SOURCE-ONLY CROSS-VIEW DISRUPTION FOR TRANSFERABLE GEOLOCATION PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models can infer geographic location from ordinary images, creating privacy risks even after explicit metadata is removed. We study whether transferable protection requires constructed geographic targets or detector-defined regions, and formulate GL-SD, a source-only cross-view representation-suppression objective over an ensemble of three surrogate encoders. GL-SD uses neither a decoy location nor detector-defined attack regions. On a strict 2,972-image paired evaluation against GPT-4.1, GL-SD achieves 5,473.5 km mean privacy gain, compared with 4,361.7 km for a paper-exact GeoShield baseline. Its paired advantage is 1,111.9 km (95% image-bootstrap CI: 921.3–1,303.4), and adversarial accuracy within 2,500 km falls from 37.7% to 25.6%. The advantage independently replicates on a previously unevaluated Claude-4.5 holdout of 2,472 images (+554.9 km, CI: 365.5–742.9) and Gemini-2.5 holdout of 2,496 images (+621.5 km, CI: 436.0–809.6). Both methods satisfy the same $\ell_\infty = 8/255$ budget; on the 492-image fidelity subset, GL-SD incurs a small measured PSNR decrease relative to GeoShield (33.63 versus 33.87 dB). A late paired-view refinement further improves positive privacy-gain coverage, although its mean increment over GL-SD is not statistically significant. On a separate 474-image common subset, ordinary global Feature-PGD yields only 652.8 km mean privacy gain, compared with 4,046.7 km for GeoShield, showing that generic global feature displacement does not explain the strength of structured geolocation attacks. A canonical 482-image component ablation further identifies a conditional effect: local disruption is weak alone but provides a significant paired increment when added to the global objective. The advantage remains significant at $4/255$, $8/255$, and $16/255$ after Holm correction, exposing a consistent privacy–quality trade-off. Finally, frozen mechanism audits show that several plausible surrogate-side signals do not reliably predict per-image black-box transfer, establishing useful boundaries on explanations of geolocation attack transferability.

1 INTRODUCTION

Image geolocation by vision–language models (VLMs) creates a privacy channel that survives metadata removal. A single photograph may reveal its origin through a distributed collection of cues—scene layout, architecture, vegetation, road markings, signs, and fine-grained textures. A practical defense must suppress those cues under a small pixel budget and, crucially, must transfer from accessible surrogate encoders to closed black-box VLMs (Vo et al., 2017; Haas et al., 2024; Mendes et al., 2024).

GeoShield provides an important baseline by decomposing geographic privacy protection into semantic guidance, spatial evidence localization, and perturbation optimization (Liu et al., 2026). Its implementation, however, relies on semantic targets and detector-defined regions fixed before optimization. Such guidance can be incomplete when evidence is distributed outside the selected regions, and surrogate-space control does not necessarily imply black-box transfer.

We ask whether strong transfer can instead be obtained without constructing a surrogate geographic target, by directly suppressing representations of the source image across sampled views. This leads to GL-SD. Its implementation contains global and local objectives, but we do not claim this structure as new: GeoShield already performs joint global–local optimization. Our question is whether this

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.

structure remains effective under a source-only objective and what each component contributes under controlled decomposition. The global term covers scene-level layout and context; the local term broadens the attack to details that may be diluted in a whole-image embedding. Both are source-only: the attack moves protected views away from their corresponding clean representations rather than toward a chosen decoy. We optimize an ensemble of CLIP-B/16, CLIP-B/32, and LAION CLIP and reserve GPT-4.1 for black-box evaluation (Radford et al., 2021).

The central result is obtained on the strict intersection of 2,972 successful examples from the frozen 2,997-image IM2GPS3K evaluation. Under the same $\ell_\infty = 8/255$ budget, GL-SD attains 5,473.5 km mean privacy gain, versus 4,361.7 km for paper-exact GeoShield. The paired improvement is 1,111.9 km with a 95% bootstrap confidence interval of 921.3–1,303.4 km. Provider-specific paired evaluations also favor GL-SD on a previously unevaluated Claude-4.5 holdout and on Gemini-2.5. These gains cannot be explained by sample composition, clean predictions, or attack budget.

We also test whether generic visual feature optimization is already sufficient. A three-CLIP, global-only Feature-PGD baseline is matched to the attack budget and evaluated using the same frozen clean predictions. On 474 jointly valid images it obtains 652.8 km mean PG, versus 294.6 km for random noise and 4,046.7 km for GeoShield. Optimization matters, but generic global feature displacement accounts for only a small fraction of the effect of a geolocation-oriented structured attack.

Our contributions are:

- We reformulate transferable geolocation privacy protection as source-only cross-view representation suppression, removing the need for a target image, surrogate semantic target, or detector-defined region during optimization.
- Through a matched decomposition of global-only, local-only, and joint objectives, we identify a bounded conditional effect: local source disruption is weak alone but provides a significant paired increment when added to the global objective. We do not interpret this as a general claim that global–local attacks are new or universally superior.
- We evaluate the formulation against random noise, global Feature-PGD, and paper-exact GeoShield under matched perturbation budgets, common-image intersections, and paired uncertainty estimates. The GL-SD advantage over GeoShield is established at full IM2GPS3K scale on GPT-4.1 and independently replicated on Claude-4.5 and Gemini-2.5. A three-budget privacy–quality analysis and synchronized runtime benchmark quantify fidelity and cost. Frozen negative audits further delimit several unsupported explanations of black-box transfer.

The resulting paper makes a deliberately bounded claim: GL-SD is a strong empirical source-disruption method under the evaluated protocol. We do not claim reliable targeted redirection, universal superiority across all VLMs, or a single internal mechanism that explains transfer.

