{I} - u_m u_m^\top. \tag{12}$$

For a normalized perturbed-view feature z, we maximize

$$J_0(z) = -\operatorname{sim}_{\cos}(z, g_m) - z^\top u_m + \left(Q_m^{\text{text}} z\right)^\top \left(Q_m^{\text{text}} g_m\right). \tag{13}$$

The first two terms reduce alignment with the clean feature and the description-based reference direction. The last term encourages agreement with the clean feature outside that direction. All three terms have unit weight, and the projected features are not renormalized. The references remain fixed throughout optimization. Without clue regions to guide local attention, we use two global views and randomly sampled local views, with no clue tracking or dynamic allocation. We average the scores over views and encoders within each branch. Starting from zero perturbation, we perform 300 updates using separate global and local Adam states. We take the sign of their combined direction and apply a step of one pixel level, clipping the perturbation to $[-8, 8]$ after each update. The final image is clipped to $[0, 255]$. This procedure does not use the candidate-weight checks described for images with usable clues. If the required cached description is missing, the image is recorded as an incomplete case rather than processed with an undefined reference.

From clue descriptions to masks. For each candidate clue, the VLM describes the visible detail, explains why it may help infer location, and provides a bounding box. The box uses corner coordinates normalized to $[0, 1]$. If no clue is found, the model returns an empty list with a brief explanation. The VLM also reports how confident it is about the predicted location. We record this score but do not use it to keep or discard clues. Before segmentation, we check that the required descriptions are present and that each box encloses a nonempty region within the image. These checks confirm that the descriptions and boxes can be used as segmentation inputs. They do not test whether a clue actually helps locate the image.

Although a bounding box indicates where a clue appears, it may also contain surrounding content. We use SAM 3 to obtain a mask that follows the clue region more closely. Each clue is processed separately, using its name as a text prompt and its box as a positive geometric prompt. The box is expressed by its center coordinates, width, and height, and the prompts are reset between clues. The official SAM 3 image processor resizes the input to 1008×1008 and maps its predictions back to the original image dimensions. For each clue, we merge all instance masks with SAM 3 confidence above 0.5 into one binary mask. Masks from different clues may overlap.

Preparing clues for optimization. Clue extraction and segmentation use the full source image. For perturbation optimization, images are processed to 640×640 pixels, with the corresponding masks kept spatially aligned. For each usable clue, we use its mask to select a square image crop

containing the visible clue region, without adding extra margins. If this crop extends beyond the image boundary, we shift it inward without changing its size. We then resize it to 640×640 and apply the preprocessing required by each encoder. The clues and masks remain fixed during optimization; DET changes how local views are allocated among them. When the VLM returns a valid empty clue list, we use global crops and randomly sampled local crops without clue-based tracking or allocation. The same procedure applies when segmentation and preprocessing leave no usable masks. Missing or invalid VLM responses and missing required mask files are handled differently: processing stops and the failure is reported. These missing or invalid outputs must be resolved before optimization can proceed.

A.2 Evidence-Space Construction

Extracting visual features for individual clues. To preserve differences among geographic clues, ESC first extracts a separate visual feature for each clue. Simply cropping around a clue may still include nearby objects and background details, mixing their appearance with that of the selected region. We therefore retain the pixels inside the clue mask and replace the surrounding content with the image's mean color. Let $x_0 \in [0, 255]^{3\times H\times W}$ denote the clean image used for optimization, M_i the binary mask of clue i, and $\mu(x_0)$ the mean RGB color of the full image. The clue image is

$$x_i^{\text{clue}} = M_i \odot x_0 + (1 - M_i) \odot \mu(x_0). \tag{14}$$

This leaves the selected clue's pixels unchanged while reducing visual detail outside its mask. We then take the square crop associated with the clue, using the bounds described in Appendix A.1. Pixels within this crop but outside the mask remain mean-filled. The resulting view focuses feature extraction primarily on the geographic clue, while some non-geographic information may still be retained. We encode each view with four frozen CLIP encoders: ViT-B/16, ViT-B/32, ViT-g-14-laion2B-s12B-b42K, and ViT-B-32-laion2B-s34B-b79K. All four encoders receive the same clue view, but each represents it in its own feature space. Before encoding, we resize the shorter side to 224 pixels using bicubic interpolation, take a 224×224 center crop, and apply CLIP's RGB-channel normalization. The perturbation is optimized on the 640×640 image, while these smaller views provide the encoder inputs. After encoding, we normalize each feature to unit length. We use $\bar{f}_m$ to denote encoder m together with its input preprocessing and feature normalization, and write $c_{m,i}$ for the resulting feature of clue i. These vectors remain separate and form the columns of the evidence matrix constructed next.

