GEOGRAPHIC FLOW ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern image-geolocation models can recover location directly from visual content, turning images into *latent geographic metadata* and creating new privacy risks. We study how to protect images against generative geolocation models using imperceptible adversarial perturbations. We introduce **Geographic Flow Attacks (`GeoFA`)**, a unified framework that organizes attacks by *where* they act, *what* they deviate from or towards, and *how* deviation is measured. This reveals a mismatch in training-loss attacks: inference follows the model's predicted velocity field, not the training target. We therefore propose **Geographic Flow Attack by Trajectory Deviation (`GeoFA-TD`)**, which maximizes angular deviation between clean and perturbed velocities to redirect the sampling trajectory without backpropagating through it. Across YFCC4K and OSV-5M, `GeoFA` attacks outperform the state of the art. `GeoFA-TD` outperforms all attacks on OSV-5M and transfers from spherical RFM to diffusion. Code and models will be released.

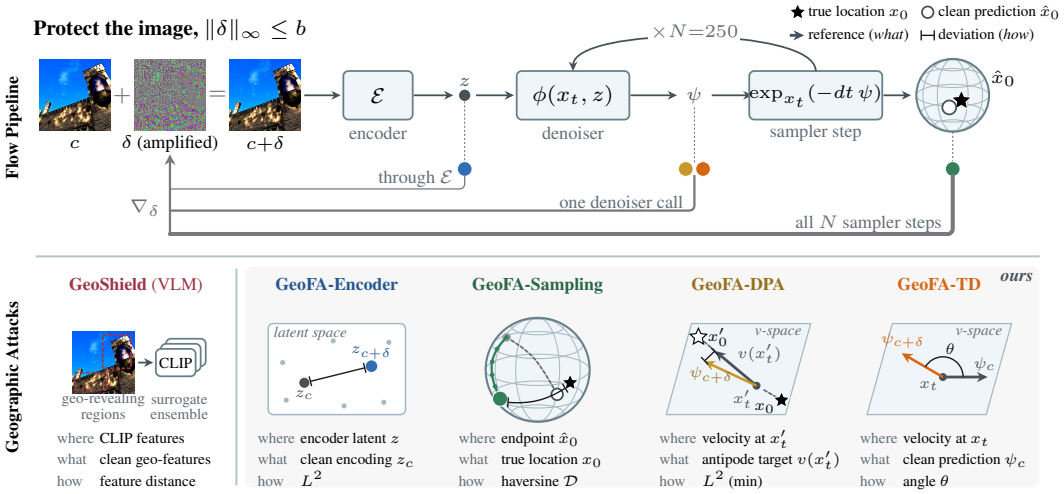


Figure 1: **Geographic Flow Attacks (GeoFA).** *Top:* A perturbation δ is added to the image c, the encoder $\mathcal{E}$ turns it into a condition z, and the flow model transports noise to a location $\hat{x}_0$ on the sphere over N sampler steps. An attack can target the encoder output (GeoFA-Encoder), the model's velocity prediction (GeoFA-DPA, GeoFA-TD), or the sampled location (GeoFA-Sampling); grey paths show how far its gradient travels back to δ. *Bottom:* GeoShield attacks black-box VLMs through geo-features, whereas we focus on flow models. Our GeoFA are each defined by *where* they act, *what* they deviate from or towards, and *how* they measure that deviation (Table 1).

1 INTRODUCTION

Images reveal more than what they visibly depict. A photograph shared online may unintentionally disclose where someone lives, works, travels, or regularly spends time; images taken at hospitals, schools, shelters, places of worship, or private residences can reveal particularly sensitive information. Traditionally, users could mitigate this risk by removing GPS metadata before sharing an image. This assumption is rapidly becoming obsolete: modern image-geolocation models can reconstruct geographic location directly from visual content, turning seemingly innocuous pixels into

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.

a form of *latent geographic metadata*. Protecting images from automatic geolocation is therefore becoming an increasingly important privacy problem.

This threat is amplified by recent progress in generative geolocation. Specialized models based on Diffusion and Flow Matching now infer geographic coordinates directly from images and have repeatedly advanced the state of the art (Dufour et al., 2025; Wang et al., 2026b). In particular, Riemannian Flow Matching on the sphere (RFM-$\mathcal{S}^2$) explicitly respects the topology of the Earth and achieves particularly strong performance (Dufour et al., 2025). Therefore, removing explicit location metadata is no longer sufficient: the image itself can contain enough information for a model to recover its geographic origin.

Existing geo-privacy methods primarily target VLMs, e.g. by disrupting geographic semantics in joint image-language representations (Liu et al., 2026) (see Table 1 for a comprehensive view). However, no such protection exists for specialized generative geolocation models despite their strong performance. Their prediction is not produced by a single classifier or retrieval operation but by iteratively transporting a sample through a learned geographic vector field.

A natural attack is to directly maximize the distance between the predicted and true location. For Diffusion and Flow Matching models, however, this requires backpropagating through the full sampling trajectory, making optimization costly, unstable, and sampler-dependent. Prior work therefore relies on surrogate objectives based on the encoder or training loss (Salman et al., 2023; Kang et al., 2023; Liang et al., 2023; Zheng et al., 2025; Shao et al., 2026). Yet these objectives were designed for generative models of images, not for models whose output is a point on the Earth.

Instead, in this work we organize these surrogates into **Geographic Flow Attacks (`GeoFA`)**, a unified framework characterized by three questions: (1) *where* to attack (encoder latents, velocity predictions, or sampled endpoints); (2) *what* to deviate from or towards (the clean encoding, training target, clean prediction, or an adversarial target); and (3) *how* to measure this deviation; summarized in Table 1. This view exposes a key mismatch in existing training-loss attacks: they deviate from the training target, while at inference the sampler follows the model's *own predicted velocity field*.

Motivated by this observation, we introduce **Geographic Flow Attack by Trajectory Deviation (`GeoFA-TD`)**, which attacks the sampling trajectory by maximizing the angular deviation between clean and perturbed velocity predictions. Unlike L^2 deviations, angular deviations change the *direction* of the trajectory rather than merely its speed, while remaining invariant to velocity scale.

All attacks in our `GeoFA`framework outperformn the state-of-the-art VLM-based geoprivacy defense GeoShield (Liu et al., 2026), and notably achieve near 100% attack success on both YFCC4K and OSV-5M even under small L_∞ perturbation budgets of $1/255$, and up to $8/255$[1]. The proposed `GeoFA-TD` further outperforms all other attacks on OSV-5M, and scores near 100%, on par with `GeoFA`-Sample and `GeoFA`-DPA attacks on YFCC4K. Furthermore, unlike all other attacks, `GeoFA-TD` transfers at almost no cost to the diffusion counterpart of RFM-$\mathcal{S}^2$ (same encoder) although it was not trained against it, as opposed to the drop for `GeoFA`-Sample and `GeoFA`-DPA.

Our contributions are: (1) We introduce **`GeoFA`**, a unified framework for privacy-preserving attacks against geographic generative models, organized around *where* to attack, *what* to deviate from, and *how* to measure that deviation. (2) We propose **`GeoFA-TD`**, which attacks the model's own sampling trajectory by maximizing angular deviation of its predicted velocity field, rather than reproducing the training objective. (3) We show across two geolocation benchmarks and multiple generative formulations that `GeoFA` attacks outperform the current state-of-the-art GeoShield, even under small L_∞ perturbation budgets. (4) We show that `GeoFA-TD` performs near 100 % accuracy on YFCC4K and OSV-5M, and outperforms any other attack on OSV-5M, while transferring across other models and maintaining success rate. We hope that this work will pave the way for privacy in geolocalization.

2 Related Work

Global visual geolocation (Hays & Efros, 2015) aims at predicting an image's localization using only its pixels. Recent work treats it as a generative task rather than regression or classification (Dufour et al., 2025; Yi & Shan, 2025). Dufour et al. (2025) train a Riemannian Flow Model RFM-$\mathcal{S}^2$

[1]For images normalized to $[0, 1]$, an L_∞ budget of $1/255$ (resp. $8/255$) limits the change of each pixel channel to at most 1 (resp. 8) intensity levels in the standard 8-bit representation.

Table 1: **GeoFA unified framework.** Attacks are characterized by *where* they act, *what* they deviate from or towards, and *how* this deviation is measured. Gray rows denote geographic adaptations evaluated in this work; bold rows are our proposed objectives. To our knowledge, prior geo-privacy methods do not consider specialized Diffusion or Flow-Matching geolocation models.

Setting	Model	Attack	Eq.	Where?	What?	How?	Access
General	VLM	M-Attack (2025)		enc. latent	target emb.	cosine (min)	CLIP ens.
		SSA-CWA (2024a)		enc. latent	target emb.	cosine (min)	surr. ens.
		AnyAttack (2025a)		enc. latent	target emb.	cosine (min)	surr. enc.
	Diffusion FM	Encoder (2023)		enc. latent	target latent	L^2 (min)	encoder
		Sample (2023)		sampler output	target image	L^2 (min)	full sampler
		AdvDM (2023)		den. output	training target	L^2 (max)	enc.+den.
		ACE (2025)		den.+enc.	target image	L^2 (min)	enc.+den.
		UniDef (2026)		den. output	training target	L^2 (max)	enc.+den. (fwd)
Geographic	VLM	GeoShield (2026)		enc. latent	clean geo-features	feat. dist. (max)	CLIP ens. + VLM/det.
	Diffusion FM	GeoFA-Encoder (2023)	5	enc. latent	clean encoding	L^2 (max)	encoder
		GeoFA-Sampling (2023)	6	endpoint	true location	haversine (max)	full sampler
		GeoFA-AdvDM (2023)	7	velocity	training target	L^2 (max)	enc.+den.
		GeoFA-ACE (2025)	8	velocity+latent	target image	L^2 (min)	enc.+den.
		GeoFA-UniDef (2026)	7	velocity	training target	L^2 (max)	enc.+den. (fwd)
		GeoFA-DPA (2024b)	9	velocity	antipode	L^2 (min)	enc.+den.
		GeoFA-L^2 (Ours)	10	**velocity**	**clean prediction**	**L^2 (max)**	enc.+den.
		GeoFA-TD (Ours)	13	**velocity**	**clean prediction**	**angular (max)**	enc.+den.
Robotic	Diffusion FM	DPA (2024b)		den. output	fixed target	L^2 (min)	enc.+den.

Access levels are nested: *encoder* $\subset$ *encoder+denoiser* $\subset$ *full sampler*.
(fwd.) denotes forward queries only, no backpropagation through the denoiser.

on the sphere conditioned on images; LocDiffusion (Wang et al., 2026b) extends this to the spherical harmonics space. Visual Geolocation has risen as an emerging capability of VLMs (Luo et al., 2025; Zhang et al., 2025c), fuelling efforts to adapt foundational models for this task (Yi & Shan, 2025).

Adversarial attacks on generative models. Diffusion (DM) and Flow Matching (FM) Models' strong generative performance (Ho et al., 2020; Song et al., 2021) has sparked concerns regarding data privacy, copyright and potential malicious use (Liang et al., 2023; Salman et al., 2023). Adversarial attacks on diffusion address this. Salman et al. (2023) craft attacks on the image encoder and approximate a direct sampling attack by backpropagating through a fraction of diffusion steps. DiffAttack Kang et al. (2023) optimizes a deviated-reconstruction loss at intermediate diffusion steps. Recent efforts focus on the training loss. AdvDM (Liang et al., 2023) maximizes it directly. ACE (Zheng et al., 2025) targets the score of a carefully designed target image. UniDef (Shao et al., 2026) refines this by targeting the global denoising level. Diffusion Policy Attack (DPA) (Chen et al., 2024b) extends to diffusion policies for robotic arms, again by designing a *suitable* target, but does not adapt existing attacks from the image literature. However, these refer to image diffusion; hence, here we traspose all attacks to a compact Riemannian geographic manifold (Section 4).

Visual geographic privacy has received specific attention given the aforementioned privacy concerns (Mendes et al., 2024). Existing general attacks against VLMs alter image features to align them with a target image (Li et al., 2025; Zhang et al., 2025b; Guo et al., 2023). The adversarial untargeted attack GeoShield Liu et al. (2026) protects the geo-privacy of images against VLMs by targeting geo-revealing regions and image features, while preserving other image semantics. BPPM (Wang et al., 2026a) uses generative obfuscation methods to regenerate privacy-revealing regions. Our work provides geographic generative counterparts to those VLM-centered privacy ones.

3 Background: Geographic Riemannian Flow Matching

Here, we formalize geographic Riemannian Flow Matching (RFM) on the sphere Dufour et al. (2025). Diffusion and Euclidean Flow Matching with our attacks are in Appendix B.

Notation. Let c be an image captured at geographic location $x_0^c \in \mathcal{S}^2$. The geographic model consists of an image encoder $\mathcal{E}$ and a velocity denoiser ϕ, predicting the velocity from x_t conditioned on the encoded image via Adaptive Layer Normalization (Xu et al., 2019): $\psi(x_t \mid c){=}\phi(x_t, \mathcal{E}(c))$, where $x_t \in \mathcal{S}^2$ is a noisy geographic coordinate and $\psi(x_t \mid c)$ is a tangent velocity at x_t.

