# WHEN GEOGRAPHIC PRIORS BECOME BIAS: CAUSAL STEERING IN REMOTE-SENSING VLMS

**Anonymous authors**
Paper under double-blind review

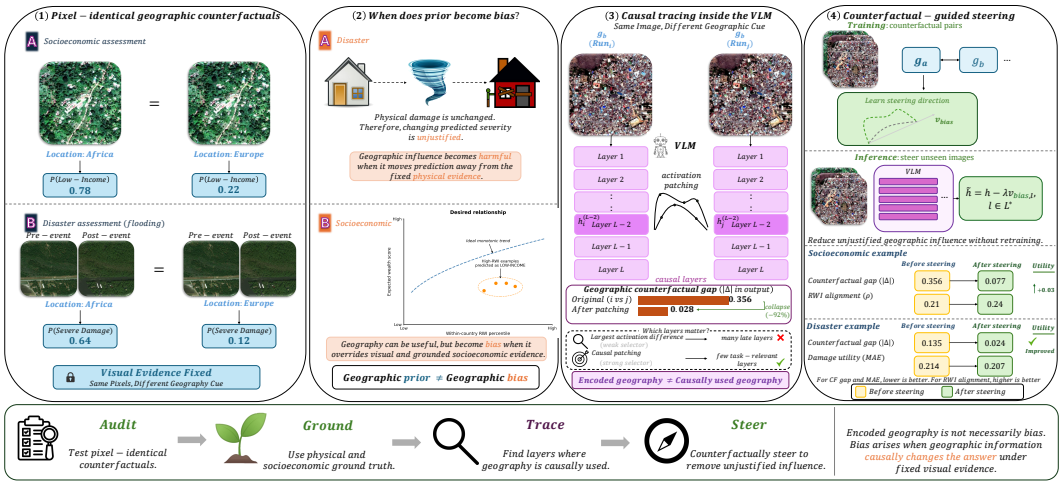



Figure 1: **From geographic prior to geographic bias in remote-sensing VLMs.** Our framework follows four stages: **(1)** pixel-identical geographic counterfactuals isolate the effect of stated geography while keeping visual evidence fixed; **(2)** task-specific grounding distinguishes useful geographic priors from unsupported geographic bias; **(3)** causal activation patching identifies representations through which geography actually changes the prediction rather than merely being encoded; and **(4)** counterfactual-guided steering intervenes on these causally localized representations at inference time, reducing unsupported geographic influence while preserving task utility. Together, the framework follows *Audit* → *Ground* → *Trace* → *Steer*.

## ABSTRACT

Geographic context can be a useful prior in remote sensing, but it becomes problematic when it overrides fixed visual evidence. We study this transition from geographic prior to geographic bias in remote-sensing vision-language models (VLMs) using pixel-identical geographic counterfactuals, where only the stated location is changed while the visual input remains unchanged. This isolates geographic influence from genuine differences in landscape, infrastructure, or physical damage. Across socioeconomic assessment and disaster-damage estimation, we find that geographic cues can substantially shift predictions even when the underlying imagery is identical, with external RWI and damage annotations revealing when such shifts are unsupported. We then use activation patching to identify transformer representations that causally mediate these effects and construct bias-weighted steering directions for inference-time intervention. Steering consistently reduces geographic counterfactual gaps on unseen images while preserving task-relevant utility. To the best of our knowledge, this is the first framework to jointly *audit, causally localize, and selectively steer* geographic bias in remote-sensing VLMs under controlled pixel-identical counterfactuals. Our results show that geographic priors can become causal shortcuts, but that these effects can be systematically identified and mitigated without retraining.

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.

# 1 INTRODUCTION

Remote-sensing VLMs increasingly connect overhead imagery with natural-language concepts, enabling open-vocabulary recognition, visual question answering, spatial reasoning, and grounded dialogue (Liu et al., 2024; Kuckreja et al., 2024). These capabilities are becoming relevant to societally consequential applications such as poverty estimation (Jean et al., 2016; Yeh et al., 2020), socioeconomic mapping (Chi et al., 2022), and disaster assessment (Wang et al., 2026). Geography is naturally informative in remote sensing. Climate, infrastructure, urban morphology, vegetation, land use, and socioeconomic conditions all vary spatially, making geographic priors potentially useful. The presence of geographic information inside a model is therefore not itself evidence of bias. The problem arises when geographic context begins to *override* rather than complement the visual evidence. Existing geographic evaluations cannot cleanly establish this distinction. Prior work has studied geographic dataset imbalance (Shankar et al., 2017), geographically diverse benchmarks (Ramaswamy et al., 2023), geographic bias metrics (Wang et al., 2025), and representation disparities in satellite-based poverty estimation (Aiken et al., 2023). However, comparisons across different regions inherently confound geography with real differences in the observed scenes. If two images receive different predictions, the cause may be geographic bias, but it may also be genuine variation in infrastructure, environment, or socioeconomic conditions. We therefore ask a more controlled question: **When does a useful geographic prior become geographic bias?**

We study this question using **pixel-identical geographic counterfactuals**. For a fixed visual input, we change only the stated geographic context. The image, objects, and physical evidence remain identical. Any resulting prediction shift therefore isolates the influence of geographic context rather than differences between locations. This idea is related to counterfactual fairness (Kusner et al., 2017) and counterfactual auditing of social bias in VLMs (Howard et al., 2024; 2025), but introduces an important distinction for remote sensing: geography is not universally irrelevant. We therefore treat geographic bias as *counterfactual geographic influence that is unsupported by the fixed visual evidence*.