Constructing the evidence space and reference features. Different geographic clues may share some visual features while differing in others. ESC needs to represent both aspects, so we keep the clue features as separate columns of $C_m = [c_{m,1}, \ldots, c_{m,K}]$ for each encoder. We apply SVD directly to this matrix without subtracting the mean clue feature. This choice lets the construction consider components shared across clues, rather than describing only how each clue differs from their average. The resulting orthogonal directions are ordered by the feature energy they capture. However, the number of directions available depends on the image: an additional clue does not necessarily provide an independent feature direction. We therefore determine the subspace size from the singular values, subject to an upper limit. For $\sigma_{m,1} \geq \sigma_{m,2} \geq \cdots$, we retain

$$r_m = \min\left(r_{\max}, \#\left\{j : \sigma_{m,j} \geq 10^{-5}\sigma_{m,1}\right\}\right), \qquad r_{\max} = 4. \tag{15}$$

The relative threshold removes numerically negligible directions, while $r_{\max}$ limits the subspace to four dimensions. This is an upper bound rather than a fixed dimension for every image. For example, an image with only one usable clue provides at most one direction. More generally, the retained rank depends on both the number of clues and the dependence among their features, and may differ across encoders. We use the first r_m left singular vectors as the columns of U_m and define

$$P_m = U_m U_m^\top, \qquad Q_m = \mathbf{I} - P_m. \tag{16}$$

These projections separate a feature into components inside and outside the retained evidence space. The first provides the geographic reference direction, while the second helps construct a reference for preserving scene content. To build the preservation reference, we first need an image that places less emphasis on the selected clues. Using the original image directly would retain the same visual details that the geographic objective seeks to change. We therefore replace the selected clue regions

with the image's mean color. With $M = \bigvee_i M_i$ denoting the union of the visible clue masks, the remaining-content image is

$$x^{\mathrm{rem}} = (1 - M) \odot x_0 + M \odot \mu(x_0). \tag{17}$$

We use the masks without enlarging their boundaries, leaving the pixels outside the selected regions unchanged. Although the selected clue pixels have been replaced, the feature of this image may still contain components along the evidence directions. We therefore also apply the complementary projection when constructing the preservation reference. Let $g_m = \bar{f}_m(x_0)$ and $n_m = \bar{f}_m(x^{\mathrm{rem}})$ be the normalized features of the clean and remaining-content images. The two references are

$$e_m = \frac{P_m g_m}{\|P_m g_m\|_2}, \qquad q_m = Q_m n_m. \tag{18}$$

The geographic reference e_m describes the clean image's direction within the evidence space. The scene-preservation reference q_m retains the part of the remaining-content feature outside that space. These references serve different roles: e_m is normalized to specify a direction, whereas q_m keeps its projected magnitude as well as its direction. The preservation objective therefore uses the remaining feature component without rescaling it to unit length. We compute the evidence space and both references once from the clean image and keep them fixed throughout optimization. Within each encoder, global and local views share these references, while DET adjusts which clues receive more local views.

Guiding perturbation optimization. With these references in place, we can guide feature changes while retaining a reference for the scene content users wish to preserve. Moving away from the clean representation alone does not distinguish geographic clues from other visual details. We therefore give this change a geographic focus through the reference direction constructed by ESC. At the same time, the complementary scene reference guides content preservation. For the normalized feature $z_{m,v}$ of a perturbed view, we define two scores that optimization seeks to increase:

$$J_{\mathrm{sup}}(z_{m,v}) = -\operatorname{sim}_{\cos}(z_{m,v}, g_m) - \operatorname{sim}_{\cos}(z_{m,v}, e_m), \tag{19}$$