Riemannian Flow Matching on $\mathcal{S}^2$. Starting from the true location x_0^c, noise $\epsilon \sim \mathcal{U}(\mathcal{S}^2)$ is sampled uniformly on the sphere. The forward path is the geodesic interpolation

$$x_t = \exp_{x_0}\big(\kappa(t)\log_{x_0^c}(\epsilon)\big), \qquad v(x_t) = \dot{\kappa}(t)\, D(x_t) \quad , \tag{1}$$

where the scheduling function $\kappa(0) = 0$ and $\kappa(1) = 1$, $\log$ and $\exp$ are the logarithmic and exponential maps on $\mathcal{S}^2$, and $D(x_t)$ denotes the tangent direction of the corresponding geodesic. The model ψ is trained to match this target velocity:

$$\mathcal{L}^{\text{RFM}} = \mathbb{E}_{x_0^c, c, \epsilon, t}\big[\|\psi(x_t \mid c) - v(x_t)\|_{x_t}^2\big] \quad , \tag{2}$$

where $\|\cdot\|_{x_t}$ and $\langle\cdot,\cdot\rangle_{x_t}$ denote the norm and inner product in the tangent space at x_t.

Sampling. At inference, sampling starts from $\epsilon \in \mathcal{S}^2$ and integrates the learned velocity field backwards from $t = 1$ to $t = 0$:

$$x_{t-dt} = \exp_{x_t}\big(-dt\,\psi_w(x_t \mid c)\big) \quad . \tag{3}$$

where ψ_w denotes the velocity field under Classifier-Free Guidance (CFG), with guidance scale w. The resulting endpoint $\hat{x}_0^c$ is the predicted geographic location.

Two properties will be central to our method: at inference, the sampler integrates the model prediction ψ rather than the training target v; and each predicted velocity determines the direction of the next sampling step on $\mathcal{S}^2$. Sections 5.1 and 5.2 exploit these two observations.

4 GeoFA: A Unified Framework for Geographic Flow Attacks

We now introduce **Geographic Flow Attacks (`GeoFA`)**, a unified framework for protecting images against geographic generative models. We formulate GeoFA for RFM-$\mathcal{S}^2$.

Attack objective. Following Goodfellow et al. (2015), let c be an image captured at location $x_0^c \in \mathcal{S}^2$. We perturb the image as $c+\delta$, where δ is constrained by $\|\delta\|_\infty \leq b$ to preserve visual similarity. Let $\hat{x}_0^{c+\delta}$ denote the location predicted by the geographic RFM sampler from the perturbed image. Our objective is to maximize the geographic error:

$$\delta^* = \underset{\|\delta\|_\infty \leq b}{\arg\max}\, \mathcal{D}\big(\hat{x}_0^{c+\delta}, x_0^c\big) \quad , \tag{4}$$

with $\mathcal{D}$ the haversine distance: the distance between the perturbed prediction and the true coordinates of c on the sphere. Directly optimizing Eq. 4 requires backpropagating through the full sampling trajectory. This is computationally expensive, produces a long gradient path, and makes the attack dependent on the sampler and its number of integration steps. We thus consider cheaper adversarial objectives, denoted by $\mathcal{L}_{\text{adv}}(\delta)$. All attacks use the same L_∞ constraint and PGD optimization (Madry et al., 2018); they differ only in the quantity they perturb and how deviation is measured.

The `GeoFA` framework. `GeoFA` organizes these objectives through three questions: **where** to attack (encoder latents, velocity predictions, or sampled endpoints); **what** to deviate from or towards (the clean encoding, training target, clean prediction, or an adversarial target); and **how** to measure this deviation (e.g., with L^2, haversine distance, or angular similarity). Table 1 summarizes this.

4.1 GeoFA Objectives

We now instantiate `GeoFA` with objectives adapted from prior adversarial attacks. `GeoFA`-Sampling directly optimizes the geographic error of Eq. 4; all remaining methods replace it with cheaper surrogate objectives that avoid backpropagating through the full sampling trajectory. They all use



Figure 2: **Qualitative comparison** of three images from YFCC4k under several attack baselines and GeoFA-TD at budget $8/255$. All attacks except for GeoShield display no visual artifacts.

the same L_∞ budget and PGD optimization; they differ only in *where* they attack, *what* they deviate from or towards, and *how* deviation is measured:

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-Encoder}}(\delta) = \|\mathcal{E}(c+\delta) - \mathcal{E}(c)\|^2 \quad , \tag{5}$$

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-Sampling}}(\delta) = \mathcal{D}\Big(\hat{x}_0^{c+\delta,\psi}, x_0^{c,\psi}\Big) \quad , \tag{6}$$

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-AdvDM}}(\delta) = \mathbb{E}_{\epsilon,t}\big[\|\psi(x_t \mid c+\delta) - v(x_t)\|_{x_t}^2\big] \quad , \tag{7}$$

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-ACE}}(\delta) = \mathbb{E}_{\epsilon,t}\big[\|\psi(x_t \mid c+\delta) - \psi(x_t \mid c')\|_{x_t}^2\big] + \alpha\|\mathcal{E}(c+\delta) - \mathcal{E}(c')\|_2^2 \quad , \tag{8}$$

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-DPA}}(\delta) = \mathbb{E}_{\epsilon,t}\big[\|\psi(x_t' \mid c+\delta) - v(x_t')\|_{x_t}^2\big] \quad . \tag{9}$$

The first three objectives are maximized, while the targeted objectives GeoFA-ACE and GeoFA-DPA are minimized. Here, we describe them briefly and Appendix C gives more details.

GeoFA-Encoder. Here, *where* is the encoder space, *what* is the clean encoding, and *how* is L^2 distance. It requires access only to $\mathcal{E}$ and can transfer across models sharing the same encoder. GeoShield (Liu et al., 2026) similarly attacks geographic information in image representations.

GeoFA-Sampling. This directly optimizes the geographic privacy objective of Eq. 4: *where* is the sampled endpoint, *what* is the true location, and *how* is haversine distance. Unlike the remaining objectives, it requires backpropagation through the full sampling trajectory.

GeoFA-AdvDM and GeoFA-UniDef. AdvDM (Liang et al., 2023) attacks the training objective: *where* is velocity space, *what* is the training target $v(x_t)$, and *how* is L^2 distance. GeoFA-UniDef adapts UniDef (Shao et al., 2026), which targets the same objective but replaces backpropagation through the denoiser with a finite-difference gradient estimate; details are given in Appendix C.6.

GeoFA-ACE and GeoFA-DPA. Targeted attacks instead optimize towards an adversarial target. GeoFA-ACE adapts ACE (Zheng et al., 2025) using a target image c', while GeoFA-DPA adapts DPA (Chen et al., 2024b) using a target location x_0'. For geographic prediction, we choose x_0' as the antipode of x_0^c, the maximally distant point on $\mathcal{S}^2$.

5 Geographic Flow Attack By Trajectory Deviation

We introduce **Geographic Flow Attack by Trajectory Deviation (GeoFA-TD)**, which directly disrupts the sampling dynamics followed at inference. GeoFA-TD deviates from the model's own predicted velocity field (Section 5.1) (rather than maximizing the training loss) and maximizes the angular change in velocity direction (rather than measuring this deviation with L^2) (Section 5.2). These yield a sampler-independent objective that avoids backpropagation through the full trajectory.

5.1 Deviate from the predicted velocity

Training-loss attacks such as GeoFA-AdvDM maximize deviation from the target velocity $v(x_t)$. At inference, however, $v(x_t)$ is never used: the sampler integrates the model prediction $\psi(x_t \mid c)$. We therefore replace the training target with the clean predicted velocity:

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-}L^2}(\delta) = \mathbb{E}_{\epsilon,t}\left[\|\psi(x_t \mid c+\delta) - \psi(x_t \mid c)\|_{x_t}^2\right] \quad . \tag{10}$$

This changes only *what* is attacked in the GeoFA framework: velocity space and the L^2 metric remain unchanged. The relation to the training-loss objective is explicit. Let $r(x_t) = v(x_t) - \psi(x_t \mid c)$ denote the clean model residual and $\Delta_\delta(x_t) = \psi(x_t \mid c+\delta) - \psi(x_t \mid c)$.

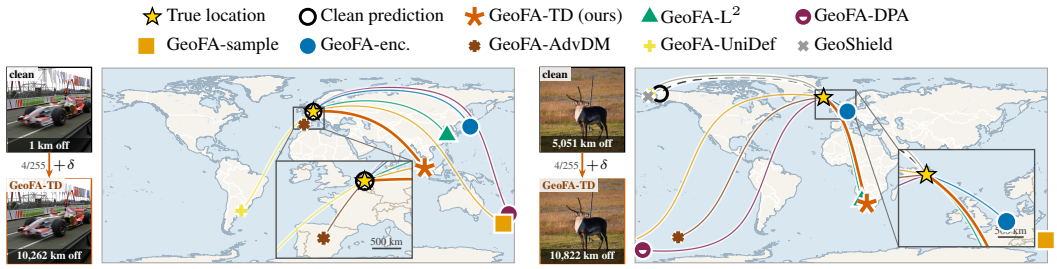


Figure 3: Our attacks on two random YFCC4K images ($b = 4/255$). Failure cases of GeoShield and GeoFA-AdvDM.

Proposition 1 (Training loss contains a residual term) *For any perturbation δ,*

$$\mathcal{L}_{\text{adv}}^{\textit{GeoFA-AdvDM}} = \mathcal{L}_{\text{adv}}^{\textit{GeoFA-}L^2} - 2\mathbb{E}_{\epsilon,t}\left[\langle \Delta_\delta(x_t), r(x_t)\rangle_{x_t}\right] + C \quad , \tag{11}$$

where $C = \mathbb{E}_{\epsilon,t}[\|r(x_t)\|_{x_t}^2]$ independent of δ.

The proof is given in Appendix A. Thus, GeoFA-AdvDM differs from directly perturbing the model prediction by a residual-dependent term. When this term vanishes in expectation, both objectives have the same expected dependence on δ; otherwise, the training-loss attack is additionally driven by the clean model's fitting error. This mismatch can become particularly relevant under classifier-free guidance, where the field followed at inference departs further from the training target.

5.2 Deviate in direction, not magnitude

Section 5.1 replaced the training target with the model's own prediction, but still measures deviation in L^2. An L^2 difference can increase either because the perturbed velocity changes *magnitude* or because it changes *direction*. Writing $\psi_\delta = \psi(x_t \mid c+\delta)$, $\psi_c = \psi(x_t \mid c)$ and θ for the angle between them, the two contributions separate exactly:

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-}L^2}(\delta) = \mathbb{E}_{\epsilon,t}\Big[\underbrace{\left(\|\psi_\delta\|_{x_t} - \|\psi_c\|_{x_t}\right)^2}_{\text{magnitude}} + \underbrace{2\,\|\psi_\delta\|_{x_t}\|\psi_c\|_{x_t}\,(1-\cos\theta)}_{\text{direction}}\Big] \quad . \tag{12}$$

Inflating the magnitude term raises the L^2 objective without moving the trajectory off the clean path. As the angular term saturates while the magnitude term does not, gradient ascent on Eq. 12 is biased towards inflating the magnitude rather than turning the field.

We therefore discard magnitude and attack direction alone, minimizing the cosine similarity of the clean and perturbed velocities in the tangent space at x_t. This yields **Geographic Flow Attack by Trajectory Deviation** (GeoFA-TD):

$$\delta^{\text{GeoFA-TD}} = \underset{\|\delta\|_\infty \le b}{\arg\min}\, \mathbb{E}_{\epsilon,t}\left[\frac{\langle \psi(x_t \mid c+\delta),\, \psi(x_t \mid c)\rangle_{x_t}}{\|\psi(x_t \mid c+\delta)\|_{x_t}\,\|\psi(x_t \mid c)\|_{x_t}}\right] \quad . \tag{13}$$

Thus, GeoFA-TD redirects the velocity actually followed by the sampler, while remaining invariant to its magnitude and avoiding backpropagation through the sampling trajectory.

6 Experiments

Here, we first review our experimental setup in Section 6.1, and then present our main results in Section 6.2. Since robustness is paramount when crafting attacks, we evaluate ablations, including sensitivity to compression, blur, varying sampling steps, as well as transferability (Section 6.3).

6.1 Experimental setup

Datasets. We evaluate attacks on 4 000-image subsets of two public datasets, each with its own encoder. (1) **OpenStreetView-5M (OSV-5M)** (Astruc et al., 2024), road-level images taken from

Table 2: Attack success rate (ASR, %, ↑) on OSV-5M and YFCC4k for three budgets, and attack cost. $\text{ASR}_{>d}$ is the fraction of images whose predicted location is pushed *more than* d km from the ground truth. All attacks use no restart (single attempt per image). Time is the wall-clock time of one optimisation step and VRAM the peak GPU memory allocated during a step, measured on one NVIDIA RTX A5000. GeoShield/GeoFAattacks run 100/60 steps per image respectively. Every Ours-GeoFA-$*$ row instantiates our framework. Best in **bold**, second underlined, third in *italics*.