We investigate this phenomenon in two complementary settings. In **socioeconomic assessment**, geographic context may contain legitimate prior information, so we additionally ground model predictions against fine-grained Relative Wealth Index measurements (Chi et al., 2022). In **disaster-damage assessment**, the setting is stricter: for an identical pre-/post-event image pair, the underlying physical damage cannot change with the stated location. Using xBD/xView2 (Gupta et al., 2019), we find that geographic cues nevertheless alter predicted damage severity. Together, these tasks allow us to distinguish geographic sensitivity from geographic influence that conflicts with external or physically invariant evidence. Behavioral auditing alone does not reveal how such effects arise internally. We therefore use counterfactual activation patching (Heimersheim & Nanda, 2024; Meng et al., 2022) to identify representations that *causally mediate* the geographic prediction gap. Importantly, strong geographic encoding need not imply causal relevance: a representation may contain geographic information without using it to determine the task prediction. This causal structure then enables targeted intervention. Building on activation-level steering (Li et al., 2023; Rimsky et al., 2024), we construct bias-weighted directions from counterfactual representation differences at causally selected layers and apply them at inference time.

The goal is not to erase geographic knowledge, but to selectively suppress the component associated with unsupported counterfactual shifts while preserving task-relevant utility. Figure 1 summarizes this process from controlled auditing and external grounding to causal localization and selective steering. Our contributions are:

We introduce a **pixel-identical geographic counterfactual framework** for identifying when geographic priors become unsupported prediction shortcuts, and ground these effects using independent socioeconomic and physical-damage evidence.

We introduce **causal localization of geographic influence** in remote-sensing VLMs, using counterfactual activation patching to distinguish representations that merely encode geography from those that causally influence task predictions.

We propose **counterfactual-guided causal steering**, an inference-time intervention that constructs bias-weighted directions from causally selected representations and reduces geographic counterfactual effects on unseen images without retraining while preserving task-relevant utility.

## 2 Related Work and Problem Formulation

**Geographic Bias in Remote-Sensing Models** Geographic variation is intrinsic to remote sensing. Climate, infrastructure, settlement morphology, land use, and socioeconomic conditions vary spatially, making geographic information potentially useful for prediction while uneven geographic representation can introduce systematic disparities. Prior work has documented geographic imbalance in visual datasets (Shankar et al., 2017), developed geographically diverse benchmarks such as GeoDE (Ramaswamy et al., 2023), and proposed explicit measures of geographic bias (Wang et al., 2025). Geographic disparities have also been observed in satellite-model transfer (Zhang et al., 2022) and satellite-based poverty estimation (Aiken et al., 2023). The problem becomes more subtle for foundation models. Language models encode substantial geographic knowledge while exhibiting systematic geographic associations (Manvi et al., 2024), and remote-sensing VLMs such as RemoteCLIP (Liu et al., 2024) and GeoChat (Kuckreja et al., 2024) combine visual evidence with linguistic and world knowledge. Their predictions may therefore depend simultaneously on what is visible and on geographic priors acquired during pretraining. This makes conventional cross-region evaluation insufficient for identifying geographic bias. Different predictions for images from Africa and Europe may reflect an undesirable geographic shortcut, but they may equally reflect real differences in infrastructure, environment, or socioeconomic conditions. A controlled evaluation must therefore separate the effect of *stated geography* from variation in the visual scene itself.

**Pixel-Identical Geographic Counterfactuals** Counterfactual fairness asks whether a prediction changes when an attribute is intervened upon while relevant properties of the evaluated instance remain fixed (Kusner et al., 2017). Counterfactual examples have similarly been used to expose social and intersectional biases in vision-language models (Howard et al., 2024; 2025), while VLMs have been shown to rely on learned semantic expectations that may be weakly supported by visible evidence (Vo et al., 2025). We adapt this principle to geography through *pixel-identical geographic counterfactuals*: the visual input and task remain fixed while only the stated geographic context changes. Prediction differences can therefore be attributed to the geographic cue rather than changes in buildings, land cover, image quality, or physical damage. Crucially, geographic sensitivity is not itself evidence of bias. Geography may contain legitimate task information, particularly for socioeconomic reasoning. We therefore distinguish a geographic prior that complements task-relevant evidence from geographic influence that overrides or conflicts with that evidence. In our formulation, **geographic prior $\neq$ geographic bias**: geographic dependence becomes problematic when a change in geographic context alters the model's judgment without corresponding support from the fixed visual or independently grounded evidence. Section 3 formalizes this distinction.

**Causal Tracing of Geographic Influence** Counterfactual auditing reveals *whether* geography affects a prediction, but not *how* that effect is mediated internally. Causal mediation and intervention methods have been used to study bias and factual knowledge in language models (Vig et al., 2020; Meng et al., 2022), while activation patching provides a direct test of whether a particular internal representation contributes causally to an observed behavior (Heimersheim & Nanda, 2024). Geographic representations have also recently been studied through activation-level interventions in language models (De Sabbata et al., 2026). For our setting, the important distinction is between *encoded geography* and *causally used geography*. A representation may strongly encode location while having little influence on the task prediction. We therefore use controlled geographic counterfactuals not merely to detect representation differences, but to test whether intervening on those representations actually changes the geographic prediction effect. This provides a causal criterion for identifying the internal representations relevant to the observed behavior.