$$J_{\mathrm{pres}}(z_{m,v}) = z_{m,v}^{\top} q_m. \tag{20}$$

Here, $\operatorname{sim}_{\cos}$ denotes cosine similarity, and $z_{m,v}^{\top} q_m$ is the inner product between the current feature and the scene-preservation reference. The two suppression terms serve different purposes. The first encourages departure from the original image representation, while the second reduces alignment with the geographic reference direction derived from the selected clues. The preservation score complements these changes by encouraging alignment with the remaining-scene feature outside the evidence space. We retain the projected magnitude of q_m, so this term uses an inner product rather than cosine similarity. We evaluate both scores across global and local views. Global views provide broader scene context, while local views expose details around the selected clues. Although these views show different parts of the image, they share the same fixed full-image references within each encoder. Their objectives therefore remain tied to the same scene as the sampled crops change. We average each score over the views and the four encoders within each branch, then combine the global and local branch averages with equal weight. This prevents the larger number of local crops from automatically increasing the local branch's weight. DET can thus redistribute local views among clues without changing the total weight assigned to that branch.

A.3 Dynamic Evidence Tracking and View Allocation

ESC provides a fixed description of the geographic evidence, but the response of each clue changes as perturbations develop. DET follows these changes and adjusts where local views are sampled. Its allocation considers both how closely a clue still resembles its clean reference and how much this similarity has decreased during recent updates.

Tracking clue retention and recent progress. To decide which clues need further attention, DET considers both their current feature retention and their response to recent updates. We first compare the same fixed crop around each clue in the clean and perturbed images. Keeping the crop boundaries unchanged allows this comparison to follow changes in the image rather than changes in the sampled region. Unlike the mean-filled views used by ESC, these crops retain the surrounding scene

content. ESC uses isolated clue features to construct the evidence space, whereas DET follows how each clue's contextual representation changes during optimization. Let $h^0_{m,i}$ and $h^t_{m,i}$ denote the normalized clean and current features of clue i under encoder m. We compare them within the evidence space and across the full feature representation:

$$s^t_{m,i} = 0.75\ \mathrm{sim}_{\cos}\left(U_m^\top h^t_{m,i}, U_m^\top h^0_{m,i}\right) + 0.25\ \mathrm{sim}_{\cos}\left(h^t_{m,i}, h^0_{m,i}\right). \tag{21}$$

Here, $\mathrm{sim}_{\cos}$ denotes cosine similarity. The first term emphasizes changes along the evidence directions, while the second also accounts for changes in the full contextual feature. A higher score indicates that the current feature remains closer to its clean reference. We combine the scores across the four encoders using the median and disagreement penalty in Eq. 9. We denote this aggregated retention score at the k-th tracking refresh by E^k_i. Retention alone, however, does not describe recent progress. Two clues can have similar current scores even when recent updates have reduced one score more than the other. DET therefore also considers the observed decrease since the previous refresh, relative to the number of local views assigned to the clue. Let N^{k-1}_i count these views. We compute

$$b^k_i = \frac{\max(E^{k-1}_i - E^k_i, 0)}{\max(N^{k-1}_i, 1)}. \tag{22}$$

Only a decrease in retention counts as positive progress. Dividing by the view count expresses this progress relative to the local-view allocation. Since all views update the same image, this quantity describes observed progress rather than the effect of a local crop alone. To avoid basing allocation entirely on the latest interval, we combine this feedback with its history. We first divide each b^k_i by the largest current value across clues, obtaining $\widetilde{b}^k_i$. If the largest value is at most 10^{-8}, we set all normalized progress scores to zero. We then update the running progress score as

$$A^k_i = 0.5A^{k-1}_i + 0.5\widetilde{b}^k_i, \qquad A^0_i = 1. \tag{23}$$

This gives equal weight to the previous estimate and the newly observed progress. Finally, we combine the running score with current retention to obtain the clue priority:

$$p^k_i = \max(E^k_i, 10^{-6})(A^k_i + 0.25). \tag{24}$$

The resulting priority favors clues whose features still resemble their clean references and have recently responded to optimization. The offset of 0.25 preserves a baseline contribution from retention even when recent progress is small. These priorities guide the next allocation of local views, allowing DET to consider both what remains unchanged and where continued optimization has shown progress.