		$b = 1/255$		$b = 4/255$		$b = 8/255$		Time↓	VRAM↓
	Attack	>750	>2500	>750	>2500	>750	>2500	(ms)	(GiB)
OSV-5M	GeoShield (2026)	55.43	34.88	98.47	92.42	99.75	96.78	651	7.80
	Ours-GeoFA-$\mathcal{E}$ (2023)	95.17	80.20	99.88	98.33	99.95	98.70	**124**	**1.85**
	Ours-GeoFA-Sample (2023)	99.17	97.25	*99.92*	*99.58*	99.95	*99.53*	232	2.93
	Ours-GeoFA-AdvDM (2023)	94.45	77.42	96.70	82.85	96.55	83.05	133	*1.91*
	Ours-GeoFA-ACE (2025)	95.95	83.78	97.30	87.33	97.12	87.28	193	1.93
	Ours-GeoFA-DPA (2024b)	98.47	*96.00*	99.95	99.75	**100.00**	99.72	133	1.90
	Ours-GeoFA-UniDef (2026)	80.60	65.25	89.15	76.95	88.60	75.88	*135*	**1.85**
	Ours-GeoFA-L^2	*98.72*	93.50	98.95	94.75	98.97	95.08	193	1.93
	Ours-GeoFA-TD	**99.97**	**99.60**	**100.00**	**99.95**	**100.00**	**99.97**	193	1.93
	Clean	23.80	n/a	*budget-independent*				–	–
YFCC4k	GeoShield (2026)	44.10	27.82	65.15	48.08	81.53	66.07	651	7.80
	Ours-GeoFA-$\mathcal{E}$ (2023)	89.10	73.78	95.53	84.00	96.17	83.95	**124**	**1.85**
	Ours-GeoFA-Sample (2023)	**97.05**	92.25	99.25	97.90	**99.38**	97.97	232	2.93
	Ours-GeoFA-AdvDM (2023)	82.78	63.27	87.83	70.00	88.30	70.58	133	*1.91*
	Ours-GeoFA-ACE (2025)	78.30	57.65	88.42	64.80	88.72	65.72	193	1.93
	Ours-GeoFA-DPA (2024b)	*95.30*	*91.42*	**99.28**	**98.08**	99.33	**98.10**	133	1.90
	Ours-GeoFA-UniDef (2026)	64.97	49.33	75.65	61.75	76.80	62.33	*135*	**1.85**
	Ours-GeoFA-L^2	85.08	78.58	87.53	82.42	87.83	82.92	193	1.93
	Ours-GeoFA-TD	96.45	**94.33**	*98.67*	*97.83*	*98.58*	*97.72*	193	1.93
	Clean	42.00	25.73	*budget-independent*				–	–

cars across 225 countries, encoded with a ViT-L (Dosovitskiy et al., 2021) finetuned on StreetClip (Haas et al., 2023). (2) **Yahoo Flickr Creative Commons 4K (YFCC4k)** (Vo et al., 2017), generic images, encoded with DinoV2 (Oquab et al., 2024). Attack hyperparameters are given in App. D.

Attack metrics. Following Liu et al. (2026), we report **Final Step Displacement** (FSD), the haversine distance (km) between the perturbed prediction and the ground truth coordinates of c, and **Attack Success Rate** (ASR), the fraction of images whose FSD exceeds a threshold, which captures an attack's regularity rather than its raw strength. Thresholds are 750 (country) and 2500 (continent) km, following the Visual Geolocation literature (Haas et al., 2024; Liu et al., 2026). Predictions are sampled with 250 steps and CFG scale $w = 10$, collapsing them to the modes of the distribution, consistent with our privacy focus on an adversary inferring a single precise location.

6.2 Comparison to the state of the art

General Performance. Table 2 reports ASR on OSV-5M and YFCC4k. Our best GeoFA attacks (GeoFA-TD/DPA/Sample) outperform the state-of-the-art GeoShield, at all budgets and on both datasets, despite much lower training costs. GeoShield is weakest at low budget, with almost no effect at $b = 1/255$ on YFCC4K, and trails on OSV-5M despite reaching high ASR at $b = 4, 8/255$: its VLM-oriented objective does not transfer to geographic RFM despite explicitly targeting geo-revealing features, and leaves visible artifacts (Figure 24).

GeoFA-TD, GeoFA-Sample and GeoFA-DPA score best across budgets. GeoFA-TD consistently outperforms all others on OSV-5M, and closely tracks GeoFA-DPA/Sample on YFCC4K, even surpassing them at the 2500 km threshold at low budget. We note that although GeoFA-Sample has a

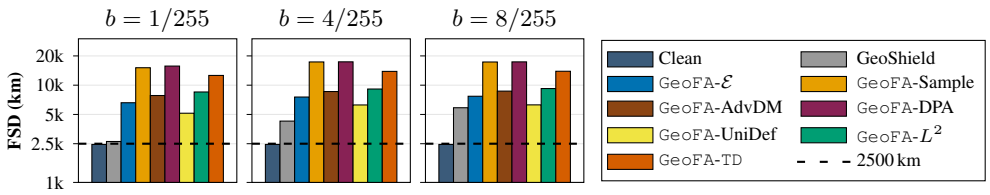


Figure 4: **FSD on YFCC4k (log scale.)** Dashed line at 2500 km.

strong performance, it remains costlier to train, less scalable to heavier models, and less robust (see Sec. 6.3). GeoFA-AdvDM and GeoFA-Unidef score significantly below our best attacks in ASR and FSD, confirming that simply maximizing training loss is insufficient.

Figure 4 displays FSD of our attacks. Here too, GeoFA-TD/Sample/DPA surpass all others. GeoFA-Sample/DPA are slightly above GeoFA-TD, although above 10000 km, such differences matter less than an attack's regularity in surpassing the 2500 km threshold.

Cross-paradigm Transferability We test whether perturbations trained against one geographic generative model transfer to another sharing the same image encoder. Figure 5 reports attacks trained on RFM-$\mathcal{S}^2$ and evaluated on Diffusion. Among our three strongest attacks, only GeoFA-TD transfers nearly fully; GeoFA-Sample and GeoFA-DPA drop by 10–30% across budgets. Since the encoder is shared, this suggests that GeoFA-TD targets transferable geo-revealing features rather than model-specific sampling dynamics.

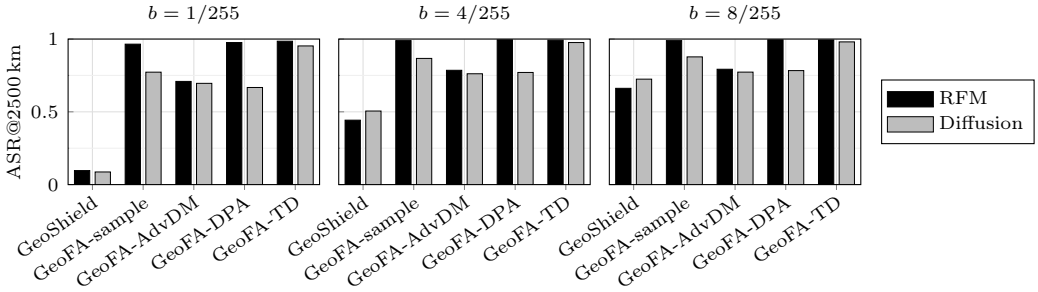


Figure 5: **Transferability of attacks, YFCC4k.** We show ASR of attacks trained against RFM $\mathcal{S}^2$ and evaluated against Diffusion and Flow Matching variants from Dufour et al. (2025). The image encoder is the same across models. GeoFA attacks are trained with 5 restarts.

Dataset and encoder variability. Table 2 shows that attack success is highly encoder dependent. StreetClip (used on OSV-5M) seems easier to attack for some attacks, especially GeoShield, than DinoV2 (used on YFCC4k). This is likely due to the fact that GeoShield is trained on Clip ensembles. It confirms Chen et al. (2024b)'s observation that the bottleneck of attacks is the encoder, and that the encoder's robustness dictates overall robustness. This does not however imply that attacking the encoder is enough, as shown by the weak performance of GeoFA-Encoder.

Failure cases. Figure 3 shows our attacks on random YFCC4K images. Successful attacks stay on landmasses even when FSD is high, implying the model still predicts a realistic position. On the right panel, the model is initially wrong by 5000 km about the clean image: GeoFA-TD/L^2 deviate from the predicted location, while GeoFA-Sample/DPA deviate from the true one. GeoFA-Sample and GeoFA-DPA display similar patterns throughout our experiments. We hypothesize this is because both target the antipode (GeoFA-DPA directly, GeoFA-Sample indirectly by maximizing distance to the true location), whereas GeoFA-TD distorts the trajectory.

6.3 Ablations

Robustness of attacks. Figure 6 reports ASR under JPEG compression, Gaussian blur, varying sampling steps, and CFG scale. Most attacks degrade under blur, compression, and more sampling

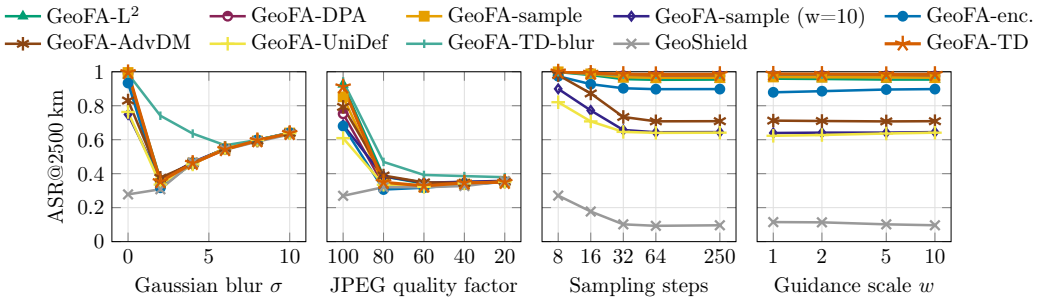


Figure 6: **Robustness of attacks,** $b = 1/255$**, YFCC4k**. ASR under Gaussian blur, JPEG compression and varying sampling steps (250 elsewhere). `GeoFA-TD` and `GeoFA-`L^2 have the lowest decline in success rates across thresholds. `GeoFA`attacks are run with 5 restarts.

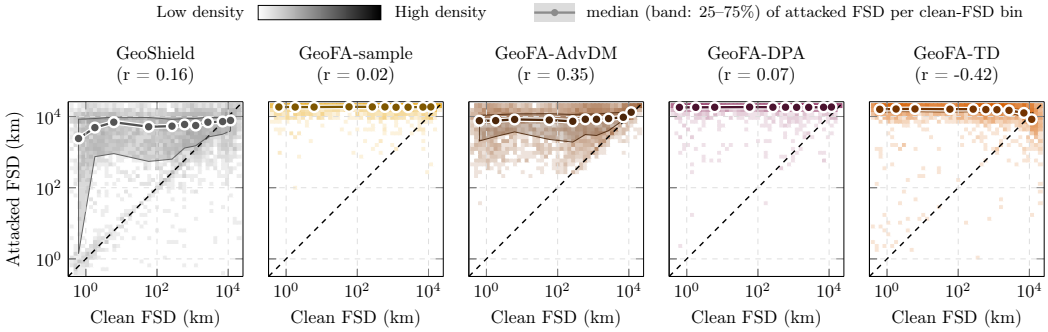


Figure 7: **Clean vs. attacked FSD heatmaps**, $b = 8/255$, YFCC4k. FSD of RFM $\mathcal{S}^2$ on attacked against clean images, with correlation coefficient r and median attacked-FSD band per clean-FSD bin. A positive r indicates an attack performs better when the model is already wrong. `GeoFA-TD` performance is uncorrelated with clean FSD.

steps, with similar trends across methods. `GeoFA`-Encoder, `GeoFA`-AdvDM, `GeoFA`-UniDef, and `GeoFA`-Sample (w=10) drop by about 30 points as sampling steps increase, while others remain more stable. Surprisingly, `GeoFA`-Sample remains robust despite being trained with only 16 steps. Higher CFG slightly improves ASR. Finally, `GeoFA-TD`-blur, trained with Gaussian noise $\sigma \in [0, 4]$ on half the optimization steps, substantially improves robustness across corruptions.

Privacy-critical subset. Nearly half of YFCC4k images are already mislocalized beyond the region level, so average FSD partly reflects images that need no protection. The relevant cases are images whose locations the model infers accurately and confidently. Only `GeoFA-TD` and `GeoFA-`L2 show a strong negative correlation between attacked and clean FSD (Fig. 7): they achieve larger displacements when clean predictions are more accurate. Their effectiveness also increases with localizability (Figs. 12, 13, and 14). Thus, their gains are concentrated on images most exposed to geolocation, consistent with their objectives targeting the predicted velocity that steers sampling.

7 Conclusion

We introduced the task of preventing visual geolocation of images via conditional geographic Riemannian Flow Matching. We proposed a unified framework within which we adapted and evaluated attacks from the image and policy Diffusion literature, and introduced GeoFA-`TD`, an attack tailored to generative models on spherical coordinates. GeoFA-`TD` is grounded in the intuition that attacking is not symmetric to training. It deviates from the fields the sampler actually integrates, targeting angles rather than L^2 norm deviations.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, and the rest of the required disclosure tasks (generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results) are not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, creation of artifacts, draft parts of a research paper, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, and format references

We have reviewed all AI-assisted work. We checked LLM-generated code, LLM-generated suggestions have been tested and backed-up by experimental results. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

REFERENCES

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. Openstreetview-5m: The many roads to global visual geolocation. *CVPR*, 2024.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=AcJrSoArlh.