**Counterfactual-Guided Steering** Representation-level interventions can alter model behavior at inference time without parameter updates (Li et al., 2023), and contrastive activation directions can control specific behavioral properties (Rimsky et al., 2024). However, suppressing geographic information indiscriminately is undesirable in remote sensing because geography may provide legitimate predictive information. The relevant objective is therefore not geographic erasure, but *selective intervention*: suppress the component of geographic influence associated with unsupported prediction shifts while retaining useful task-relevant information. Existing geographic-disparity analyses (Shankar et al., 2017; Ramaswamy et al., 2023; Wang et al., 2025) identify geographic differences but do not localize their causal mechanism, while counterfactual auditing (Howard et al., 2024;

2025) does not by itself provide a targeted intervention. Generic activation steering (Li et al., 2023; Rimsky et al., 2024), in turn, does not distinguish useful geographic priors from unsupported geographic shortcuts. Our formulation connects these pieces: controlled counterfactuals establish when geography changes a prediction beyond the available evidence, causal interventions identify the representations mediating that effect, and inference-time steering selectively targets those representations. The resulting goal is not to make the model geographically unaware, but to prevent geographic context from overriding the evidence on which the task prediction should depend.

## 3 METHODOLOGY

Our framework identifies when geographic context affects a prediction beyond what is supported by fixed visual evidence, localizes the representations mediating this effect, and selectively intervenes on them. It consists of four stages: **Audit, Ground, Trace, and Steer**. Causal-layer selection and steering-direction estimation use training data, intervention hyperparameters are selected on validation data, and final evaluation is performed on held-out images.

**Task-Level Prediction Scores** We use constrained candidate scoring to obtain directly comparable predictions across geographic counterfactuals. For a candidate $c$ consisting of $T_c$ tokens, we compute the length-normalized log-likelihood $\ell_\theta(c \mid x, g) = \frac{1}{T_c} \sum_{t=1}^{T_c} \log p_\theta(c_t \mid x, g, c_{<t})$. These scores are normalized over the task-specific candidate set:

$$q_\theta(c \mid x, g) = \frac{\exp\left(\ell_\theta(c \mid x, g)\right)}{\sum_{c' \in \mathcal{C}} \exp\left(\ell_\theta(c' \mid x, g)\right)}. \tag{1}$$

Here, $q_\theta$ is a candidate-normalized score rather than a calibrated probability. For socioeconomic assessment, we use $\mathcal{C}_{\text{soc}} = \{c_{\text{low}}, c_{\text{mid}}, c_{\text{aff}}\}$, corresponding to low-income, middle-income, and affluent, and define $s_{\text{soc}} = q_\theta(c_{\text{low}} \mid x, g)$. For disaster assessment, $\mathcal{C}_{\text{dis}} = \{c_{\text{none}}, c_{\text{minor}}, c_{\text{major}}, c_{\text{dest}}\}$, corresponding to no damage, minor damage, major damage, and destroyed, and define $s_{\text{dis}} = q_\theta(c_{\text{major}} \mid x, g) + q_\theta(c_{\text{dest}} \mid x, g)$. We use $s_\theta$ to denote the corresponding task score.

**Stage I: Audit** For each visual input $x_i$, we construct a pixel-identical geographic counterfactual pair $(x_i, g_a)$ and $(x_i, g_b)$, changing only the stated location. The signed geographic effect is $\Delta_i = s_\theta(x_i, g_a) - s_\theta(x_i, g_b)$. Because the visual evidence is identical, $\Delta_i$ isolates the influence of the geographic cue on the task prediction. We characterize this effect through its magnitude, direction, prediction flips, and distributional change across examples. Geographic sensitivity is not treated as geographic bias.

**Stage II: Ground** We assess whether the geographic effect is supported by task-relevant evidence. For disaster assessment, the physical damage label of a fixed pre-event/post-event pair is invariant to the stated location, $y_i^{\text{damage}}(g_a) = y_i^{\text{damage}}(g_b)$. A prediction change therefore reflects a different interpretation of identical physical evidence. For socioeconomic assessment, geography provide legitimate information. We use the Relative Wealth Index (RWI) (Chi et al., 2022) as external grounding evidence. For each site $i$, convert its matched RWI value $r_i$ to the within-country percentile $\rho_i = \text{PercentileRank}(r_i \mid \text{country}(i))$ and evaluate whether model scores preserve this ordering.

**Stage III: Trace** We identify layers that causally mediate geographic cue effects relative to a cue-absent execution. Let $g_\emptyset$ denote the same prompt with the geographic statement omitted, and let $h_{i,\ell}^g$ and $h_{i,\ell}^\emptyset$ denote the task-relevant representations under $g \in \mathcal{G}$ and $g_\emptyset$. We define $\delta_{i,\ell}^g = h_{i,\ell}^g - h_{i,\ell}^\emptyset$ and $\Delta_i^g = s_\theta(x_i, g) - s_\theta(x_i, g_\emptyset)$. To measure causal relevance, we patch $h_{i,\ell}^g \leftarrow h_{i,\ell}^\emptyset$ and recompute $s_{i,\ell}^{g \leftarrow \emptyset}$. For each geography, $C_\ell^g = \frac{1}{N} \sum_i \left[ |\Delta_i^g| - \left| s_{i,\ell}^{g \leftarrow \emptyset} - s_\theta(x_i, g_\emptyset) \right| \right]$. We macro-average $C_\ell = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} C_\ell^g$ and select $\mathcal{L}^* = \text{TopK}_{\ell \in \{1,\ldots,L\}}(C_\ell)$. The reported relative patching reduction is $R_\ell^{\text{patch}} = \frac{100}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left( 1 - \frac{\sum_i |s_{i,\ell}^{g \leftarrow \emptyset} - s_\theta(x_i, g_\emptyset)|}{\sum_i |\Delta_i^g|} \right)$. Top-$K$ reduction is computed identically after jointly patching all layers in $\mathcal{L}^*$. The cue-absent execution is an operational non-geographic anchor and is not assumed to be intrinsically unbiased.