Allocating and sampling views. The priorities guide where DET spends additional local views, while every usable clue continues to receive attention. We keep a fixed budget of six local views and two global views per update. During the first 20 updates, local views are assigned through balanced round-robin allocation. This gives each clue initial attention and provides a period over which DET can observe feature changes before using them to guide allocation. Dynamic allocation begins before update 21 and is refreshed every 20 updates thereafter. At each refresh, we first place the priorities on a common scale by subtracting their mean and dividing by their population standard deviation across clues. A softmax with temperature 0.5 then converts them into normalized weights. When the priorities are equal or numerically indistinguishable, we use uniform weights. To maintain attention beyond the highest-priority clues, we mix the softmax weights with a uniform distribution, assigning 25% of the total weight uniformly. We also cap each weight at $3/K$, where K is the number of usable clues. Any excess is redistributed among weights below the cap in proportion to their remaining capacity, followed by normalization. These weights determine how the available local views are shared. Since the default setting contains at most five clues and six local views, we first assign one view to every clue. We distribute the remaining views according to the weights using largest-remainder rounding, as detailed in Algorithm 1. Specifically, we take the whole-number part of each weighted share and assign any leftover views to the largest fractional remainders. Exact ties are resolved by rotating the clue order across refreshes, so the same ordering does not repeatedly favor one clue. The resulting counts remain fixed until the next refresh, but crop positions are sampled anew at every update. Thus, repeated attention to a clue can involve different views of its visible region. To sample a 224×224 local crop, we first choose a pixel uniformly from the assigned clue mask. We then uniformly select a valid crop position within the optimization image

Algorithm 1: Allocating six local views among geographic clues

Input: Normalized clue weights w; $1 \leq K \leq 5$ clues; local-view budget $N_\ell = 6$; refresh index k
Output: View counts n, with $\sum_{i=1}^{K} n_i = N_\ell$

$n_i \leftarrow 1$ for every clue
$R \leftarrow N_\ell - K$
if $R > 0$ **then**
  $u_i \leftarrow Rw_i$ for every clue
  $n_i \leftarrow n_i + \lfloor u_i \rfloor$ for every clue
  $B \leftarrow R - \sum_{i=1}^{K} \lfloor u_i \rfloor$
  Rank clues by $u_i - \lfloor u_i \rfloor$ in descending order
  Resolve exact ties using a cyclic clue order determined by k
  Add one view to each of the first B clues
return n

that contains this pixel. Each crop therefore intersects its assigned clue without requiring the entire clue to fit inside the crop. Different local views may overlap. Alongside these clue-focused views, we sample two global crops at every update to retain broader scene context. Their area fractions are drawn from $[0.5, 0.9]$ of the optimization image, and their aspect ratios are drawn from $[3/4, 4/3]$. Both crops are resized to 640×640. The global views remain active throughout optimization, providing scene-level context as DET shifts local attention among clues.

The allocated views determine which image regions contribute to each update. We use their gradients to change the original and geographic feature representations while guiding content preservation. Each update must also remain within the allowed pixel perturbation budget. Algorithm 2 summarizes the full NOWHERE procedure, including evidence construction, dynamic view allocation, and perturbation optimization.

Perturbation updates. Changing the image representation and preserving scene content do not always favor the same pixel changes. We therefore adjust the preservation guidance according to its interaction with the suppression objective. Let g and h denote the suppression and preservation gradients summed across the global and local branches. When $g^\top h < 0$, we remove the component of h that opposes g and distribute this correction across branches. We then try decreasing weights γ on the adjusted preservation gradients to find a balance between the two goals. The global and local branches maintain separate Adam states, and every trial starts from copies of the same current states. Their update directions are added without taking the sign. Before projection, the proposed update u must satisfy $g^\top u > 0$ and $h^\top u \geq 0$, where h remains the original preservation gradient. We then project the candidate image onto the perturbation budget and valid pixel range. Using the same sampled views, we check whether the suppression score increases while the preservation score remains unchanged or improves, allowing a small numerical tolerance. The first accepted trial determines both the image update and the preservation weight $\lambda_t = \gamma$. If no trial passes, we set $\lambda_t = 0$ and use only the current suppression gradients, retaining the existing Adam history without applying these acceptance checks again. We perform 300 updates from zero perturbation, with $\|\delta\|_\infty \leq 8$ and pixel values in $[0, 255]$. All gradients come from the frozen surrogate encoders, without target-model queries.