Yipu Chen, Haotian Xue, and Yongxin Chen. Diffusion policy attacker: Crafting adversarial attacks for diffusion-based policies. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 119614–119637. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3800. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d83fd70a31c64e020844ec80705ba87f-Paper-Conference.pdf.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=SSKZPJCt7B.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23016–23026, June 2025.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–11, 2015.

Jun Guo, Xingyu Zheng, Aishan Liu, Siyuan Liang, Yisong Xiao, Yichao Wu, and Xianglong Liu. Isolation and induction: Training robust deep neural networks against model stealing attacks. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 4178–4189, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612092. URL https://doi.org/10.1145/3581783.3612092.

Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, 2023. URL https://arxiv.org/abs/2302.00275.

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

James Hays and Alexei Efros. Large-scale image geolocalization. *Multimodal Location Estimation of Videos and Images*, pp. 41–62, 10 2015. doi: 10.1007/978-3-319-09861-6_3.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=QB7ot7p6j7.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=9xXjWwAoUF.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

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 Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i42.40877. URL https://doi.org/10.1609/aaai.v40i42.40877.

Weidi Luo, Tianyu Lu, Qiming Zhang, Xiaogeng Liu, Bin Hu, Yue Zhao, Jieyu Zhao, Song Gao, Patrick McDaniel, Zhen Xiang, and Chaowei Xiao. Doxing via the lens: Revealing location-related privacy leakage on multi-modal large reasoning models, 2025. URL https://arxiv.org/abs/2504.19373.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17240–17292, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.957. URL https://aclanthology.org/2024.emnlp-main.957/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Mingwen Shao, Lingzhuang Meng, Xiang Lv, Mengyao Wu, Xinyuan Chen, Qiao Zhang, Chang Liu, Yuanjian Qiao, and Chao Dong. Unidef: Universal defense against unauthorized image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8631–8640, June 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Hao Wang, Zihao Wang, Qiang Xu, Dong Wang, and Kaiju Li. Preserving location privacy for buildings in street view images. *IEEE Transactions on Privacy*, 3:26–38, 2026a. doi: 10.1109/TP.2026.3666140.

Zhangyu Wang, Zeping Liu, Jielu Zhang, Zhongliang Zhou, Qian Cao, Nemin Wu, Lan Mu, Yang Song, Yiqun Xie, Ni Lao, and Gengchen Mai. Locdiff: Identifying locations on earth by diffusing in the hilbert space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=ghybX0Qlls.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2f4fe03d77724a7217006e5d16728874-Paper.pdf.

Qiang Yi and Lianlei Shan. Geolocsft: Efficient visual geolocation via supervised fine-tuning of multimodal foundation models. *ArXiv*, abs/2506.01277, 2025. URL https://api.semanticscholar.org/CorpusID:279075895.

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 Computer Vision and Pattern Recognition Conference*, pp. 19900–19909, 2025a.

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 Computer Vision and Pattern Recognition Conference*, pp. 19900–19909, 2025b.

Zheyuan Zhang, Runze Li, Tasnim Kabir, and Jordan Boyd-Graber. Navig: Natural language-guided analysis with vision language models for image geo-localization, 2025c. URL https://arxiv.org/abs/2502.14638.

Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=agHddsQhsL.

A PROOFS

Proof of Proposition 1. Write $v = \psi_c + r$ with $\psi_c = \psi(x_t \mid c)$, $\psi_\delta = \psi(x_t \mid c+\delta)$, and expand pointwise in the tangent space at x_t:

$$\|\psi_\delta - v\|_{x_t}^2 = \|(\psi_\delta - \psi_c) - r\|_{x_t}^2 = \|\psi_\delta - \psi_c\|_{x_t}^2 - 2\langle \psi_\delta - \psi_c,\, r\rangle_{x_t} + \|r\|_{x_t}^2. \tag{14}$$

We then take expectations over (ϵ, t), yielding the final result.

Remark on the cross term. For a fixed image with known x_0, r is nonzero pointwise: it encodes both the model's fit error and the multimodality of its location posterior. Hence, the AdvDM gradient carries a δ-dependent bias towards the data's velocity, a quantity the sampler never integrates.

B DIFFUSION AND EUCLIDEAN FLOW MATCHING

We formulate the paper in the RFM setting of Section 3. This appendix gives the two variants of Dufour et al. (2025) that we also evaluate, and the substitution under which every attack of Sections 4–5 applies to them unchanged.

Geographic Diffusion. With $(\epsilon, t) \sim \mathcal{N}(0, I) \otimes \mathcal{U}[0,1]$ and the schedule κ of Section 3, coordinates are noised as $x_t = \sqrt{1-\kappa(t)}\, x_0 + \sqrt{\kappa(t)}\, \epsilon$, the network minimises the denoising loss $\mathcal{L}^{DM} = \mathbb{E}_{x_0,c,\epsilon,t}\big[\|\psi(x_t \mid c) - \epsilon\|^2\big]$, and locations are inferred by Denoising Diffusion Implicit Model (DDIM) sampling (Song et al., 2021), which iterates

$$\hat{x}_t = \frac{1}{\sqrt{1-\kappa(t)}} \Big(x_t - \sqrt{\kappa(t)}\, \psi(x_t \mid c) \Big) \tag{15}$$

from $t = 1$ to $t = 0$. The factorisation $\psi(x_t \mid c) = \phi(x_t, \mathcal{E}(c))$ and Classifier-Free Guidance equation 17 are unchanged.

Euclidean Flow Matching. Euclidean FM replaces the geodesic interpolation equation 1 by the linear one, $x_t = (1-\kappa(t))\, x_0 + \kappa(t)\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, so that the target field is the constant $v(x_t) = \dot{\kappa}(t)\, (\epsilon - x_0)$ and the exponential map of equation 3 degenerates into addition, $x_{t-dt} = x_t - dt\, \psi_w(x_t \mid c)$. Coordinates are then predicted in $\mathbb{R}^3$ (or in latitude–longitude) and projected back onto the sphere, which is precisely the geometry that RFM avoids.

Correspondence. Every attack in Section 4 is written in terms of three objects: the encoder $\mathcal{E}$, the field $\psi(x_t \mid c)$ that the sampler integrates, and the training target. Table 3 gives the substitution; the encoder is common to all three settings, so encoder attacks (Section **??**) are unaffected.

Table 3: Substitution carrying the GeoFA framework from RFM to the other two settings.

	RFM on $\mathcal{S}^2$	**Diffusion**
integrated field	velocity $\psi(x_t \mid c)$	noise prediction $\psi(x_t \mid c)$
training target	$v(x_t)$	ϵ
metric	$\langle \cdot, \cdot \rangle_{x_t}$	Euclidean $\langle \cdot, \cdot \rangle$
sampling step	$\exp_{x_t}(-dt\, \cdot)$	Eq. equation 15

B.1 LOCALIZABILITY

Dufour et al. (2025) generative modeling of an image's location allows to compute distributional metrics of an image. We compute localizability as the entropy of the predicted location

$$\text{Localizability}(c) = \int_{\mathcal{S}^2} p(y \mid c) \log_2 p(y \mid c)\, dy. \tag{16}$$

This metric allows quantifying, when the model is a good predictor of an image's location, whether an image can be *uniquely* localized (low entropy), or whether it could correspond to multiple different locations.

B.2 Classifier Free Guidance

Riemannian Flow Matching, Flow Matching and Diffusion enable Classifier free guidance to sharpen estimates at sampling time, via

$$\psi_w(x_t \mid c) = (1 + w)\,\psi(x_t \mid c) - w\,\psi(x_t \mid \varnothing), \tag{17}$$

where $\psi(x_t \mid \varnothing)$ is the unconditional prediction and w the guidance scale. This sampling method is used throughout our exeperiments. By default, we set $w = 10$.

B.3 LocDiff: extending to Spherical Harmonics

LocDiff (Wang et al., 2026b) represents image's geographic locations as dirac functions on the sphere. Their spherical harmonic coefficients then provide a sparse, linearly decodable encoding space which captures multi-scale information unlike raw geographic coordinates. LocDiffusion is trained using a Kullback-Leibler (KL) loss between the distributions on the sphere induced by the true and predicted spherical harmonic coefficients.

C Details on GeoFA

C.1 Extended review of attacks against diffusion models

Recent advances in image generation, in particular Diffusion and Flow Matching Models (Ho et al., 2020; Song et al., 2021), have sparked concerns regarding the privacy of training data as well as potential malicious use to generate harmful content or infringe copyright (Liang et al., 2023; Salman et al., 2023). Adversarial attacks on Diffusion Models (DM) address this concern. Salman et al. (2023) propose attacks that perturb images to prevent AI-powered editing, highlighting the difficulty of the task due to the iterative nature of Diffusion. They craft attacks on the image encoder and approximate a **direct sampling attack** by backpropagating only through a few diffusion steps to avoid memory errors. Kang et al. (2023) propose DiffAttack, which optimizes a deviated-reconstruction loss at intermediate Diffusion steps, again backpropagating through a fraction of the steps. Recent efforts have focused on the denoising training loss. AdvDM (Liang et al., 2023) maximizes that loss to prevent a DM from extracting features of an image. ACE (Zheng et al., 2025) improves it by targeting the score of a carefully designed target image. UniDef (Shao et al., 2026) refines this approach by targeting the Diffusion process at the global denoising level, and estimates the global denoising jacobian of the loss via finite differences which improves transferability. Diffusion Policy Attack (DPA) (Chen et al., 2024b) extends this **training loss centric approach** to Diffusion policies for robotic arms, again by designing a *suitable* target. To the best of our knowledge, no attack has yet been designed against geographic generative models. All of the above are formulated for Euclidean, image-space Diffusion, so transposing them to a compact Riemannian manifold is itself part of our contribution. DPA (Chen et al., 2024b) is the closest work to ours, as it attacks diffused coordinates conditioned on images. It does not, however, adapt existing attacks from the image Diffusion literature to its setting.

C.2 Encoder attacks

The simplest form of attack against image conditioned Riemannian Flow Matching targets the image encoder $\mathcal{E}$. The only channel conveying visual information is $\mathcal{E}$. Encoder attacks hence seem both the most direct and generalizable way to perturb an image. They can be used to attack multiple different models sharing an encoder: GeoShield (Liu et al., 2026) attacks an ensemble of Clip based image-text encoders. Following the geo-privacy literature (Liu et al., 2026) and unlike the image Diffusion baseline (Salman et al., 2023), we mainly consider untargeted attacks here as they are deemed more fit to distorting the predicted location.

A targeted alternative aims at a latent target $\mathcal{E}^T$ under budget b:

$$\mathcal{L}_{\text{adv}}^{\text{GeoFA-t-Encoder}}(\delta) = \|\mathcal{E}(c + \delta) - \mathcal{E}^T\|^2 \qquad (\min) \quad . \tag{18}$$

C.3 SAMPLING-CENTRIC ATTACKS

Such attacks directly optimize the final objective (Eq. 4). This approach is closest to adversarial attacks against Classification Models (Goodfellow et al., 2015; Madry et al., 2018), where one can directly optimize the end goal by maximizing classification error. These attack are usually difficult and costly to train directly, which can be mitigated by backpropagating only through a few sampling steps. Since Dufour et al. reach optimal performance at only 16 timesteps with 36M parameters model, we can afford to train GeoFA-Sampling through the full 16-step sampler. We stress that this is not viable as model size increases.

C.4 TRAINING LOSS CENTRIC ATTACKS – GEOFA-ADVDM AND GEOFA-UNIDEF

This surrogate is inherited from the recent SOTA literature on adversarial attacks against Diffusion Models: for Diffusion, the likelihood-weighted denoising loss upper-bounds the KL divergence between the model and data distributions (Song et al., 2021), which motivates maximizing it as a proxy for distorting the model's predicted distribution (Liang et al., 2023; Shao et al., 2026). We adopt it in the RFM setting. GeoFA-AdvDM (equation 7) adapts AdvDM (Liang et al., 2023) to the RFM framework. GeoFA-UniDef adapts UniDef (Shao et al., 2026), which maximizes the same quantity (up to a weighting factor), but replaces the backward pass through the denoiser by a finite-difference Jacobian estimate, so as not to overfit a single denoiser. GeoFA-UniDef is therefore distinguished from GeoFA-AdvDM by its gradient estimator rather than by its objective $\mathcal{L}_{\text{adv}}^{\text{GeoFA-UniDef}}$. UniDef is however designed for attacks on the diffused modality (images in their case, coordinates in ours), not the conditioning modality as in our case. We hence do not expect its direct transposition to geographic RFM to yield strong performance, and discuss it in Appendix C.6.

C.5 TARGETED ATTACKS – GEOFA-ACE AND GEOFA-DPA

Another option is to set a target either in pixel or coordinate space. An image space target c' corresponds to ACE (Zheng et al., 2025) and yields GeoFA-ACE, which minimizes the distance between the predicted velocities of the perturbed and the target image, jointly with an encoder term weighted by α. A coordinate space target x_0', with $x_t' = x_0' + \epsilon$, corresponds to DPA (Chen et al., 2024b) (in their case for a target policy) and yields GeoFA-DPA. It is closer to our setting, since we diffuse coordinates, not images, and the natural choice for x_0' in geographic FM is the *antipod* of x_0, that is the furthest point on earth. Both methods follow the intuition that a well chosen target allows to train an attack more efficiently than naively maximizing training loss, since one deliberately targets a *bad* example.