**Stage IV: Grounded Counterfactual-Guided Steering** We weight each contrast by its cue-induced increase in grounded task error. For disaster assessment, $e_i^{\mathrm{dis}}(g) = |\widehat{D}_i(g) - D_i|$, where $\widehat{D}_i(g) = \sum_{c \in \mathcal{C}_{\mathrm{dis}}} q_\theta(c \mid x_i, g) d(c)$ is the expected ordinal damage severity. For socioeconomic assessment, $e_i^{\mathrm{soc}}(g) = |\mathcal{P}_i|^{-1} \sum_{j \in \mathcal{P}_i} \max\left(0, m - \mathrm{sign}(r_i - r_j)[W_i(g) - W_j(g)]\right)$, where $\mathcal{P}_i$ contains valid same-country comparisons. For either task, we compute $b_i^g = \max\left(0, e_i(g) - e_i(g_\emptyset)\right)$ and $v_\ell^{g,\mathrm{bias}} = \frac{\sum_i b_i^g \delta_{i,\ell}^g}{\sum_i b_i^g + \epsilon}$. Examples for which geography does not increase grounded error therefore receive zero weight. At inference, we apply $\widetilde{h}_\ell(x, g) = h_\ell(x, g) - \lambda v_\ell^{g,\mathrm{bias}}$ for $g \in \mathcal{G}$ and $\ell \in \mathcal{L}^*$, while the cue-absent execution remains unchanged. Let $G_{i,g}^{\mathrm{B}} = |s_\theta(x_i, g) - s_\theta(x_i, g_\emptyset)|$ and $G_{i,g}^{\mathrm{S}} = |s_\theta^{\mathrm{steer}}(x_i, g) - s_\theta(x_i, g_\emptyset)|$. The macro mitigation is

$$M = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left(1 - \frac{\sum_i G_{i,g}^{\mathrm{S}}}{\sum_i G_{i,g}^{\mathrm{B}}}\right), \qquad \sum_i G_{i,g}^{\mathrm{B}} > 0. \tag{2}$$

Higher $M$ indicates stronger mitigation relative to the cue-absent anchor.

**Utility preservation.** We jointly select the number of causal layers $K$ and steering strength $\lambda$ on validation data. We define the normalized task utilities as

$$\widetilde{U}_{\mathrm{soc}}(K, \lambda) = \left\langle \frac{\rho_m(K, \lambda)}{\rho_m(0)} \right\rangle_m, \qquad \widetilde{U}_{\mathrm{dis}}(K, \lambda) = \left\langle \frac{\mathrm{MAE}_m(0)}{\mathrm{MAE}_m(K, \lambda)} \right\rangle_m, \tag{3}$$

where $\langle \cdot \rangle_m$ denotes averaging across models. Both equal 1 without steering and increase with utility. The validation gap is macro-averaged across the four models and two tasks: $\overline{G}_{\mathrm{CF}}^{\mathrm{val}}(K, \lambda) = \frac{1}{8} \sum_{m=1}^{4} \sum_{t \in \{\mathrm{soc}, \mathrm{dis}\}} G_{\mathrm{CF},m,t}^{\mathrm{val}}(K, \lambda)$. We then select

$$(K^*, \lambda^*) = \underset{K \in \mathcal{K},\, \lambda \in \Lambda}{\arg\min} \left[ \overline{G}_{\mathrm{CF}}^{\mathrm{val}}(K, \lambda) + \frac{\beta}{2} \sum_{t \in \{\mathrm{soc}, \mathrm{dis}\}} \max\left(0, 1 - \widetilde{U}_t(K, \lambda)\right) \right]. \tag{4}$$

This penalizes mitigation that degrades either task. Algorithm 1 summarizes the complete procedure.

**Algorithm 1** Cue-Absent Counterfactual-Guided Causal Steering

**Require:** VLM $f_\theta$, datasets $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{val}}$, cues $\mathcal{G}$, cue-absent condition $g_\emptyset$, grounded error $e_i(\cdot)$
**Ensure:** Causal layers $\mathcal{L}^*$, directions $\{v_\ell^{g,\mathrm{bias}}\}$, and strength $\lambda^*$