B Hyperparameter Sensitivity

Evaluation setting. We investigate the sensitivity of NOWHERE to two hyperparameters: the projection similarity weight λ_e and the cross-model disagreement penalty β. Protection is evaluated on Claude-Opus-4.8 and GPT-5.6-Terra using localization accuracy within 25 km (Acc@25) and average geolocation error. Lower Acc@25 and larger average error indicate stronger protection. We vary one hyperparameter at a time while keeping the evaluation images and other attack settings unchanged. The tested values are $\lambda_e \in \{0, 0.25, 0.5, 0.75, 1\}$ and $\beta \in \{0, 0.125, 0.25, 0.5, 1\}$. Unless varied, these parameters are fixed at $\lambda_e = 0.75$ and $\beta = 0.25$.

Projection similarity weight. The parameter λ_e balances evidence-subspace similarity and full-feature similarity when estimating clue retention. The endpoints correspond to relying exclusively

Algorithm 2: NOWHERE with dynamic local-view allocation

Input: Clean image x; frozen surrogate encoders $\{f_m\}_{m=1}^{4}$; $T = 300$
Output: Protected image

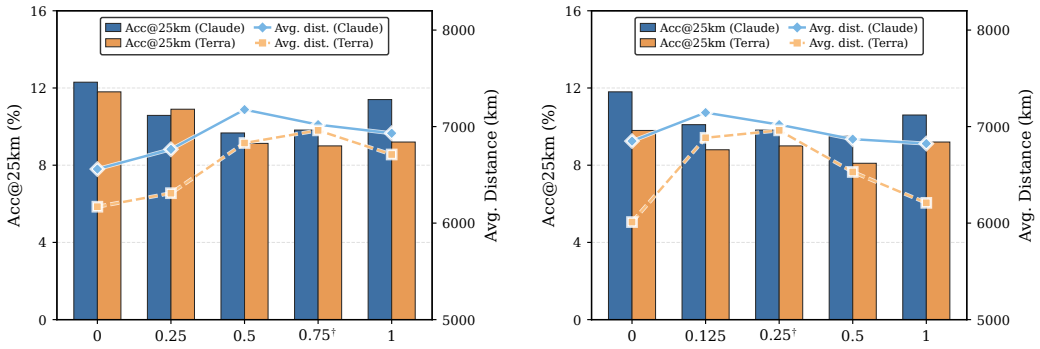
Identify geographic clues and obtain their masks with SAM 3
Align the image and masks to the 640×640 image x_0
if no usable clue remains **then**
    Protect x_0 using the no-clue procedure in Appendix A.1
    return the protected image
Construct fixed ESC references and clean clue-context features
Initialize $\delta \leftarrow 0$ and separate global and local Adam states
Initialize clue retention, progress estimates, and view counts
for $t \leftarrow 0$ **to** $T - 1$ **do**
    if $t \geq 20$ and $t \bmod 20 = 0$ **then**
        Update clue retention and recent progress
        Convert the resulting priorities into allocation weights
        Allocate six local views using Algorithm 1, with $k = t/20$
        Store the current retention scores and reset interval view counts
    Sample two global views and six local views using warm-up or current allocation
    Update the per-clue view counts
    Compute suppression and preservation gradients for each branch
    Adjust conflicting preservation gradients using the combined branch gradients
    Try decreasing preservation weights from copies of the current Adam states
    For each trial, sum the two Adam directions without taking the sign and project the candidate image
    if a trial passes the gradient and image-score checks **then**
        Accept the first passing candidate and its Adam states
        Set λ_t to its preservation weight
    else
        Set $\lambda_t = 0$ and update using current suppression gradients with the retained Adam history
        Project the image without applying the acceptance checks
return the quantized protected image with the perturbation budget enforced

on full-feature similarity ($\lambda_e = 0$) or projected similarity ($\lambda_e = 1$). As shown in Fig. 6(a), intermediate weights provide better protection than either endpoint in the tested settings. Claude achieves its lowest Acc@25 and largest average error at $\lambda_e = 0.5$, whereas Terra performs best on both metrics at $\lambda_e = 0.75$. Increasing the weight to 1 weakens protection on both models. These observations support combining the two similarity signals rather than relying exclusively on either representation. The different preferred values also indicate that the best balance is model-dependent; we use $\lambda_e = 0.75$ as the common setting.