2 Related Work

Image geolocation and location privacy. Image geolocation has progressed from retrieval and classification systems to models that combine large-scale visual representations with geographic reasoning. IM2GPS3K remains a standard geographically diverse evaluation set (Vo et al., 2017), while recent systems demonstrate substantially stronger country-, city-, and coordinate-level inference (Haas et al., 2024). This progress also creates a privacy risk: removing EXIF metadata does not remove location evidence encoded in image content (Mendes et al., 2024; Yang et al., 2026). GeoShield directly addresses this threat with bounded adversarial perturbations designed to transfer to black-box VLMs (Liu et al., 2026). Importantly, GeoShield's PSAE already jointly optimizes global random crops and local regions. We therefore do not claim novelty for global–local optimization itself. Our work retains its defensive threat model but asks whether target construction and detector-defined guidance can be removed in favor of source-only cross-view suppression. Concurrent geo-privacy defenses instead disrupt hierarchical reasoning or introduce semantic typographic signals (Zhang et al., 2026; Zhu et al., 2025). These approaches alter reasoning or add visible tex-

tual content, whereas GL-SD studies bounded, source-only pixel perturbations transferred through visual encoders.

Transferable adversarial attacks. Transfer attacks optimize accessible surrogate models and rely on shared decision geometry to affect an unseen target. Momentum and transformation diversity are established tools for reducing surrogate-specific overfitting (Dong et al., 2018; Xie et al., 2019); model ensembles provide another route to transferable gradients (Chen et al., 2024). Recent attacks on closed multimodal systems increasingly optimize intermediate visual features rather than a target model's unavailable output loss (Jia et al., 2025). GLEAM also combines global and local transformations to improve multimodal attack transferability (Liu et al., 2025). GL-SD follows the broader feature-space and transformation-diversity literature but studies a different, bounded question: target-free protection against coordinate-level geolocation inference. It uses a deliberately simple three-encoder ensemble, with momentum and standard input-diversity augmentation disabled, so that the source-only objective can be decomposed under a common execution path.

Global and local visual evidence. Global image embeddings capture scene context but can dilute small location cues. Conversely, detector-defined or saliency-selected regions impose a specific hypothesis about where evidence resides. Randomized local views offer broader support without requiring the black-box model to expose its evidence map; randomized masking has also proved useful for black-box attribution (Petsiuk et al., 2018). GL-SD is not an attribution method, and coupling global and local views is not claimed as a new attack structure. Our controlled audits further show that identifying an interpretable or surrogate-important region is not, by itself, sufficient evidence that attacking that region improves black-box geolocation privacy.

3 METHOD

Let $x \in [0,1]^{H\times W\times 3}$ be a clean image with ground-truth coordinate y, let $x' = \Pi_{[0,1]}(x+\delta)$ be its protected version, and constrain $\|\delta\|_\infty \le \epsilon$. The black-box geolocator is never queried during optimization.

3.1 THREAT MODEL

The defender has white-box access to a fixed ensemble of generic visual encoders but has neither gradients nor logits from the deployed geolocator. The defender may preprocess an image once before publication and seeks to increase the geolocator's great-circle error without changing metadata or querying the protected service. The adversary is therefore image-specific, transfer-based, and query-free with respect to the evaluated black box. We do not assume a desired decoy coordinate, a geographic text description, or an external detector that identifies where location evidence occurs.

3.2 SOURCE-ONLY CROSS-VIEW DISRUPTION

Let $a \sim \mathcal{A}$ denote a global random-resized view and $c \sim \mathcal{C}$ a 224×224 local crop. For surrogate encoder f_m, the global source-disruption objective minimizes similarity between a protected global view and the cached clean-image representation:

$$\mathcal{L}_{\mathrm{G}}(x', x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{a\sim\mathcal{A}}[\cos(f_m(a(x')), f_m(x))]\,. \tag{1}$$

At initialization, GL-SD independently samples a clean local reference $c_0 \sim \mathcal{C}$ and caches $f_m(c_0(x))$. During optimization it draws fresh adversarial local views $c_t \sim \mathcal{C}$, giving

$$\mathcal{L}_{\mathrm{L}}(x', x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{c_t\sim\mathcal{C}}[\cos(f_m(c_t(x')), f_m(c_0(x)))]\,. \tag{2}$$

Thus "unpaired" has a precise meaning: the clean reference and the current protected view have the same marginal view family but do not share spatial parameters. This deliberately discourages a perturbation that succeeds only for one fixed correspondence. It also distinguishes GL-SD from ordinary paired augmentation consistency.

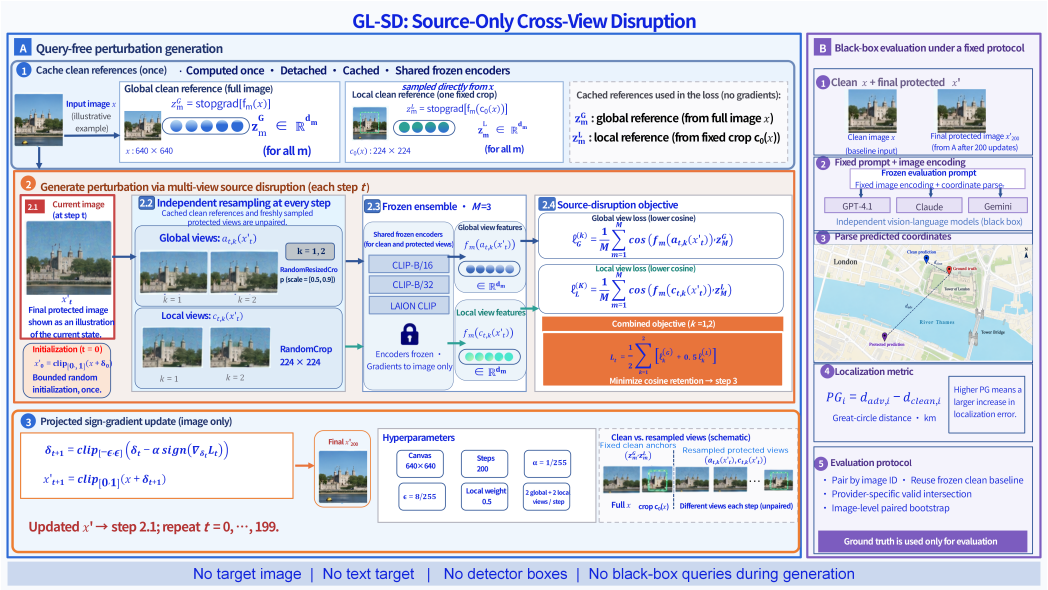

Figure 1: Overview of GL-SD. Clean global and local representations are cached once; independently resampled protected views are evaluated by the frozen three-encoder ensemble and combined in the source-retention objective. Projected sign-gradient updates modify only the protected image, and black-box VLMs are queried only after optimization. The scene shown in the diagram is an illustrative example used to visualize the data flow.