Loss-centric attacks are both theoretically grounded and cheap to train via Monte Carlo sampling over (ϵ, t). Targeted attacks are a first deviation from them, while keeping their loss form.

C.6 DETAILS ON GEOFA-UNIDEF

UniDef's (Shao et al., 2026) Consistent Distribution Deviation objective integrates the denoising loss over the whole trajectory with a schedule-induced weight $w(t)$, evaluated on a single noise sample per trajectory. Since $t \sim \mathcal{U}[0,1]$, that time integral is the expectation over t; taking $w \equiv 1$ as we do for AdvDM, and restoring the expectation over ϵ that Monte Carlo estimation performs in any case, it coincides with equation 7. In our setting the two attacks are therefore distinguished by their gradient estimator rather than by their objective. This equivalence is specific to our setting. This is due to the fact that UniDef is designed for attacking the modality that is diffused, that is images in their case, coordinates in ours. We do not attack the coordinates directly, making a straightforward transposition of UniDef impossible. We therefore analyse GeoFA-UniDef results carefully, and point to future work for designing specifically transferable geographic flow attacks targeted at the global denoising level. We however note that our attacks transfer well across generative architectures (diffusion,riemannian flow matching) despite the absence of an explicit transferability mechanism (see Figure 5).

We propose a simple interpretation of UniDef in our framework. We note $e_\delta = \mathcal{E}(c+\delta)$, $\Delta_t(\delta) = \phi(x_t, e_\delta) - v_t$, with v_t the true training target of the path through the ground-truth location, $\mathcal{L}^{\text{GeoFA-UniDef}} = \mathbb{E}_{\epsilon,t}\|\Delta_t(\delta)\|_2^2$, and fix $z = \mathcal{E}(c)/\|\mathcal{E}(c)\|_2$ following Shao et al. (2026). A

Hutchinson-style projection on z and a symmetric finite difference of step h give, up to a positive constant,

$$\widehat{\nabla}_\delta \mathcal{L}^{\text{GeoFA-UniDef}} = \mathbb{E}_{\epsilon,t}\left[\left(\frac{\partial e_\delta}{\partial \delta}\right)^\top z \left\langle \frac{\phi(x_t, e_\delta + hz) - \phi(x_t, e_\delta - hz)}{2h},\ \Delta_t(\delta)\right\rangle\right] \quad , \tag{19}$$

estimated by Monte Carlo over (ϵ, t) and used in the sign update of PGD. The denoiser ϕ is only ever evaluated forward; the gradient reaches δ through the encoder alone.

C.7 Extending GeoFA-TD to Diffusion and other topologies

We formulate GeoFA-TD for RFM since it performs best (Dufour et al., 2025) and is the most specific setup. Our attack can however be easily applied against a DM by replacing flows by the predicted noise given that these vectors dictate the sampling trajectory in equation 3 and equation 15. Indeed, the angle GeoFA-TD maximizes is that of the vector the sampler actually steps along, which in Diffusion is the predicted noise, so the objective equation 13 is unchanged up to the substitution of Appendix B. Figure 5 confirms this empirically: perturbations trained against RFM $\mathcal{S}^2$ retain their success rate against the Diffusion and Euclidean Flow Matching variants of Dufour et al. (2025). GeoFA-TD applies to LocDiffusion (Diffusion in spherical harmonics space) (Wang et al., 2026b) in the same way. The arguments in Section 5.2 carry over wherever the sampled domain is compact.

D Experimental Details

Attacked models and data. The attacked model is, for each dataset, the public Riemannian flow-matching checkpoint of Dufour et al. (2025) on $\mathcal{S}^2$ (PLONK_YFCC, PLONK_OSV_5M). The image encoder $\mathcal{E}$ is shared with the diffusion and Euclidean flow-matching checkpoints used in the transfer experiment: DINOv2 ViT-L/14 (Oquab et al., 2024) on YFCC4k, StreetCLIP ViT-L/14 (Haas et al., 2023) on OSV-5M, both taking a 336×336 input obtained by a 1:1 centre crop, bicubic resize and the encoder's own channel normalisation. We evaluate $4\,000$ images per dataset: YFCC4k in full, and for OSV-5M a fixed, prefix-stable subset of the test split (seeded shuffle, seed 42).

Attack hyperparameters. Attacks are trained per image over 60 PGD steps, with batch-size 256 (where applicable) and learning rate 10^{-3}. These hyperparameters ensure comparable training length and memory usage across considered baselines despite their design differences. We train with budgets $b \in \{1/255,\ 4/255,\ 8/255\}$. This choice aims at demonstrating the high performance of our attack even at very low perturbation budgets, considered as highly stealthy in the literature (Kang et al., 2023; Croce et al., 2021). Budget $8/255$ allows direct comparison with GeoShield. Every trainable attack shares the same outer loop: the perturbation δ lives on the $[0, 1]$ pixel scale of the model-input square and is updated by sign-SGD, $\delta \leftarrow \mathrm{clip}(\delta - \eta\,\mathrm{sign}(\nabla_\delta \mathcal{L}), \pm b)$ with $\eta = 10^{-3}$, from a small random start $\delta_0 \sim \mathcal{U}(-10^{-4}, 10^{-4})$. Expectations over (x_t, ϵ, t) are estimated with 256 Monte-Carlo samples per step, $t \sim \mathcal{U}[0, 1]$ and $\gamma(t)$ the model's own noise schedule. The forward process is anchored at the image's *true* location x_0, which the defender knows by construction.

For GeoFA, we evaluate attacks with either one restart, or we train with five random restarts and select the best with respect to FSD. The latter allows to to make attacks more reliable. Given our attacks are cheaper to train than GeoShield (see Table **??**), such restarts allow matching compute time with it. We use the five restart perturbations in our robustness and transferability ablations. All other results are for one restart.

Evaluation protocol. The protected image $\mathrm{clamp}(c+\delta)$ is quantised to 8 bits and fed to the model through the very same preprocessing as a clean image, so that nothing of the perturbation survives outside the pixel domain. For every (image, attack, budget) we draw 128 locations with 250 sampling steps and guidance $w = 10$ from noise shared with the clean run; FSD is the mean haversine distance between these predictions and the ground truth, and $\mathrm{ASR}_{>d}$ the fraction of images whose FSD exceeds $d \in \{200, 750, 2500\}$ km. Clean rows apply the same protocol to the unprotected image.

Attack cost. We measure the wall-clock time of each optimisation step and the peak GPU memory it allocates, model weights included. Every attack runs through the same pipeline and uses the same

hyperparameters as in our main evaluation, and GeoShield uses its official implementation. Each attack runs in its own fresh process, alone on a single NVIDIA RTX A5000. We attack 10 YFCC4k images at $b = 8/255$, after one warm-up image that we discard. We report the median over all steps: 60 per image for our attacks and 100 for GeoShield. We ran the whole benchmark twice, the second time in reverse order, and the two runs agree within 0.5%. Cost does not depend on b.

Attack training with restarts. We enable multiple random restarts for `GeoFA`attacks to improve performance. In our main experiments, we train with either 1 or 5 random initializations of the perturbation, and report results accordingly. 5 starts approximately aligns training costs on GeoShield (see speed/VRAM in Table 2). We mainly report single initalization results, except for robustness and transferability ablations.

GeoShield. GeoShield (Liu et al., 2026) is run with the authors' released implementation and its `ensemble_3models` configuration: 100 FGSM-style sign steps of size $1/255$ at the budgets $\{1, 4, 8\}/255$, a CLIP ensemble (ViT-B/16, ViT-B/32, LAION ViT-G/14), global and random-crop (scale $[0.5, 0.9]$) feature-similarity losses and a geo-text term. We note that the released code ships a placeholder image description and no detector regions, so our GeoShield is its CLIP-ensemble feature attack with a constant geo-text term rather than the full VLM/detector pipeline of the original paper.

Ablations. All ablations reuse the selected perturbations; nothing is retrained. *Sampling steps:* the protected image is re-evaluated with $\{8, 16, 32, 64, 250\}$ steps at $w = 10$. *Guidance:* the same with $w \in \{1, 2, 5, 10\}$. *Robustness:* the protected image is JPEG-compressed at quality $\{20, 40, 60, 80, 100\}$ (100 = no compression) or blurred with a Gaussian kernel of $\sigma \in \{0, 2, 4, 6, 8, 10\}$ pixels before the encoder, then evaluated at 250 steps; this sweep covers every attack of the run. *Transfer:* the same protected image is evaluated against the diffusion checkpoints of Dufour et al. (2025), which share the encoder, at 250 steps and $w = 10$.

E Additional Experimental Results

In this section, we provide more comprehensive ablations and comments about our results.

Table 4 reports ASR on OSV-5M and YFCC4k *with 5 restarts*. Our best `GeoFA` attacks (`GeoFA-TD`/DPA/Sample) reach 100% with almost all budgets for the two datasets. All our `GeoFA-*` attacks outperforms the the state-of-the-art GeoShield at all budgets and on both datasets when evaluated with the same number of restarts.

E.1 YFCC4K ablations

E.1.1 Transferability to diffusion

Figure 8 displays ASR at the 2500 km threshold for all our trained attacks at varing budgets, confirming our observations in Section 6.3.

E.1.2 Attacked vs clean FSD

Figures 9, 10, 11 display clean vs attacked FSD heatmaps for all attacks across budgets $1 - 8/255$. We observe that `GeoFA-TD`/L^2 are the only attacks with a strong negative correlation, as discussed in 6.3. This remains true across budgets and when including all attacks.

E.1.3 FSD with respect to localizability

Figure 12, 13, 14 display FSD distribution with respect to localizability for all attacks across budgets $1 - 8/255$. As observed in 6.3, `GeoFA-TD` is the only attack whose FSD improves slightly as localizability increases, while others stagnate or decline.

Table 4: Attack success rate (ASR, %, ↑) on YFCC4k and OSV-5M for three budgets, and attack cost. $\text{ASR}_{>d}$ is the fraction of images whose predicted location is pushed *more than* d km from the ground truth. All Ours-GeoFAattacks select the best of 5 restarts by FSD to the true location. Time is the wall-clock time of one optimisation step and VRAM the peak GPU memory allocated during a step, measured on one NVIDIA RTX A5000. GeoShield/GeoFAattacks run 100/60 steps per image respectively. Every Ours-GeoFA-$*$ row instantiates our framework. Best per column in **bold**, second best underlined, third best in *italics*.

		$b = 1/255$		$b = 4/255$		$b = 8/255$		Time↓	VRAM↓
	Attack	>750	>2500	>750	>2500	>750	>2500	(ms)	(GiB)
YFCC4k	GeoShield (2026)	44.10	27.82	65.15	48.08	81.53	66.07	651	7.80
	Ours-GeoFA-$\mathcal{E}$ (2023)	98.10	93.27	99.98	99.02	**100.00**	99.20	**124**	**1.85**
	Ours-GeoFA-Sample (2023)	**99.83**	99.55	**100.00**	*99.92*	**100.00**	99.92	232	2.93
	Ours-GeoFA-AdvDM (2023)	94.50	83.03	97.25	89.28	97.55	89.88	133	*1.91*
	Ours-GeoFA-ACE (2025)	84.35	63.50	93.67	69.80	94.23	70.58	193	1.93
	Ours-GeoFA-DPA (2024b)	99.55	*99.17*	**100.00**	**100.00**	**100.00**	**100.00**	133	1.90
	Ours-GeoFA-UniDef (2026)	87.48	76.33	92.53	86.12	92.83	86.55	*135*	**1.85**
	Ours-GeoFA-L^2	*99.25*	98.02	*99.45*	98.55	99.50	98.52	193	1.93
	Ours-GeoFA-TD	**99.83**	**99.58**	99.98	99.95	99.98	99.92	193	1.93
	Clean	42.00	25.73	*budget-independent*				–	–
OSV-5M	GeoShield (2026)	55.43	34.88	98.47	92.42	99.75	96.78	651	7.80
	Ours-GeoFA-$\mathcal{E}$ (2023)	99.95	*98.08*	**100.00**	**100.00**	**100.00**	**100.00**	**124**	**1.85**
	Ours-GeoFA-Sample (2023)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	232	2.93
	Ours-GeoFA-AdvDM (2023)	99.12	89.30	99.78	93.55	99.75	93.67	133	*1.91*
	Ours-GeoFA-ACE (2025)	*99.55*	95.47	*99.58*	94.53	*99.38*	94.08	193	1.93
	Ours-GeoFA-DPA (2024b)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	133	1.90
	Ours-GeoFA-UniDef (2026)	95.95	89.25	98.35	95.08	98.50	95.30	*135*	**1.85**
	Ours-GeoFA-L^2	**100.00**	99.75	**100.00**	*99.83*	**100.00**	*99.85*	193	1.93
	Ours-GeoFA-TD	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	193	1.93
	Clean	23.80	n/a	*budget-independent*				–	–

E.1.4 Robustness ablations at $b = 8/255$

Figure 15 extends Figure 6 at budgets $8/255$. Patterns observed and discussed in Section 6.3 persist here.