Cross-model disagreement penalty. The parameter β reduces the priority of clues whose retention estimates disagree across surrogate models. Setting $\beta = 0$ removes this penalty. Fig. 6(b) shows that moderate nonzero penalties improve both metrics over this setting. However, the two metrics do not identify the same optimum. Both target models achieve their lowest Acc@25 at $\beta = 0.5$, while the largest average error occurs at $\beta = 0.125$ for Claude-Opus-4.8 and $\beta = 0.25$ for GPT-5.6-Terra. In particular, increasing β from 0.25 to 0.5 further reduces Acc@25 but also decreases average error. This difference reflects the complementary roles of the metrics: Acc@25 measures the frequency of relatively precise localization, whereas average error summarizes the magnitude of prediction errors. We therefore report both metrics and use $\beta = 0.25$ as a compromise rather than a uniformly optimal value.

C Visual Fidelity

Protecting location privacy should also preserve the appearance of shared photos. We therefore compare each protected image with its corresponding clean input on the Im2GPS3k dataset. We use PSNR, SSIM, and LPIPS to measure pixel fidelity, structural similarity, and perceptual differences, respectively. We compute PSNR in RGB with a data range of 255. SSIM uses an 11×11 Gaussian window with $\sigma = 1.5$, population covariance, and constants $K_1 = 0.01$ and $K_2 = 0.03$. LPIPS



(a) Projection weight λ_e.

(b) Disagreement penalty β.

Figure 6: Hyperparameter sensitivity of NOWHERE on Claude-Opus-4.8 and GPT-5.6-Terra. Bars show localization accuracy within 25 km (lower is better), and lines show average geolocation error (higher is better). Each panel varies one parameter while keeping the others at their default values. † marks the default setting.

Table 5: Visual fidelity on the Im2GPS3k dataset at a resolution of 640×640. Higher PSNR and SSIM and lower LPIPS indicate better fidelity.

Method	PSNR↑ (dB)	SSIM↑	LPIPS↓
AnyAttack	30.9590	0.8971	0.2712
M-Attack	34.2472	0.8899	0.1817
SSA-CWA	34.6774	0.9079	0.2457
ReasonBreak	32.7038	0.8600	0.3360
GeoShield	34.1756	0.8832	0.1909
NOWHERE	32.3136	0.8302	0.2625

uses AlexNet with version-0.1 calibration and RGB inputs mapped to $[-1, 1]$. All three metrics are computed for each image and then averaged; PSNR is averaged in decibels.

Table 5 shows that the methods differ across visual-quality measures. NOWHERE has lower LPIPS than AnyAttack and ReasonBreak, indicating smaller perceptual differences from the clean images. M-Attack and GeoShield achieve higher PSNR and SSIM and lower LPIPS than NOWHERE. These results complement the localization experiments in the main text: the localization scores measure how well a method protects location privacy, while the fidelity scores describe how much it changes the shared image.

D Prompts

D.1 Geographic Clue Extraction

We ask the auxiliary VLM to identify up to five geographic clues and provide their descriptions and bounding boxes. These outputs guide clue segmentation. Fig. 7 and Fig. 8 present the system prompt and user template used for this task.

D.2 Geolocation and Semantic Consistency Evaluation

We ask the target VLM to estimate where an image was taken. Figs. 9 and 10 present two prompt formats. The coordinate-only format requests latitude and longitude. The extended format also requests the country, city, region, confidence, and a brief explanation. For semantic consistency evaluation, Fig. 11 presents the prompt for generating image descriptions without identifying or inferring geographic locations.

Geographic Clue Extraction System Prompt

ROLE:
You are the single-stage visual geolocation cue detector for an academic image-privacy experiment.