The complete objective is

$$\mathcal{L}_{\text{GL-SD}} = \mathcal{L}_{\text{G}} + \lambda \mathcal{L}_{\text{L}}, \qquad \min_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\text{GL-SD}}. \tag{3}$$

We use $M = 3$ encoders (CLIP-B/16, CLIP-B/32, and LAION CLIP), $\lambda = 0.5$, two local views per step, and crop scales sampled from $[0.5, 0.9]$. Projected sign-gradient updates run for 200 steps with step size $1/255$ on a 640×640 canvas and are projected into $\mathcal{B}_\infty(x, 8/255) \cap [0,1]^{H \times W \times 3}$. Momentum and standard input-diversity (DI) augmentation are disabled in the frozen configuration; the stochastic views above remain intrinsic to the GL-SD objective. Although historical configuration files contain target-image and bounding-box fields, neither is read by the GL-SD loss.

3.3 Late paired-view refinement

F2 retains the unpaired GL-SD objective for the first 100 updates. For updates $t \ge 100$, it replays the same sampled global and local view parameters on the clean and protected branches. It changes only correspondence timing; all weights, models, steps, and budgets remain fixed. Because its paired mean gain over GL-SD is not significant, F2 remains a refinement. Appendix Section C.1 gives the exact objectives.

3.4 Conditional contribution of view families

The global term requires displacement to survive framing changes, while the local term exposes details diluted by pooled embeddings; neither assigns a semantic label to a crop. The zeroed-component experiment tests only whether the local term adds downstream benefit conditional on the global term. We call this "conditional complementarity," not synergy, because super-additivity is not established. The distributional interpretation is formalized in Appendix Section C.1.

3.5 Optimization and static control

Each iteration averages over three encoders and two sampled views, takes a projected sign-gradient step, and updates only x'. Attack generation consumes no coordinate, target photograph, prompt,

detector box, or black-box response. Our matched Feature-PGD control instead attacks one fixed whole-image embedding; G-only retains GL-SD's stochastic global-view execution but sets the local weight to zero. Exact update and control objectives are in Appendix Section C.1.

4 EXPERIMENTS

4.1 PROTOCOL

We evaluate on the frozen 2,997-image IM2GPS3K set. The primary table uses the strict intersection of 2,972 images for which paper-exact GeoShield and GL-SD both have valid GPT-4.1 responses. Every method uses the same clean reference. The evaluator uses model `gpt-4.1`, low image detail, JPEG quality 85, and a 128-token response limit. Its fixed prompt has SHA-256 `59eb5fab5d04a62c...77a34544a7405`; full hashes and source-result hashes are retained in the artifact. Failed coordinate parses are excluded only before constructing the shared intersection and are never silently assigned a zero gain.

The strict-common 2,972-image table supports the primary GeoShield/GL-SD claim. F2 and controls generated later are compared on exact common intersections while reusing the frozen clean prediction by image identifier. Random, Feature-PGD, and GeoShield share 474 valid images for the baseline analysis.

All attacks operate on a 640×640 canvas for 200 steps with step size $1/255$ and $\epsilon = 8/255$. GL-SD and F2 use the same three-surrogate ensemble described in Section 3. The black-box model supplies no gradient, prediction, or selection signal during attack generation.

The controls answer distinct questions. Random noise tests whether the pixel budget alone changes VLM outputs. Feature-PGD tests a static, gradient-optimized whole-image displacement. G-only and L-only are matched-path zeroed-component ablations: both retain the GL-SD attack loop and surrogate ensemble, while one objective coefficient is set to zero. This taxonomy prevents the strong G-only result from being confused with the weaker but conceptually different Feature-PGD control.

Comparison discipline. We distinguish three sources of variation that are otherwise easy to conflate: attack method, API validity, and sample composition. Absolute statistics are reported only on the named intersection for a table, while method differences are computed image-wise before aggregation. Confidence intervals resample images, not API fields or individual distance components. No sample is selected using its adversarial PG, and exploratory screens are excluded from confirmatory tables.

4.2 METRICS AND STATISTICS

Privacy Gain is defined per image as

$$\mathrm{PG}_i = d(\hat{y}_i^{\mathrm{adv}}, y_i) - d(\hat{y}_i^{\mathrm{clean}}, y_i), \tag{4}$$

where d is great-circle distance. We report mean and median PG, positive-PG rate, localization accuracy at fixed radii, paired bootstrap confidence intervals, and valid denominators.

4.3 MAIN COMPARISON

Table 1 reports the primary comparison. GL-SD improves mean PG by 1,111.9 km over GeoShield; an image-level paired bootstrap with 10,000 replicates gives a 95% CI of $[921.3, 1303.4]$ km. GL-SD also increases the positive-PG rate from 79.5% to 85.2% and reduces Adv@2500 from 37.7% to 25.6%.

4.4 IS ORDINARY FEATURE DISPLACEMENT SUFFICIENT?

Table 3 separates pixel noise from generic gradient-based displacement. Random noise reduces Hit@25 by 1.9 percentage points. Feature-PGD reduces it by 5.9 points and improves mean PG over Random by 358.1 km (95% CI: 96.2–621.5), confirming an optimization effect. Its median PG is nevertheless zero and its mean PG is only 652.8 km. On the same 474-image intersection,

Table 1: Strict paired GPT-4.1 results on the common 2,972-image subset. Higher Privacy Gain (PG) and PG>0 are better; lower Adv@2500 is better.

Method	Mean PG (km)	Median PG (km)	PG>0	Adv@2500
GeoShield paper-exact	4,361.7	3,363.1	79.5%	37.7%
GL-SD	**5,473.5**	**5,998.7**	**85.2%**	**25.6%**

Table 2: Image-level paired contrast on the same 2,972 images. Confidence intervals use 10,000 paired bootstrap replicates.

Contrast	Mean ΔPG [95% CI]	Median ΔPG	W/D/L
GL-SD $-$ GeoShield	1,111.9 [921.3, 1,303.4]	146.6	1,649/211/1,112

GeoShield exceeds Feature-PGD by 3,393.9 km (95% CI: 2,943.2–3,871.7). Generic global feature displacement is therefore insufficient to reproduce GeoShield-level protection. This motivates GL-SD's local source objective without assigning the entire gap to any single GeoShield component.