**for** each $x_i \in \mathcal{D}_{\mathrm{tr}}$ **do**
  Store $(s_i^c, \{h_{i,\ell}^c\}_\ell)$ from $f_\theta(x_i, c)$ for all $c \in \mathcal{G} \cup \{g_\emptyset\}$
  For each $g \in \mathcal{G}$, set $\Delta_i^g = s_i^g - s_i^\emptyset$, $\delta_{i,\ell}^g = h_{i,\ell}^g - h_{i,\ell}^\emptyset$, and $b_i^g = \max\left(0, e_i(g) - e_i(g_\emptyset)\right)$
**end for**
**for** each layer $\ell$ and $g \in \mathcal{G}$ **do**
  Patch $h_{i,\ell}^g \leftarrow h_{i,\ell}^\emptyset$ for all $i$ and obtain $s_{i,\ell}^{g \leftarrow \emptyset}$
  Compute $C_\ell^g = \frac{1}{N} \sum_i \left[ |\Delta_i^g| - |s_{i,\ell}^{g \leftarrow \emptyset} - s_i^\emptyset| \right]$
  Compute $v_\ell^{g,\mathrm{bias}} = \frac{\sum_i b_i^g \delta_{i,\ell}^g}{\sum_i b_i^g + \epsilon}$
**end for**
Set $C_\ell = |\mathcal{G}|^{-1} \sum_{g \in \mathcal{G}} C_\ell^g$; select $(K^*, \lambda^*)$ on $\mathcal{D}_{\mathrm{val}}$ and let $\mathcal{L}^*$ be the $K^*$ highest-scoring layers
**for** each test input $(x, g)$ **do**
  If $g \in \mathcal{G}$, apply $\widetilde{h}_\ell = h_\ell - \lambda^* v_\ell^{g,\mathrm{bias}}$ for $\ell \in \mathcal{L}^*$; otherwise leave activations unchanged
  Evaluate the prediction
**end for**

### 3.1 Theoretical Analysis: From Encoded Geography to Grounded Causal Steering

Our analysis formalizes why geographic information should not be suppressed merely because it is represented internally. Counterfactual representation differences can support steering (Park et al., 2023), while causal abstraction provides a framework for interpreting representation-level interventions (Geiger et al., 2025). Our key distinction is between *encoding* geography and *causally using* it for the task prediction. For image $x_i$, geography $g \in \mathcal{G}$, and layer $\ell$, define $\delta^g_{i,\ell} = h^g_{i,\ell} - h^{\emptyset}_{i,\ell}$ and $\Delta^g_i = s_\theta(x_i, g) - s_\theta(x_i, g_\emptyset)$.

**Proposition 1 (Intervention-based geographic decomposition).** The geography-specific cue effect decomposes exactly as $\Delta^g_i = M^g_{i,\ell} + R^g_{i,\ell}$, where $M^g_{i,\ell} = s_\theta(x_i, g) - s^{g \leftarrow \emptyset}_{i,\ell}$ is the effect removed by replacing the geography-conditioned representation with its cue-absent counterpart, and $R^g_{i,\ell} = s^{g \leftarrow \emptyset}_{i,\ell} - s_\theta(x_i, g_\emptyset)$ is the residual cue effect. Under differentiability, $M^g_{i,\ell}$ depends on the component of $\delta^g_{i,\ell}$ aligned with downstream task sensitivity. Geography therefore be strongly encoded at a layer without causally affecting the task prediction. This motivates selecting intervention locations by causal effect rather than representation magnitude alone. Generic steering typically treats representation contrasts as equally informative. Our construction instead couples causal localization with grounded harm: activation patching determines *where* geography affects the prediction, while grounded task error determines *which* counterfactual contrasts define the intervention direction.

**Proposition 2 (Grounded harm-weighted consensus direction).** For $b^g_i = \max(0, e_i(g) - e_i(g_\emptyset))$ and $\epsilon > 0$, the direction $v^{g,\text{bias}}_\ell = \frac{\sum_i b^g_i \delta^g_{i,\ell}}{\sum_i b^g_i + \epsilon}$ is the unique minimizer of $\sum_i b^g_i \|\delta^g_{i,\ell} - v\|_2^2 + \epsilon \|v\|_2^2$. It is therefore a regularized consensus of representation changes associated with increased grounded error, rather than an average over geographic variation irrespective of task harm.

**Proposition 3 (Local steering reduction under directional alignment).** Let $\phi^g_i(\lambda)$ denote the task score after jointly steering the selected layers and define $\widetilde{\Delta}^g_i(\lambda) = \phi^g_i(\lambda) - s_\theta(x_i, g_\emptyset)$. For $\Delta^g_i \neq 0$, assume $\text{sign}(\Delta^g_i)\phi^{g\prime}_i(0) \leq -\gamma$ for $\gamma > 0$ and $|\phi^{g\prime\prime}_i(t)| \leq B$. Before a sign reversal, $|\widetilde{\Delta}^g_i(\lambda)| \leq |\Delta^g_i| - \lambda\gamma + \frac{B\lambda^2}{2}$. Thus, under directional alignment and bounded curvature, sufficiently small positive $\lambda$ reduces the absolute cue effect. If grounded utility is locally insensitive along the same path, its change is correspondingly bounded.

Together, these results formalize the three components of grounded causal steering: intervention-based localization, harm-weighted direction estimation, and locally controlled suppression of cue-induced effects. The propositions provide characterization and sufficient conditions rather than a guarantee for arbitrary models; effectiveness of the estimated directions is therefore established empirically through the activation-patching, ablation, and steering experiments. Full proofs are provided in Appendix G.

## 4 Experiments

Figure 2 provides representative held-out examples from both evaluation settings. In each case, the visual evidence is fixed while the geographic cue changes the model prediction. Counterfactual-guided steering substantially reduces this shift and moves the prediction back toward an evidence-consistent response.