E.1.5 Additional observations

Attack stealth. We observe that the best attacks maintain high success rates even at the low perturbation budget of $1/255$, despite an anticipated decreasing performance trend as budget goes down. GeoShield suffers most from the low budget, and has almost no effect at $b = 1/255$.

E.2 OSV ablations

E.2.1 Final Step Deviation

Figure 16 shows FSD for all ou attacks on OSV-5M.

E.2.2 Transferability to diffusion

Figures 17 and 18 display ASR at the 2500 km threshold for all our trained attacks at varing budgets, confirming our observations in Section 6.3.

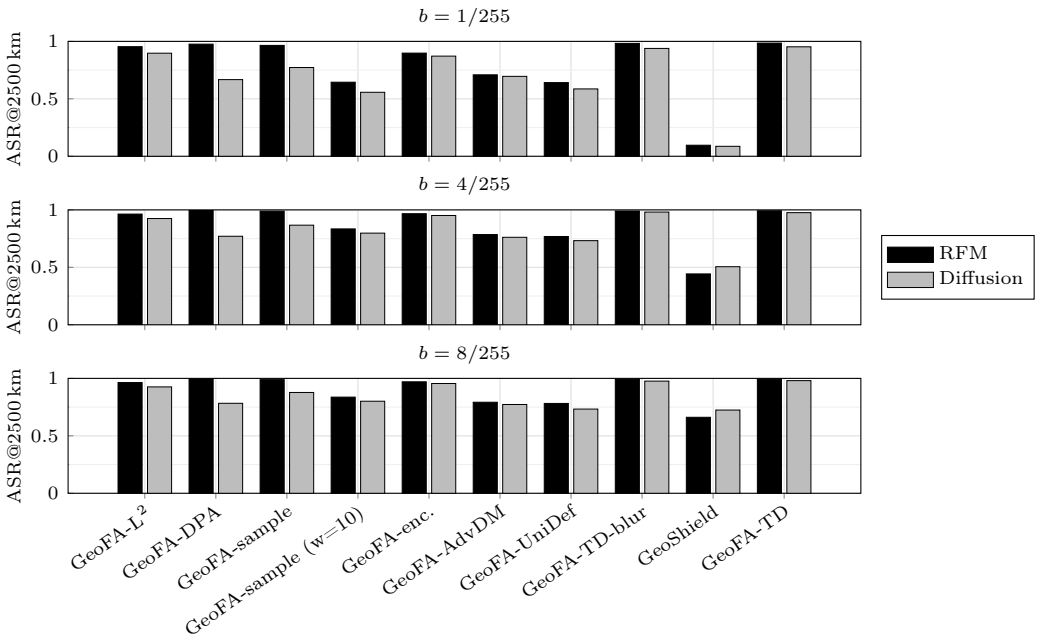


Figure 8: **Transferability to model variants of GeoFA attacks for 3 attack budgets, YFCC4k**. Best $k = 5$ start selection.

E.2.3 Attacked vs clean FSD

Figures 19, 20, 21 display clean vs attacked FSD heatmaps for all attacks across budgets $1 - 8/255$. We observe that `GeoFA-TD`/L^2 are the only attacks with a strong negative correlation, as discussed in 6.3. This remains true across budgets and when including all attacks.

E.2.4 Robustness ablations at $b = 8/255$

Figure 23 displays robustness of our attacks to blur, compression, increasing sampling steps and increasing CFG scale.

E.2.5 Qualitative comparison

Figure 24 provides a qualitative comparison of GeoShield against `GeoFA-TD` on a random OSV-5M picture for $b = 8/255$. We choose $b = 8/255$ since GeoShield's performance drops substantially below that level. We observe that GeoShield creates very visible textual artifacts, which may be linked to the text-image encoding space it attacks.

F Limitations

Our framework and evaluation method presents several limitations. First of all, we consider white box attacks, i.e. setups in which we know the model architecture and its weights. In the contrary, Liu et al. (2026) propose a black box attack which is transferable across VLMs. In that sense, our attacks remain limited in scope. We however observe a strong transferability of GeoFA-TD to the Diffusion counterpart of RFM $\mathcal{S}^2$, that it was not trained against. Specifically targeting transferability as UniDef does hence seems less justified here. Following the literature on attacks against VLMs (Li et al., 2025; Liu et al., 2026), targeting ensembles of images encoders or model architectures could further improve transfer. We however demonstrate the limited effectiveness of encoder-only attacks, especially on images of the YFCC4K dataset.

Second, we have used the datasets on which Dufour et al. (2025) trained their Diffusion Models, limiting comparability with other spans of the literature, and We could not evaluate our attacks

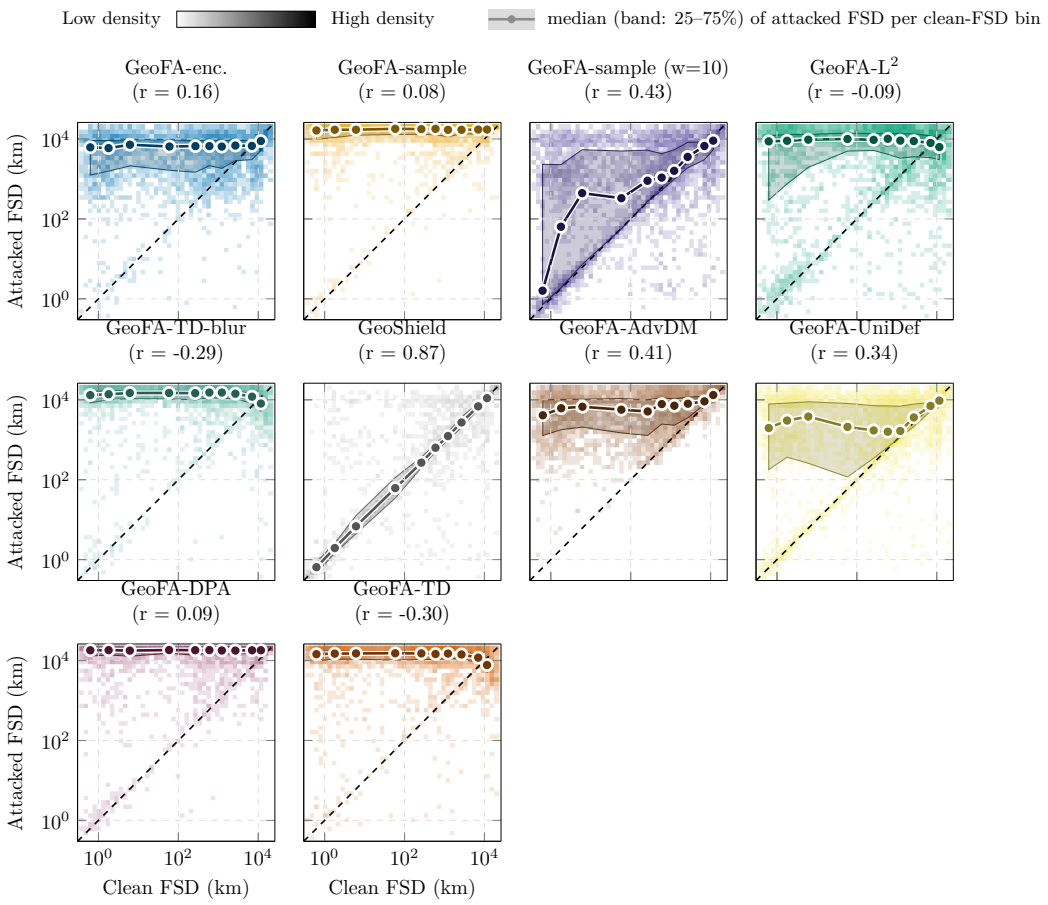


Figure 9: **Clean vs attacked FSD heatmaps for** $b = 1/255$ on YFCC4K. Single ($k = 1$) start selection.

against the most recent Diffusion based model for visual geolocation LocDiff (Wang et al., 2026b), since no open source implementation or model weights are available.

Finally, one of our contributions consists in adapting existing attacks such as AdvDM (Liang et al., 2023), ACE (Zheng et al., 2025) or DPA (Chen et al., 2024b) to geographic RFM. These attacks were initially designed against Image Diffusion. This makes their adaptation to our framework non trivial, and entails choosing the objective and parameters that fit the concept of the attack most closely. For instance, for AdvDM, we maximize the RFM training objective, which is mathematically different but conceptually equivalent to the initial objective of the attack against Image Diffusion. For DPA, we target the velocity of the antipod, which is different from targeting the score of an adversarial robotic action.

G Generation Examples

In this section, we report examples of our attack's effect on images. Figures 25 and 26 display such examples at different budgets.

Figure 25 shows a failure case of `GeoFA-TD`. Here, RFM $\mathcal{S}^2$ initially mislocalizes the clean image by 18000 km. `GeoFA`-DPA and `GeoFA`-Sample stick to the clean (wrong) prediction, while `GeoFA-TD` deviates from it by design, and ends up closer to the real position.

Figure 26 displays a more standard case. `GeoFA-TD`/Sample/Encoder distort predictions the most, followed by `GeoFA`-DPA. All other attacks remain close to the real location, which is accurately predicted by the clean model.

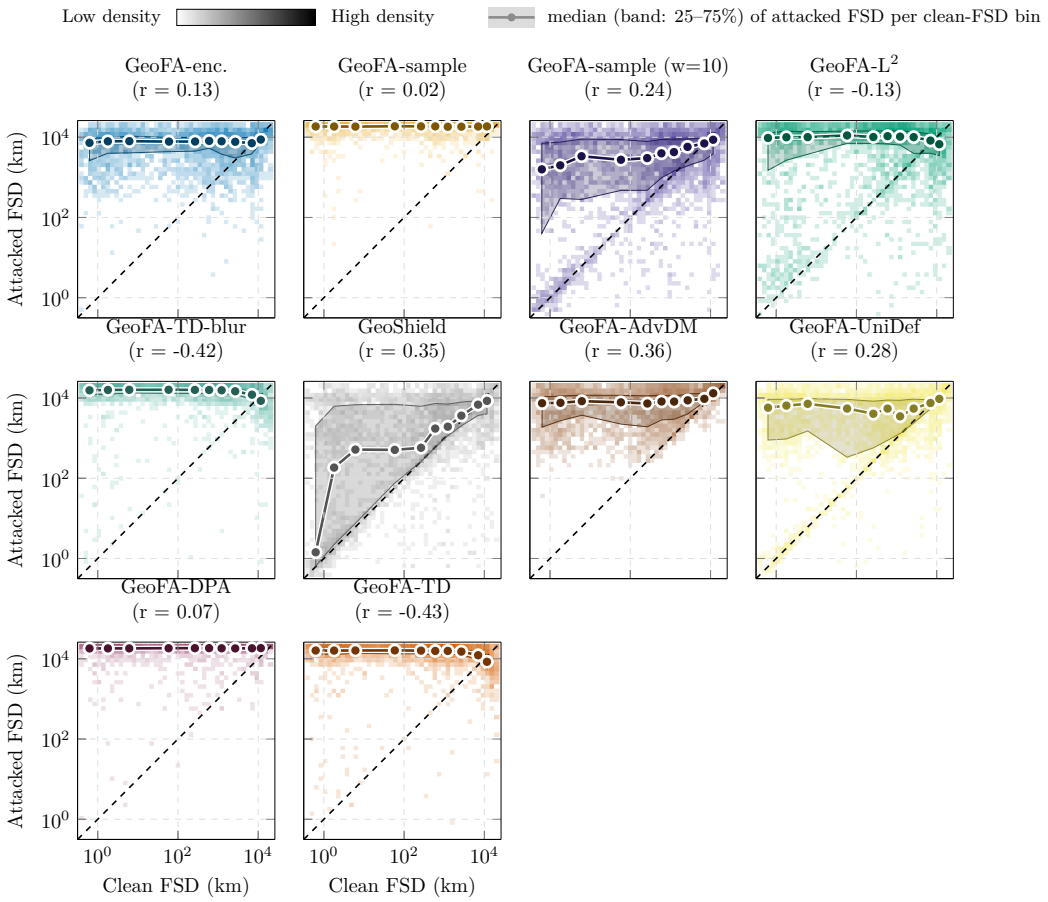


Figure 10: **Clean vs attacked FSD heatmaps for** $b = 4/255$ on YFCC4K. Single ($k = 1$) start selection.