Detailed baseline values are reported in Appendix Table 3.

4.5 Ablations

The zeroed-component experiment in Table 5 provides the direct test of the global–local construction. On 482 jointly valid images, GL-SD exceeds matched-compute G-only by 1,181.6 km (95% CI: 730.5–1,637.4) and L-only by 2,736.3 km (95% CI: 2,215.1–3,246.5). Both intervals exclude zero. Global evidence is the stronger individual component, but the local term provides significant incremental value when combined with it.

The distribution is heterogeneous. GL-SD's median paired advantage over G-only is only 0.5 km despite a large positive mean and confidence interval; the joint objective wins, ties, and loses on 251, 58, and 173 images. Against L-only, the benefit is broader (322/27/133 W/D/L and a 1,490.0 km median). We therefore claim reliable aggregate complementarity, not per-image dominance.

Why stochastic cross-view optimization matters. The confirmatory same-path control in Table 7 changes only whether four global views are frozen per image or resampled at every update. On 492 jointly valid images, Stochastic-G exceeds Fixed-G by 2,656.2 km (95% CI: 2,179.7–3,139.3), with 311/56/125 wins/ties/losses. This result supports cross-iteration resampling as an important source of transfer in the G-only path. It does not establish that resampling is the only mechanism behind the complete GL-SD advantage.

The component and same-path tables are deferred to Appendix Tables 5 and 7; Figure 3 summarizes the principal component comparisons.

F2 reaches a mean PG of 5,618.4 km and a positive-PG rate of 86.2% on the $N = 2{,}972$ strict-common subset. Relative to GL-SD, however, its paired mean increment is only 206.6 km with a 95% CI of $[-127.4, 543.3]$ km. We therefore treat F2 as a refinement and do not claim that it significantly outperforms GL-SD.

Image-fidelity statistics appear in Appendix Table 8; the F2 paired comparison is stated above and documented by the released paired JSON.

Headroom and fidelity. In the pre-defined clean-error strata, both methods have negative mean PG when clean error already exceeds 2,500 km, but GL-SD retains a +604.0 km paired advantage (95% CI: 66.7–1,153.6). All stored attacks reach the same $8/255$ budget; the full paired distribution and strata are visualized in Appendix Figure 5, with fidelity metrics in Table 8.

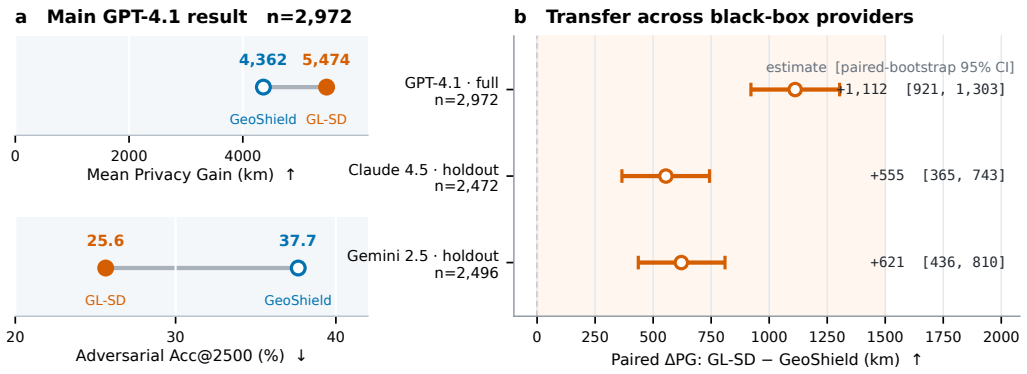

Figure 2: Main results and cross-provider transfer. **a,** Full-scale GPT-4.1 comparison on the strict-common subset of 2,972 images. GL-SD increases mean Privacy Gain and lowers residual geolocation accuracy at 2,500 km relative to paper-exact GeoShield. **b,** Provider-specific paired mean improvements in Privacy Gain. GPT-4.1 uses the full strict-common evaluation, whereas Claude-4.5 and Gemini-2.5 use their independently frozen holdout2500 subsets. Error bars denote image-level paired-bootstrap 95% confidence intervals. The panels use different evaluation scopes, and providers are analyzed independently without pooling.

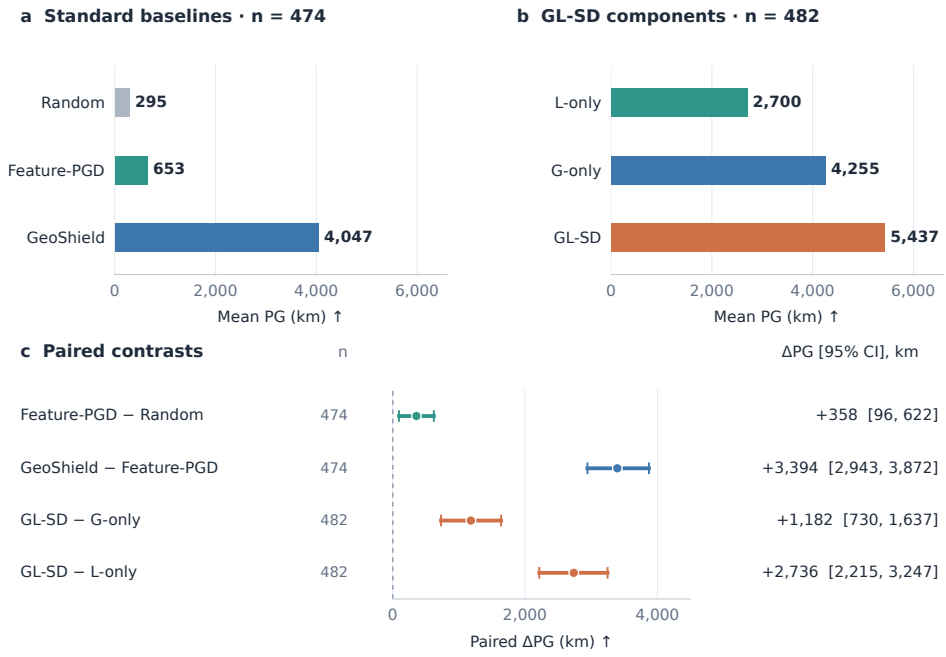

Figure 3: Standard baselines and GL-SD components on separately matched GPT-4.1 subsets. **a,** Mean Privacy Gain for Random, Feature-PGD, and GeoShield on 474 common images. **b,** Mean Privacy Gain for L-only, G-only, and GL-SD on 482 common images. **c,** Paired mean differences with image-level bootstrap 95% confidence intervals; each row uses its stated matched subset. The two subsets do not form a joint ranking.