We evaluate two complementary remote-sensing reasoning tasks: socioeconomic assessment from single-date imagery and disaster-damage assessment from paired pre-/post-event imagery. The former may legitimately use geographic priors, whereas physical damage should be invariant to the stated location. We evaluate four frozen generative VLMs: Gemma-4-31B (Team et al., 2026), Qwen2.5-VL-7B-Instruct (Team, 2025), GeoChat (Kuckreja et al., 2024), and RSCoVLM-7B (Li et al., 2026). For socioeconomic assessment, we use residential fMoW examples (Christie et al., 2018), balancing Africa, the Americas, Asia, and Europe by downsampling each region to the least represented group. African sites with valid measurements are matched to RWI (Chi et al., 2022)

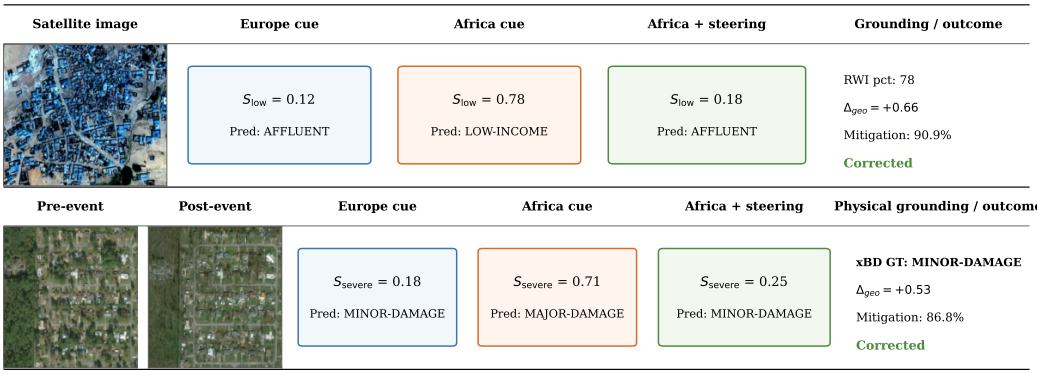



Figure 2: **Qualitative examples of counterfactual-guided steering.** Top: socioeconomic assessment on fMoW. Bottom: disaster-damage assessment on xBD. Changing only the geographic cue shifts predictions under fixed visual evidence, while steering reduces this effect. RWI percentile and xBD damage labels provide independent grounding.

for external grounding. For disaster assessment, we use xBD/xView2 (Gupta et al., 2019), similarly balancing damage classes. Our analysis contrasts Africa and Europe using the mean absolute counterfactual gap $G_{\mathrm{CF}}$. Further details are provided in C. Table 1 shows consistent geographic effects across all models and tasks. Under identical imagery, changing the cue from Europe to Africa increases the LOW−INCOME score by 0.151-0.356 and the severe-damage score by 0.058-0.118. For Gemma-4-31B, all socioeconomic examples shift in this direction and 35.8% of discrete predictions flip. Thus, stated geography can directly alter predictions even when the visual evidence is fixed.

Table 1: **Pixel-identical geographic counterfactual evaluation.** Only the stated geographic context changes between Africa and Europe. $\Delta_{\mathrm{LOW}}$ is the signed change in the candidate-normalized LOW−INCOME score, and $\Delta_{\mathrm{sev}}$ is the signed change in the severe-damage score. CF Gap measures absolute geographic sensitivity; Dir. is the fraction following the Africa > Europe direction.

| | Socioeconomic | | | | Disaster Damage | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | $\Delta_{\mathrm{LOW}}$ | **CF Gap** | **Dir. (%)** | **Flip (%)** | $\Delta_{\mathrm{sev}}$ | **CF Gap** | **Dir. (%)** | **Flip (%)** |
| Gemma-4-31B | **+0.356** | **0.356** | **100.0** | 35.8 | **+0.118** | **0.135** | **84.0** | 17.0 |
| Qwen2.5-VL-7B | +0.224 | 0.258 | 84.7 | 23.5 | +0.091 | 0.112 | 79.0 | 13.0 |
| GeoChat | +0.173 | 0.211 | 81.3 | 18.6 | +0.071 | 0.094 | 75.0 | 10.0 |
| RSCoVLM-7B | +0.151 | 0.187 | 78.4 | 15.9 | +0.058 | 0.081 | 72.0 | 8.0 |

The effect remains consistent across additional geographic pairs and semantically equivalent prompts; results are reported in Appendix D. Table 2 evaluates whether these counterfactual effects are supported by independent task evidence. For socioeconomic assessment, alignment with within-country RWI is modest, with Spearman correlations of 0.21-0.33. Moreover, 39.8%-55.2% of relatively high-RWI locations are still classified as LOW−INCOME, where classification is determined by the highest-scoring candidate. Thus, strong geographic sensitivity is not consistently supported by measured socioeconomic variation. For disaster assessment, models retain meaningful agreement with xBD damage annotations, with severity correlations of 0.54-0.66. The geographic counterfactual shifts therefore occur despite preserved damage-recognition ability, indicating that stated location can alter the interpretation of fixed physical evidence.

Table 3 shows that geographic effects concentrate in late transformer layers. Top-1 and Top-$K$ causal patching reduce the counterfactual gap by 72.6% to 91.2% and 85.3% to 96.0%, respectively, compared with 10.8% to 14.2% for random layers. On held-out test images, steering reduces the gap by 65.2% to 81.9%, improves 82.6% to 100% of examples, and reduces prediction flips by 58.7% to 84.1%, while preserving task utility. All variants use the same data, conditional cue gate, and validation objective. Standard CAA (Rimsky et al., 2024) uses unweighted mean-difference directions with layers and strength selected on validation data, without causal patching. Random direction uses norm-matched Gaussian vectors at the selected causal layers, while random layers