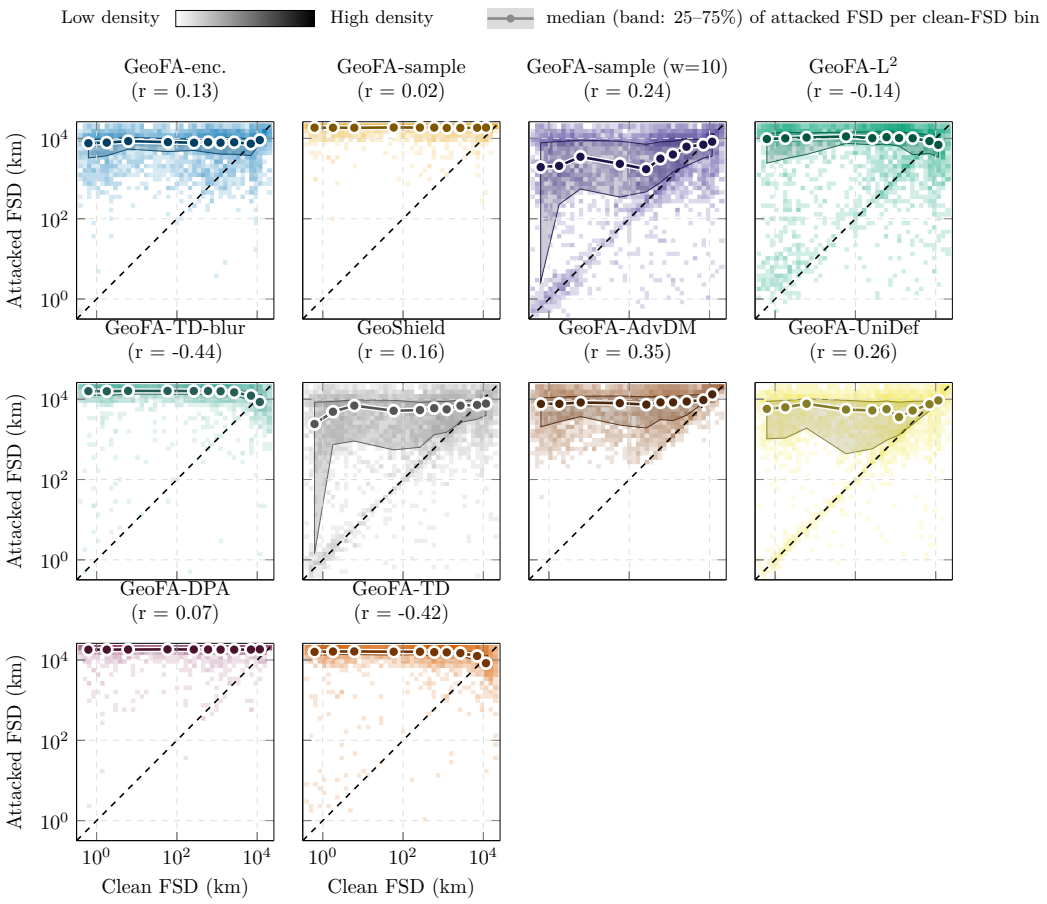


Figure 11: **Clean vs attacked FSD heatmaps for** $b = 8/255$ on YFCC4K. Single ($k = 1$) start selection.

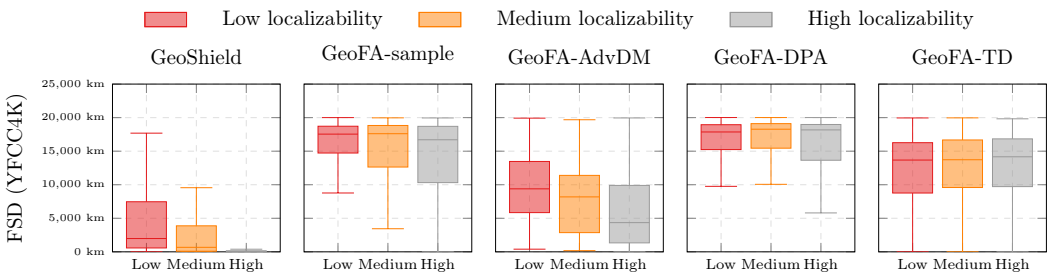


Figure 12: **Distributions of FSD** ($b = 1/255$**, single** ($k = 1$) **start selection.)** of several attacks on YFCC4K according to **localizabillity**. Low, medium and high localizability thresholds are designed using 33% and 66% quantiles of localizabiity. FSD declines with localizability, indicating attacking highly localizable images are harder to attack. GeoFA-TD and GeoFA-L^2 display the lowest decline in FSD due to this phenomenon.

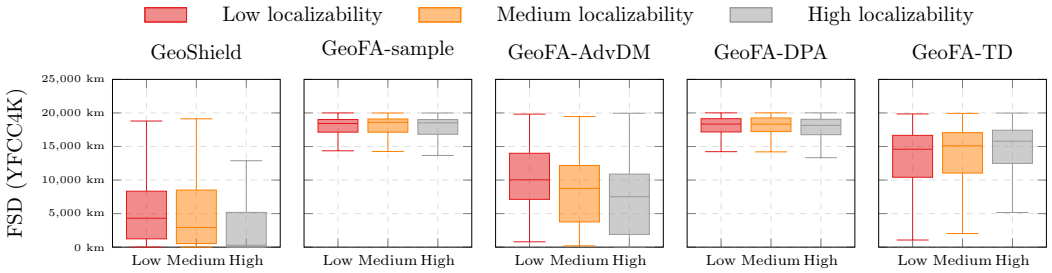


Figure 13: **Distributions of FSD ($b = 4/255$, single ($k = 1$) start selection.)** of several attacks on YFCC4k according to **localizabillity**. Low, medium and high localizability thresholds are designed using 33% and 66% quantiles of localizabiity. FSD declines with localizability, indicating attacking highly localizable images are harder to attack. GeoFA-TD and GeoFA-L^2 display the lowest decline in FSD due to this phenomenon.

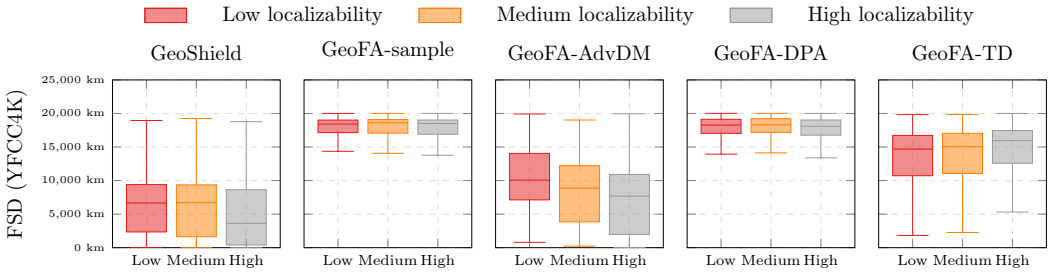


Figure 14: **Distributions of FSD ($b = 8/255$, single ($k = 1$) start selection.)** of several attacks on YFCC4k according to **localizabillity**. Low, medium and high localizability thresholds are designed using 33% and 66% quantiles of localizabiity. FSD declines with localizability, indicating attacking highly localizable images are harder to attack. GeoFA-TD and GeoFA-L^2 display the lowest decline in FSD due to this phenomenon.

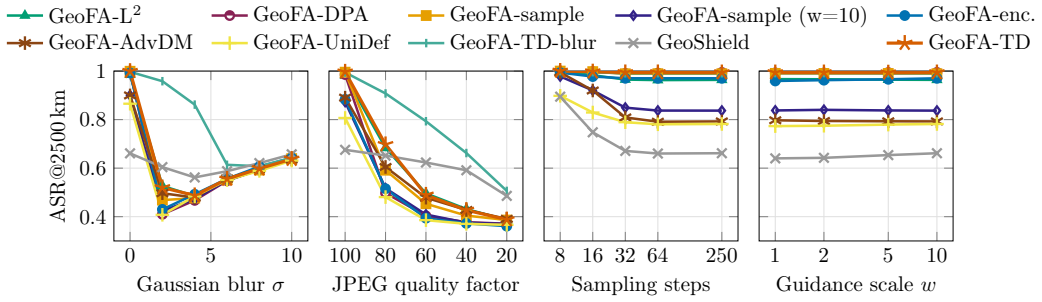


Figure 15: **Robustness to Gaussian blur, JPEG compression, sampling steps and guidance scale,** $b = 8/255$, YFCC4k. Best $k = 5$ start selection.

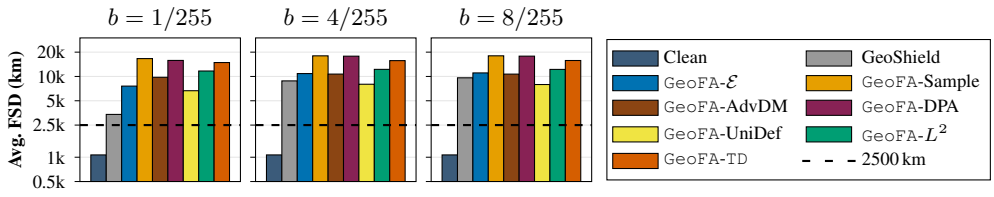


Figure 16: FSD on OSV-5M (log scale), single restart; dashed line at 2500 km.

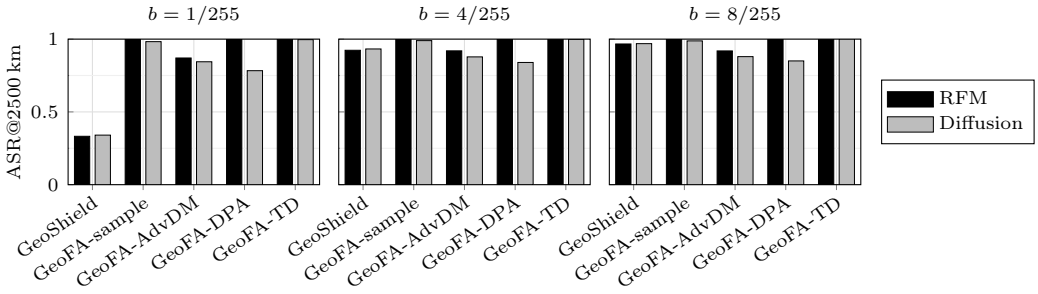


Figure 17: **Transferability to model variants of GeoFA attacks for 3 attack budgets. Best $k = 5$ start selection. OSV-5M**

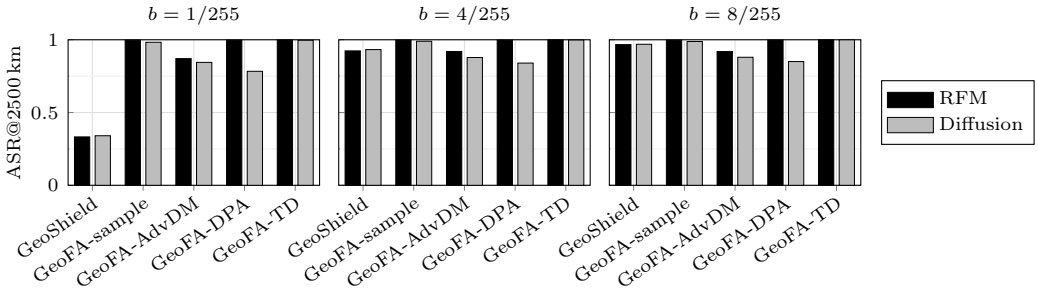


Figure 18: **Transferability to model variants of GeoFA attacks for 3 attack budgets. Best $k = 5$ start selection. OSV-5M**

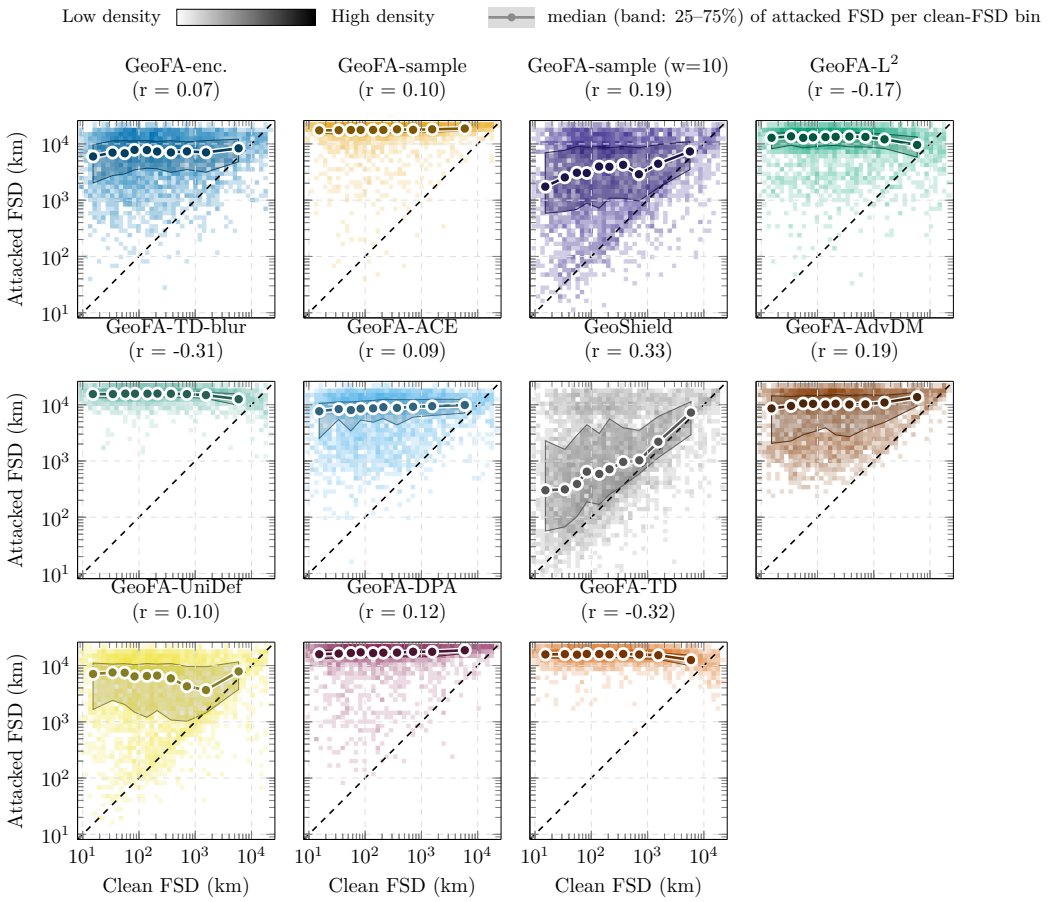


Figure 19: **Clean vs attacked FSD heatmaps for** $b = 1/255$ on OSV-5M. Single ($k = 1$) start selection.