4.6 Privacy–Quality Trade-off Across Budgets

Appendix Table 10 and Figure 4 evaluate the frozen $N = 200$ budget set; all six attack/budget cells are jointly valid on 197 images. GL-SD improves paired mean PG over GeoShield by 1,130.9,

863.8, and 1,620.5 km at $4/255$, $8/255$, and $16/255$, respectively. Positive-direction, one-sided paired sign-flip Monte Carlo tests (100,000 randomizations; image-level pairing; Holm family size three) remain significant after Holm correction ($p = 0.0034$, 0.0220, and 0.00015). The benefit is not free: relative to GeoShield, GL-SD lowers mean PSNR by 0.36, 0.24, and 0.19 dB and raises LPIPS-Alex by 0.007, 0.009, and 0.013. We therefore report a favorable but measurable privacy–quality trade-off rather than claiming identical fidelity.

Computational cost. GL-SD takes 117.76 seconds per image versus 116.23 for matched G-only. The full single-GPU benchmark is in Appendix Table 11; GeoShield's row excludes its cached semantic and detector preprocessing.

4.7 Transfer and robustness

Appendix Table 12 reports provider-specific replication. On the previously unevaluated Claude-4.5 holdout, GL-SD improves mean PG by 554.9 km with CI $[365.5, 742.9]$. On the independently frozen Gemini-2.5 holdout, the paired gain is 621.5 km with CI $[436.0, 809.6]$. We do not pool provider outputs into a pseudo-sample; each row uses its own clean reference and strict-common subset.

Figure 2b visualizes these scope-matched provider effects; Appendix Table 12 provides the complete values.

Under matched JPEG75 and $0.75\times$ resize, GL-SD retains paired advantages of +2,207.8 and +1,315.7 km, respectively. Appendix Table 13 gives the intervals and evaluation boundary.

5 Analysis and Mechanism Boundaries

The component ablation establishes objective-level complementarity under a fixed generator, not that a target VLM contains separable global and local geographic modules. Likewise, three frozen geographic probes did not predict per-image PG on the confirmation split ($\rho = 0.030$, 95% CI $[-0.410, 0.470]$, $N = 25$). GL-SD's performance therefore remains an empirical transfer result rather than proof of a scalar internal mechanism. Appendix Table 14 and Section C.4 give the complete negative-result boundaries and distributional discussion.

6 Limitations

Black-box VLM outputs are stochastic and may change with model updates. API evaluation is therefore bound to model identifiers, prompts, image hashes, and timestamps. Privacy Gain measures displacement from ground truth and does not by itself demonstrate steering toward a chosen target. F2 is treated as an optimization refinement because its mean advantage over GL-SD is not yet statistically established. Finally, bounded perturbations reduce but cannot guarantee the removal of all location information. Our strongest large-sample target is GPT-4.1 accessed through an OpenAI-compatible relay endpoint. The relay passes through model identifiers, prompts, and image hashes without modification; the exact model snapshot is verified by returned model IDs and timestamps retained in the evaluation artifact. The advantage replicates on frozen large holdouts for Claude-4.5 and Gemini-2.5; broader claims remain bounded to these exact model versions and samples. A single frozen API run cannot quantify all run-to-run variance. Feature-PGD shows that generic global displacement is weak, but does not isolate which individual GeoShield component explains the full gap. PSNR, SSIM, and LPIPS characterize image fidelity rather than human imperceptibility; a controlled human study remains necessary. Runtime comparisons use one hardware/software stack, and GeoShield's external semantic and detector preprocessing is cached rather than remeasured. Finally, GL-SD's unpaired objective compares fresh protected views with cached clean references. Its empirical benefit should not be interpreted as proof that condition coverage is the causal transfer mechanism; our mechanism audits explicitly keep that stronger claim separate.

7 CONCLUSION

We studied whether transferable image geolocation privacy can be achieved without constructing a target image, semantic surrogate, or detector-defined region. GL-SD answers this question with source-only cross-view representation suppression. We do not claim novelty for global–local optimization itself. Instead, strict paired experiments establish its gain at full IM2GPS3K scale on GPT-4.1, with provider-specific replication on Claude-4.5 and Gemini-2.5, and isolate a conditional incremental contribution of the local term. The advantage persists across three perturbation budgets after multiplicity correction, while synchronized timing and perceptual metrics expose its computational and fidelity costs. Mechanism audits clarify which stronger causal claims are not supported. This combination of attack performance and evidence discipline provides a reproducible basis for future geolocation privacy work.

REPRODUCIBILITY STATEMENT

The final submission will provide frozen sample identifiers, attack configurations, prompt hashes, image hashes, parsing rules, and paired statistical analysis code. The accompanying artifact records the exact common subsets and recomputes PG from clean and adversarial great-circle errors rather than trusting cached derived fields.

ETHICS STATEMENT

This work studies adversarial perturbations as a defensive mechanism against image geolocation and location-privacy leakage. We evaluate only authorized models and datasets and do not use the method to identify private individuals.

AI USE DISCLOSURE

Generative AI tools were used to assist with language editing, software development, experimental analysis, and manuscript organization. All AI-assisted code, analyses, and textual content were reviewed and verified by the authors. The authors take full responsibility for the methodology, experiments, claims, and final content of this paper.

REFERENCES

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

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018. doi: 10.1109/CVPR.2018.00957. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. PIGEON: Predicting image geolocations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12893–12902, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Haas_PIGEON_Predicting_Image_Geolocations_CVPR_2024_paper.html.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1944. URL https://papers.nips.cc/paper_files/paper/2025/hash/540a3bf4863672cdaebfed69fa2ce335-Abstract-Conference.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 AAAI Conference on Artificial Intelligence, volume 40, pp. 35653–35661, 2026. doi: 10.1609/aaai.v40i42.40877. URL https://ojs.aaai.org/index.php/AAAI/article/view/40877.