Table 2: **Grounding against independent task evidence.** Socioeconomic predictions are evaluated against within-country RWI; $A_{\mathrm{RWI}}$ denotes pairwise concordance between relatively high- and low-wealth locations. High-RWI LOW reports the percentage of relatively high-RWI locations for which `LOW-INCOME` is the highest-scoring candidate. Disaster predictions are evaluated against physical damage annotations from xBD.

| | Socioeconomic Grounding | | | Damage Grounding | |
|---|---|---|---|---|---|
| **Model** | **RWI $\rho\uparrow$** | **$\mathbf{A}_{\mathrm{RWI}}\uparrow$** | **High-RWI LOW (%) $\downarrow$** | **MAE $\downarrow$** | **$\boldsymbol{\rho}\uparrow$** |
| Gemma-4-31B | 0.21 | 0.64 | 55.2 | 0.214 | 0.54 |
| Qwen2.5-VL-7B | 0.25 | 0.67 | 49.6 | 0.196 | 0.58 |
| GeoChat | 0.29 | 0.69 | 44.1 | 0.181 | 0.62 |
| RSCoVLM-7B | **0.33** | **0.72** | **39.8** | **0.169** | **0.66** |

use the proposed direction at three layers sampled uniformly without replacement using seed 42. Activation-difference selection uses the layer maximizing the mean $\ell_2$ representation difference. Normalized direction unit-normalizes each steering vector. Uniform weighting sets $b_i = 1$, and single-layer steering uses only the highest-ranked causal layer. Table 4 shows that random directions, random layers, and activation-difference selection achieve only $4.5\%$, $11.7\%$, and $28.5\%$ mitigation. Standard CAA reaches $38.5\%$, compared with $53.6\%$ under causal selection with uniform weighting and $60.3\%$ using one causal layer. The full method reaches a gap of $0.048$ ($73.3\%$ mitigation) and improves $89.9\%$ of examples while approximately preserving utility. This isolates the benefits of causal localization, bias weighting, and multi-layer steering.

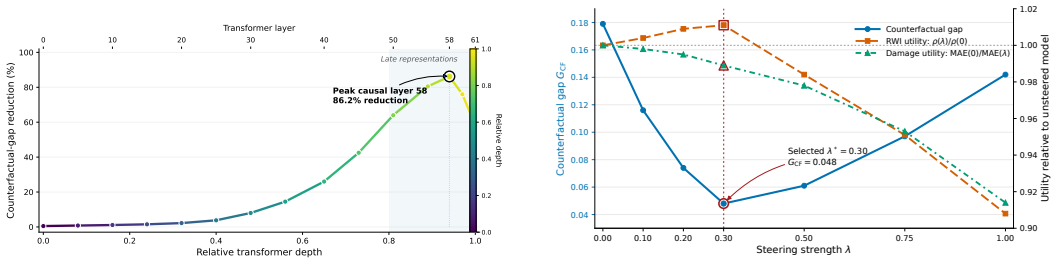
Table 3: **Causal localization and counterfactual-guided steering.** Causal-localization results are computed on the training data used for layer selection, while steering is evaluated on held-out test images. Relative gap reduction measures the aggregate counterfactual-gap reduction after activation patching, with Top-$K$ using joint patching of the selected causal layers. Socioeconomic utility is measured by RWI Spearman correlation, while disaster utility is measured by damage MAE.

| | Gemma-4-31B | | Qwen2.5-VL-7B | | GeoChat | | RSCoVLM-7B | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **Socio.** | **Disaster** | **Socio.** | **Disaster** | **Socio.** | **Disaster** | **Socio.** | **Disaster** |
| **Causal Localization** | | | | | | | | |
| Peak Relative Depth | 0.94 | 0.97 | 0.89 | 0.93 | 0.88 | 0.94 | 0.89 | 0.93 |
| Top-1 Relative Gap Reduction (%) $\uparrow$ | 86.2 | **91.2** | 78.5 | 83.4 | 74.1 | 80.3 | 72.6 | 78.8 |
| Top-$K$ Relative Gap Reduction (%) $\uparrow$ | 94.1 | **96.0** | 89.7 | 92.1 | 86.8 | 90.5 | 85.3 | 88.9 |
| Random-Layer Relative Gap Reduction (%) | 12.4 | 10.8 | 13.6 | 11.9 | 14.2 | 12.7 | 13.1 | 11.4 |
| **Counterfactual-Guided Steering** | | | | | | | | |
| $G_{\mathrm{CF}}^{\mathrm{base}}$ | 0.356 | 0.135 | 0.258 | 0.112 | 0.211 | 0.094 | 0.187 | 0.081 |
| $G_{\mathrm{CF}}^{\mathrm{steer}}\downarrow$ | 0.077 | 0.024 | 0.071 | 0.026 | 0.067 | 0.027 | 0.065 | 0.025 |
| Mitigation (%) $\uparrow$ | 78.4 | **81.9** | 72.5 | 76.8 | 68.2 | 71.3 | 65.2 | 69.1 |
| Images Improved (%) $\uparrow$ | 92.4 | **100.0** | 88.5 | 94.2 | 85.1 | 89.7 | 82.6 | 86.4 |
| Utility Before | $\rho = 0.21$ | 0.214 | $\rho = 0.25$ | 0.196 | $\rho = 0.29$ | 0.181 | $\rho = 0.33$ | 0.169 |
| Utility After | $\rho = \mathbf{0.24}$ | **0.207** | $\rho = \mathbf{0.28}$ | **0.191** | $\rho = \mathbf{0.31}$ | **0.178** | $\rho = \mathbf{0.34}$ | **0.167** |