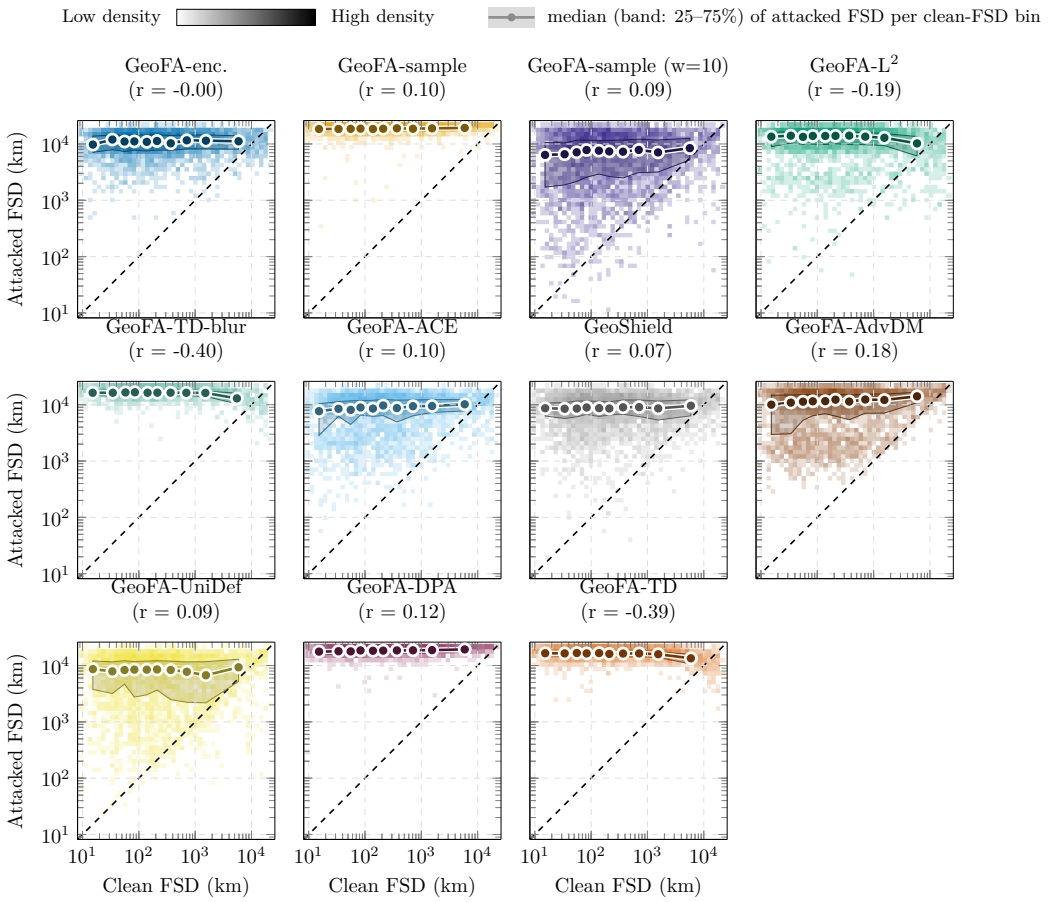


Figure 20: **Clean vs attacked FSD heatmaps for** $b = 4/255$ on OSV-5M. Single ($k = 1$) start selection.

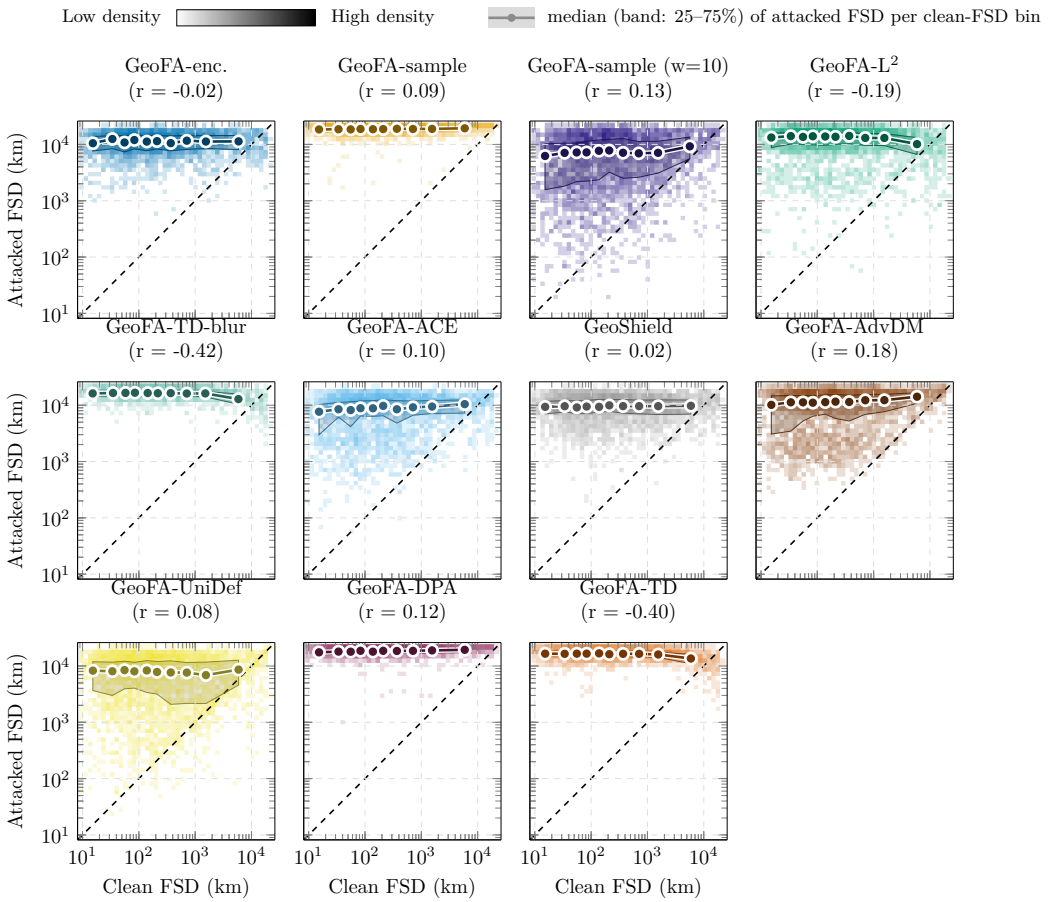


Figure 21: **Clean vs attacked FSD heatmaps for** $b = 8/255$ on OSV-5M. Single ($k = 1$) start selection.

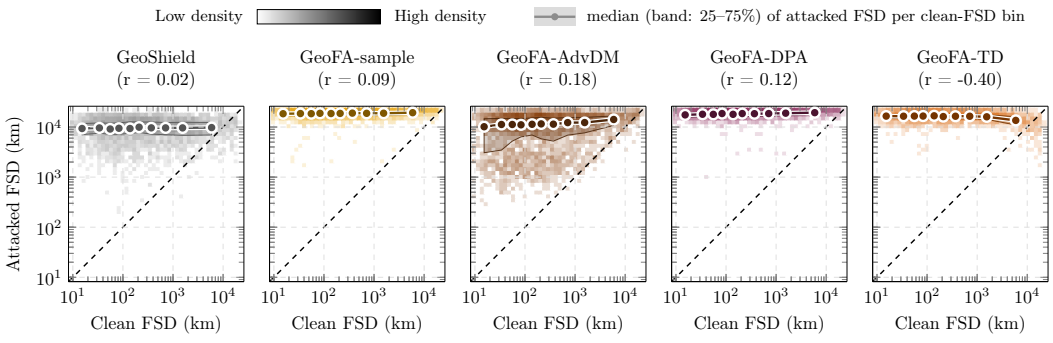


Figure 22: **Clean vs attacked FSD heatmaps for** $b = 8/255$ on OSV-5M. Single ($k = 1$) start selection (selected attacks).

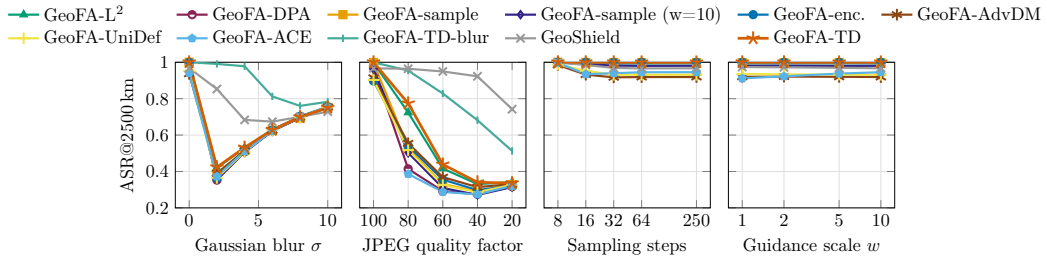


Figure 23: **Robustness to Gaussian blur, JPEG compression, sampling steps and guidance scale,** $b = 8/255$, on OSV-5M. Best $k = 5$ start selection.



Figure 24: Qualitative comparison of visual artifacts induced by GeoShield and GeoFA-TD for a random image from OSV-5M, for budget $8/255$.

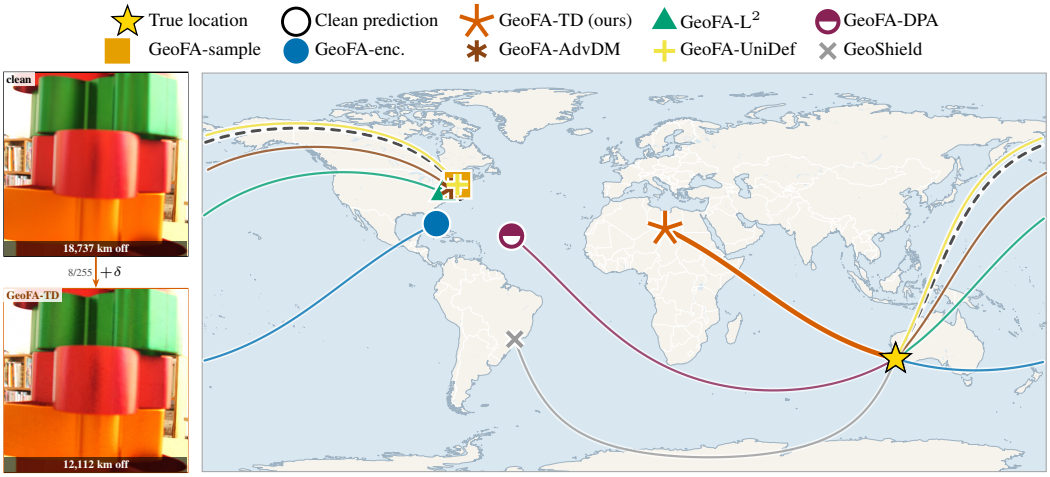


Figure 25: Attacks on a random image from YFCC4k, $b = 8/255$. RFM $\mathcal{S}^2$ mislocalizes the clean image by 18000 km.

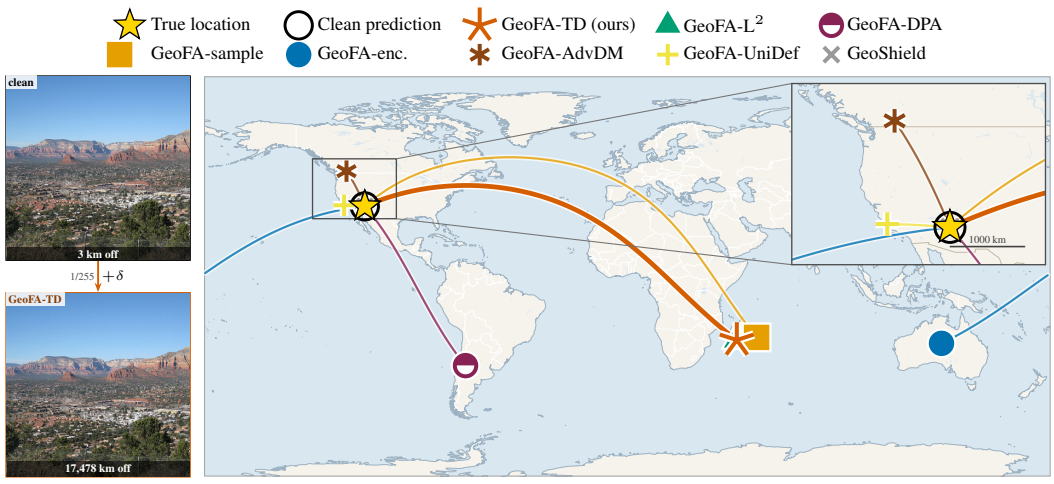


Figure 26: Attacks on a second random image from YFCC4k, $b = 8/255$. RFM $\mathcal{S}^2$ is very close to ground truth.