Yunqi Liu, Xue Ouyang, and Xiaohui Cui. GLEAM: Enhanced transferable adversarial attacks for vision–language pre-training models via global–local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1665–1674, 2025. doi: 10.1109/ICCV51701.2025.00163. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_GLEAM_Enhanced_Transferable_Adversarial_Attacks_for_Vision-Language_Pre-training_Models_via_ICCV_2025_paper.html.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17240–17292, 2024. doi: 10.18653/v1/2024.emnlp-main.957. URL https://aclanthology.org/2024.emnlp-main.957/.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, pp. 151, 2018. URL https://bmva-archive.org.uk/bmvc/2018/contents/papers/1064.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the deep learning era. In *IEEE International Conference on Computer Vision*, pp. 2640–2649, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/html/Vo_Revisiting_IM2GPS_in_ICCV_2017_paper.html.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019. doi: 10.1109/CVPR.2019.00284. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Improving_Transferability_of_Adversarial_Examples_With_Input_Diversity_CVPR_2019_paper.html.

Ruixin Yang, Ethan Mendes, Arthur Wang, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Do vision-language models respect contextual integrity in location disclosure? In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/38380a17248c4dfc4f48fc4d836d9b62-Abstract-Conference.html.

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. OpenReview preprint submitted to ICLR 2026, 2026. URL https://openreview.net/forum?id=5S6YTG9dL0. Acceptance status not independently verified.

Jiayi Zhu, Yihao Huang, Yue Cao, Xiaojun Jia, Qing Guo, Felix Juefei-Xu, Geguang Pu, and Bin Wang. Beyond pixels: Semantic-aware typographic attack for geo-privacy protection. *arXiv preprint arXiv:2511.12575*, 2025. doi: 10.48550/arXiv.2511.12575. URL https://arxiv.org/abs/2511.12575.

A Additional experimental details

GL-SD uses CLIP-B/16, CLIP-B/32, and LAION CLIP, two local views per step, crop scale $[0.5, 0.9]$, local weight 0.5, 200 steps, and step size $1/255$. All images are optimized on a 640×640

canvas under $\ell_\infty = 8/255$. GPT-4.1 evaluation uses low image detail, JPEG quality 85, temperature zero, a 128-token limit, and the prompt hash reported in Section 4. Frozen identifiers and complete configuration lineage are included in the artifact; robustness dry runs are excluded because they contain no model responses.

B HELD-OUT SCALE EXPANSION PROTOCOL

Images 501–2000 are reserved for an additional GL-SD scale replication and are partitioned into three frozen 500-image shards. Before evaluation, each shard must pass one-to-one PNG/tensor completeness, unique-ID, configuration, exact tensor-reconstruction, and $8/255$ budget checks. The three shards are reported separately as well as pooled; incomplete-shard results are never inserted into the paper. This expansion supports an absolute GL-SD generalization claim. It supports superiority over another method only if that method is generated and evaluated on the same images.

B.1 EXECUTABLE GL-SD SPECIFICATION

For each image, the implementation first caches one full-image clean feature and one randomly cropped clean local feature per surrogate. At every update it constructs two independent Monte Carlo loss samples. Each sample contains a fresh random-resized protected global view and a fresh protected 224×224 local crop. The three-surrogate losses are averaged, the two Monte Carlo samples are summed before differentiation, and the protected image is updated once. The sign convention minimizes cosine similarity to the cached clean features. F2 follows the same procedure through update 99; from update 100 onward, clean features are recomputed from views whose random parameters are exactly replayed on the protected branch.

This specification also defines what is absent. The target-image dataloader and bounding-box path remain in legacy configuration schemas but contribute no tensor to GL-SD or F2. Ground-truth coordinates are not loaded by the attack loss. We retain these statements because configuration fields alone can otherwise give the misleading impression that the method is target- or detector-conditioned.

B.2 BLACK-BOX EVALUATION AND PAIRED INFERENCE

Each clean prediction is frozen by image identifier and reused across attacks. The evaluator requests a coordinate, parses latitude and longitude using one fixed rule, and marks an example invalid if no coordinate can be recovered. For a comparison, we first intersect valid identifiers and only then compute per-image PG and method differences. A paired bootstrap draws image identifiers with replacement and recomputes the mean difference for each of 10,000 replicates. Thus the reported interval reflects image-level heterogeneity and preserves the within-image clean baseline.

B.3 INTERPRETATION OF THE DISTANCE METRIC

Positive PG means that the protected prediction is farther from ground truth than the frozen clean prediction; negative PG means that the perturbation accidentally improved localization. Because great-circle error is bounded and clean errors vary substantially, mean PG, median PG, PG> 0, and threshold accuracy capture different behavior. We therefore do not treat a gain in one summary as automatic dominance on every criterion.

C DETAILED RESULTS AND AUDITS

This appendix preserves the complete controls, secondary breakdowns, robustness results, and mechanism-boundary audit referenced in the main text. Moving these materials does not change any estimate, denominator, interval, or conclusion.

Table 3: Matched comparison on 474 images with valid Random, Feature-PGD, and GeoShield outputs. PG is recomputed from the shared frozen clean error.

Method	Clean Hit@25	Adv. Hit@25	Mean PG (km)	Median PG (km)
Random ℓ_∞ noise	44.5%	42.6%	294.6	−0.0
Global Feature-PGD	44.5%	38.6%	652.8	0.0
GeoShield paper-exact	44.5%	15.4%	4,046.7	2,327.8

Table 4: Paired baseline contrasts with 10,000 bootstrap replicates.

Contrast	N	Mean ΔPG [95% CI]	W/D/L
Feature-PGD − Random	474	358.1 [96.2, 621.5]	302/0/172
GeoShield − Random	474	3,752.0 [3,278.9, 4,226.7]	383/0/91
GeoShield − Feature-PGD	474	3,393.9 [2,943.2, 3,871.7]	332/53/89

C.1 Secondary objectives and controls

For updates $t \geq 100$, F2 samples a_t and c_t once and replays their exact parameters on the clean and protected branches:

$$\mathcal{L}_{\mathrm{G}}^{\mathrm{pair}}(t) = \frac{1}{M} \sum_{m} \cos(f_m(a_t(x')), f_m(a_t(x))) \,, \tag{5}$$

$$\mathcal{L}_{\mathrm{L}}^{\mathrm{pair}}(t) = \frac{1}{M} \sum_{m} \cos(f_m(c_t(x')), f_m(c_t(x))) \,. \tag{6}$$

The global and local branches may be viewed as a mixture $q(v) = \pi_{\mathrm{G}} q_{\mathrm{G}}(v) + \pi_{\mathrm{L}} q_{\mathrm{L}}(v)$. This is an optimization interpretation, not a claim that sampled views are the black-box model's causal evidence.