TASK:
Your task is to infer where the image was taken, then detect and localize the visible image regions
that actually helped you make that prediction. Return one concise location judgment, a brief
evidence-based reasoning summary, and a small non-redundant list of cue names with tight bounding
boxes. The reasoning summary must be concise and must not reveal hidden chain-of-thought.

SPECIAL NOTES:
Use only visible image evidence. A cue may be text, architecture, road design, vehicle details,
vegetation, geology, coastline form, terrain, or a useful spatial structure. Do not list ordinary
objects that did not affect the prediction. Do not create a cue merely to fill the list. If the image has no useful localizable cue, return an empty cue list. Boxes are approximate prompts for
SAM 3, but they must be as tight as practical and should exclude people and unrelated foreground.

Do not output neutral regions, positive/negative points, thresholds, tiers, masks, or a second-stage
verification result. Return JSON only.

Figure 7: Geographic Clue Extraction System Prompt.

Geographic Clue Extraction User Prompt

Image ID: {image_id}
Teacher model: {teacher_model}
Return at most {max_cues} non-redundant cue regions.

OUTPUT FORMAT:
Return exactly one JSON object:

```
{
 "schema_version": "rg_gcc_teacher_simple_box_v1",
 "image_id": "{image_id}",
 "teacher_model": "{teacher_model}",
 "predicted_location": "brief country/region/city judgment, or unknown",
 "prediction_confidence": 0.0,
 "reasoning_summary": "brief explanation of how the visible cues support the judgment",
 "cue_regions": [
   {
    "cue_id": "cue_00",
    "cue_name": "short visible cue name",
    "why_useful": "one short sentence explaining how this region helped geolocation",
    "bbox_xyxy_norm": [0.0, 0.0, 0.2, 0.2]
   }
 ],
 "no_useful_cue_reason": null
}
```

SPECIAL NOTES:
Coordinates are normalized to [0,1] in [x1,y1,x2,y2] order. Each box must have positive area. Use cue IDs cue_00, cue_01, ... in order. When cue_regions is empty, no_useful_cue_reason must be a short non-empty string; otherwise it must be null.

Figure 8: Geographic Clue Extraction User Prompt.

Geolocation Evaluation Prompt Coordinate-only Format

ETHICS STATEMENT:
This is a controlled academic benchmark using authorized or publicly available evaluation images.

TASK:
The task is designed only to evaluate visual geolocation capability in a research setting, not to identify or track any private individual. Inspect the image and return JSON only with exactly two keys: latitude and longitude. Both values must be numeric decimal coordinates of the most likely location. If uncertain, provide your best estimate. Do not include explanations or any extra text.

OUTPUT FORMAT:
Return exactly one JSON object:

```
{
 "latitude": xxx.xxx,
 "longitude": xxx.xxx
 }
```

Figure 9: Geolocation Evaluation Prompt Coordinate-only Format.

Geolocation Evaluation Prompt Extended Format

ETHICS STATEMENT:
This is a controlled academic benchmark using authorized or publicly available evaluation images.

TASK:
The task is designed only to evaluate visual geolocation capability in a research setting, not to identify or track any private individual. Inspect the image and return JSON only with keys: latitude, longitude, country_guess, city_guess, region_guess, coarse_geo_confidence, reason. Latitude and longitude must be numeric decimal coordinates of the most likely location. If uncertain, provide your best estimate and lower confidence. Do not include any extra text.

OUTPUT FORMAT:
Return exactly one JSON object:

```
{
 "latitude": xxx.xxx,
 "longitude": xxx.xxx,
 "country_guess": "xxxx",
 "city_guess": "xxxx",
 "region_guess": "xxxx",
 "coarse_geo_confidence": "xxxx",
 "reason": "xxxx"
 }
```

Figure 10: Geolocation Evaluation Prompt Extended Format.

Semantic Evaluation Prompts

ETHICS STATEMENT:
This is a controlled academic benchmark using authorized or publicly available evaluation images.

TASK:
Describe the visible content of this image. Focus on the main objects, their appearance, activities, and spatial relationships. Do not identify or infer the geographic location, or include place names, addresses, coordinates, or landmark names. Describe only what is visible and avoid speculation. Return only the description.

Figure 11: Semantic Consistency Evaluation Prompt.