Starting from a bounded random perturbation, GL-SD averages the loss over the three encoders and two sampled views, takes a sign-gradient step, and projects onto the ℓ_∞ ball and valid image range. Clean features are detached; only x' receives gradients. The static Feature-PGD control minimizes

$$\mathcal{L}_{\mathrm{FPGD}}(x', x) = \frac{1}{M} \sum_{m=1}^{M} \cos(f_m(x'), f_m(x)) \tag{7}$$

with the same encoders, canvas, steps, step size, and budget. It contains no local view, mask, text, target image, or consistency term.

C.2 Full-scale and control results

The corresponding full-scale and control results are reported in Tables 3–9 below.

C.3 Budget, cost, transfer, and robustness details

Budget, runtime, transfer, and robustness results are summarized in Tables 10–13 and Figures 4–5.

C.4 Mechanism-boundary audit

We distinguish three levels of evidence: an intervention can change a surrogate representation; the change can recur across surrogates; and the change can predict black-box PG. Only the third supports a transfer-mechanism claim. Cross-model causal effects can be matched while their residual variation remains unrelated to PG; functional correspondence is therefore not equivalent to transfer relevance. The tested scalar surrogate proxies did not reliably identify transfer under our protocols, but this does not imply that transfer lacks shared structure.

Table 5: Canonical GPT-4.1 component ablation on the strict common 482-image intersection. Higher PG and lower Adv@2500 are better.

Method	Mean PG (km)	Median PG (km)	PG>0	Adv@2500
G-only	4,255.0	2,638.3	78.0%	39.6%
L-only	2,700.3	233.3	65.4%	56.6%
GL-SD	**5,436.6**	**5,562.2**	**83.0%**	**27.0%**

Table 6: Paired component contrasts with 10,000 bootstrap replicates.

Contrast	Mean ΔPG [95% CI]	Median	W/D/L
GL-SD $-$ G-only	1,181.6 [730.5, 1,637.4]	0.5	251/58/173
GL-SD $-$ L-only	2,736.3 [2,215.1, 3,246.5]	1,490.0	322/27/133
G-only $-$ L-only	1,554.7 [1,049.4, 2,052.6]	151.8	284/43/155

The method claim is independent of these negative mechanism results. GL-SD was selected by held-out black-box performance, not by filtering images or choosing a proxy after observing PG. Conversely, its performance is not used to retroactively validate a failed internal explanation.

The component ablation gives causal evidence only at the attack-objective level: under a fixed generator, removing either view-family loss reduces mean downstream PG. It does not establish that the target VLM internally separates global and local geographic modules, or that every sampled crop contains an interpretable geographic clue. A useful training objective need not expose a scalar surrogate score that predicts which individual image will transfer.

Finally, the near-zero median GL-SD–G-only difference motivates a distributional interpretation. The local term produces large gains on a subset but is neutral or harmful elsewhere. F2's higher positive-PG rate without a significant mean advantage suggests a related mean–tail trade-off. We report mean, median, sign rate, and W/D/L together so that neither a few large gains nor a majority of small ties determines the conclusion alone.

Table 7: Same-path view-resampling control on the strict-common $N = 492$ set. The objective, views per update, surrogate ensemble, and optimization budget are identical; only cross-iteration resampling changes.

Method	Mean PG	Median PG	PG> 0
Fixed-G	2,062.6	12.9	56.1%
Stochastic-G	4,718.7	3,990.8	79.3%
Paired difference	+2,656.2 km, 95% CI [2,179.7, 3,139.3]		

Table 8: Perturbation-level image quality on strict-common492, computed from stored float tensors before evaluation encoding. These metrics do not by themselves establish human imperceptibility.

Method	PSNR ↑	SSIM ↑	MAE_{255} ↓	$\ell_\infty/255$
GeoShield paper-exact	33.869	0.8773	4.495	8.0
GL-SD	33.631	0.8733	4.664	8.0
GL-SD + F2	33.620	0.8732	4.671	8.0

Table 9: Privacy gain by frozen clean-error headroom on the GPT-4.1 strict-common 2,972-image set. The main estimate retains all strata.

Clean error stratum	N	GeoShield PG		GL-SD PG	
		Mean	Median	Mean	Median
≤ 200 km	1,802	5,036	3,880	6,275	6,815
$(200, 750]$ km	449	5,201	5,007	6,602	7,180
$(750, 2500]$ km	360	5,146	5,385	5,771	6,030
> 2500 km	361	-832	-304	-228	-2

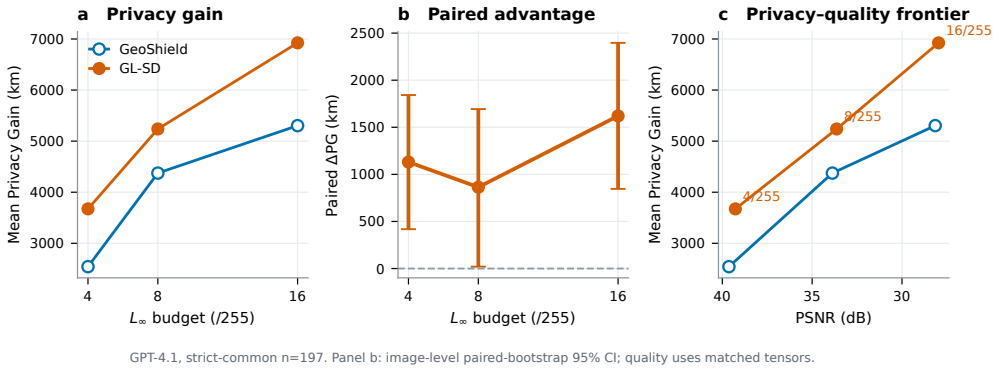

Figure 4: Budget sensitivity on the strict-common 197-image GPT-4.1 subset. Panels show mean Privacy Gain, paired GL-SD-minus-GeoShield differences with pointwise image-bootstrap 95% confidence intervals, and the privacy–PSNR frontier. Higher PG and PSNR are better.

Table 10: Privacy–quality trade-off on the frozen EPS-N200 strict-common set ($N = 197$). Quality columns show GeoShield/GL-SD. p_{Holm} adjusts the three positive-direction, one-sided paired sign-flip Monte Carlo tests (100,000 randomizations). All tensor and decoded image ℓ_∞ maxima equal the stated budget.

Budget	GeoShield PG	GL-SD PG	ΔPG	p_{Holm}	PSNR (GS/GL)	LPIPS (GS/GL)
4/255	2,542.6	3,673.5	1,130.9	0.0034	39.63/39.27	0.059/0.065
8/255	4,374.4	5,238.2	863.8	0.02197	33.87/33.64	0.160/0.169
16/255	5,303.6	6,924.1	1,620.5	0.00015	28.15/27.96	0.326/0.339

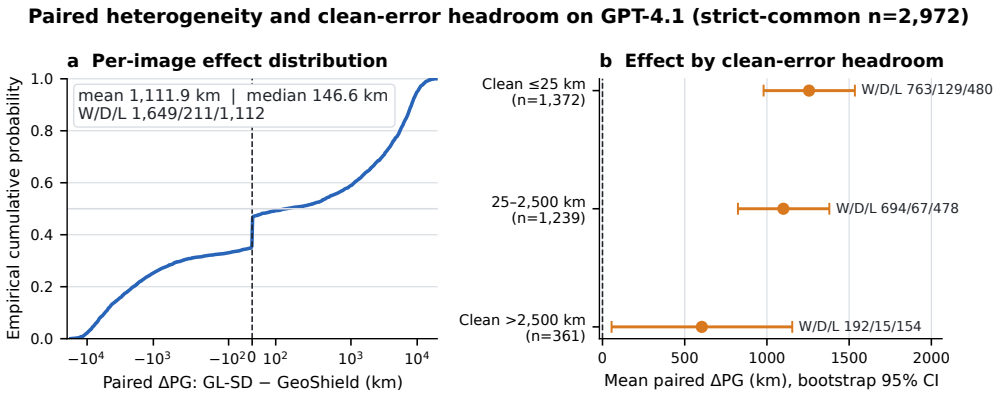

Figure 5: Paired effect heterogeneity and clean-error headroom on GPT-4.1. **a,** Empirical cumulative distribution of the per-image Privacy Gain difference between GL-SD and paper-exact GeoShield on the full strict-common subset of 2,972 images. The symmetric-log horizontal scale preserves both tails; the dashed line marks no difference. **b,** Mean paired difference within strata defined by the frozen clean geolocation error. Error bars are image-level paired-bootstrap 95% confidence intervals (10,000 replicates); W/D/L denotes wins, ties, and losses. The figure supports an aggregate improvement with substantial image-level heterogeneity, not per-image dominance.

Table 11: COST-N50 on one A100 with five warm-up images excluded and 50 timed images per method. Times include attack-local I/O. GeoShield uses frozen cached descriptions and boxes, so its row is not full end-to-end preprocessing time. Forward counts are top-level surrogate calls.

Method	Mean s/img	Median	P10–P90	Peak GiB	Forward / backward
Feature-PGD	27.59	27.62	27.42–27.83	7.23	603 / 200
DI-Feature-PGD	27.70	27.67	27.48–27.93	7.23	603 / 200
MI-Feature-PGD	27.60	27.55	27.41–27.86	7.22	603 / 200
G-only	116.23	115.60	115.03–119.66	9.71	2406 / 200
GL-SD	117.76	117.36	116.96–118.96	9.71	2406 / 200
GeoShield (cached prep.)	62.13	61.82	60.73–64.50	8.07	1206–1248 / 200

Table 12: Provider-specific paired evaluation. Each row uses its own frozen clean predictions and strict-common subset. Claude is a held-out replication.

Black-box	N	GeoShield PG	GL-SD PG	Paired ΔPG [95% CI]
GPT-4.1	2,972	4,361.7	5,473.5	1,111.9 [921.3, 1,303.4]
Claude-4.5 holdout	2,472	2,143.1	2,698.1	554.9 [365.5, 742.9]
Gemini-2.5 holdout	2,496	2,872.6	3,494.1	621.5 [436.0, 809.6]

Table 13: Real post-processing evaluation on the frozen main100 subset. PG is paired to the clean image under the same transformation.

Transform	N	GeoShield PG	GL-SD PG	Paired ΔPG [95% CI]
JPEG quality 75	98	3,258.2	5,466.0	2,207.8 [1,080.7, 3,316.2]
Resize $0.75\times$	98	3,939.2	5,254.9	1,315.7 [469.9, 2,170.2]

Table 14: Mechanism-boundary audit. A surrogate observation supports a transfer mechanism only if it predicts held-out, per-image black-box PG.

Hypothesis	Surrogate observation	Black-box test	Supported conclusion
Static semantic/spatial guidance	GNFD descriptions and Geo-EE boxes are executable	Main20 estimates have wide intervals and no independent gain	Static anchors and boxes are not established as independent transfer sources
Value/token/attention routing	Interventions produce measurable internal effects	Effect, routed-flow, and response-subspace scores fail held-out PG prediction	Internal controllability is insufficient evidence of transfer relevance
Cross-model causal correspondence	Matched interventions agree across some CLIP variants	Normalized depth explains much correspondence; residuals do not predict PG	Causal agreement does not imply black-box transfer relevance
Geographic task displacement	Three frozen geographic probes detect displacement	Confirmation split: $N = 25$, $\rho(\Delta\mathrm{disp}, \Delta\mathrm{PG}) = 0.030$, 95% CI $[-0.410, 0.470]$	Under this probe and split, displacement does not reliably predict image-level PG
Targeted control	Strong directional control exists on an optimized surrogate	Held-out control is near zero; automatic targeting underperforms F2	Surrogate control does not establish transferable redirection
Cross-paradigm dynamics	Three of four geo probes are valid	SAM2 validity fails the preregistered gate	No four-paradigm shared-transition claim is made