Table 4: **Comparison with generic activation steering and component ablations.** CF Gap and Images Improved are macro-averaged across the eight model-task settings. Mitigation is computed from the unrounded macro-averaged gaps. Utility changes are macro-averaged across models within each task, where $\Delta\rho_{\mathrm{RWI}} = \rho_{\mathrm{steer}} - \rho_{\mathrm{base}}$ and $\Delta\mathrm{MAE} = \mathrm{MAE}_{\mathrm{steer}} - \mathrm{MAE}_{\mathrm{base}}$. All hyperparameters are selected on validation data. Standard CAA uses validation-selected mean-difference directions without causal localization.

| **Method** | **CF Gap $\downarrow$** | **Mitigation (%) $\uparrow$** | **Images Improved (%) $\uparrow$** | **$\Delta\rho_{\mathrm{RWI}}\uparrow$** | **$\Delta\mathrm{MAE}\downarrow$** |
|---|---|---|---|---|---|
| No steering | 0.179 | 0.0 | N/A | 0.000 | 0.000 |
| Standard CAA (Rimsky et al., 2024) | 0.110 | 38.5 | 70.5 | $+0.004$ | $+0.001$ |
| Random direction | 0.171 | 4.5 | 51.8 | $-0.007$ | $+0.004$ |
| Random layers | 0.158 | 11.7 | 57.3 | $-0.005$ | $+0.005$ |
| Largest activation-difference layer | 0.128 | 28.5 | 66.4 | $-0.002$ | $+0.003$ |
| Normalized direction | 0.096 | 46.4 | 74.6 | $+0.011$ | $-0.001$ |
| Causal layers + uniform weighting ($b_i = 1$) | 0.083 | 53.6 | 78.9 | $+0.019$ | $-0.003$ |
| Single causal layer | 0.071 | 60.3 | 82.1 | $+0.014$ | $-0.002$ |
| **Full causal steering** | **0.048** | **73.3** | **89.9** | $\mathbf{+0.023}$ | $\mathbf{-0.004}$ |

We further evaluate SigLIP2 and RemoteCLIP to test whether geographic sensitivity arises in frozen vision-language representations. Task-specific text prototypes are scored by image-text similar-




**(a) Layerwise causal localization** **(b) Steering strength and utility trade-off**

Figure 3: **Tracing and controlling geographic influence.** (a) Single-layer activation patching for Gemma on socioeconomic assessment. Causal effects concentrate in late layers and peak at layer 58 (relative depth 0.94). (b) Validation-set sweep of the mean counterfactual gap and task-specific utility. RWI utility is normalized as $\rho(\lambda)/\rho(0)$, while damage utility is normalized as $\mathrm{MAE}(0)/\mathrm{MAE}(\lambda)$, so higher values are better for both. Ratios are computed per model and then averaged. The selected $\lambda^* = 0.30$ substantially reduces geographic sensitivity while keeping both utilities within 1.1% of their unsteered values.

ity and normalized across candidates under *No Geography*, *Geography-Conditioned*, and *Shuffled Geography* conditions. Table 5 shows counterfactual gaps and flip rates of 0.158 and 14.1% for SigLIP2, compared with 0.104 and 8.9% for RemoteCLIP. Removing geography provides the strongest RWI grounding, with correlations of 0.28 and 0.34, while shuffling cues weakens it. Thus, remote-sensing specialization reduces but does not eliminate geographic influence at the representation level. Figure 3(a) localizes the effect to late layers, peaking at relative depth 0.94, while Figure 3(b) shows the validation steering sweep. At $\lambda^* = 0.30$, the gap decreases from 0.179 to 0.048, with normalized RWI and damage utilities of 1.011 and 0.989. Larger strengths increase the gap and degrade both utilities, indicating over-steering. Further analyses appear in D, E, and F.

Table 5: **Frozen dual-encoder geographic ablation.** SigLIP2 and RemoteCLIP are evaluated with no geographic cue, geography-conditioned prototypes, and shuffled geographic cues. We report geographic counterfactual sensitivity and socioeconomic grounding against within-country RWI. Lower sensitivity and higher grounding are better.

| | | Geographic Counterfactual Sensitivity | | | | | Socioeconomic Grounding | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Prompt Condition** | $\Delta_{\mathrm{LOW}}$ | $\mathrm{G}_{\mathrm{CF}}\downarrow$ | **Dir. (%)** | **Flip (%)** $\downarrow$ | **JS** $\downarrow$ | **RWI** $\rho\uparrow$ | $\mathrm{A}_{\mathrm{RWI}}\uparrow$ |
| SigLIP2 | No Geography | – | – | – | – | – | **0.28** | **0.69** |
| | Geography-Conditioned | +0.142 | 0.158 | 79.3 | 14.1 | 0.052 | 0.25 | 0.66 |
| | Shuffled Geography | +0.131 | **0.151** | 76.8 | **12.7** | **0.048** | 0.20 | 0.61 |
| RemoteCLIP | No Geography | – | – | – | – | – | **0.34** | **0.74** |
| | Geography-Conditioned | +0.087 | 0.104 | 71.5 | 8.9 | 0.034 | 0.32 | 0.72 |
| | Shuffled Geography | +0.079 | **0.098** | 69.6 | **8.1** | **0.031** | 0.27 | 0.67 |

## 5 Conclusion and Future Work

We study when stated geography becomes an unsupported decision shortcut in remote-sensing VLMs. Pixel-identical counterfactuals isolate geographic influence from visual variation, while RWI and damage annotations provide task-specific grounding. Socioeconomic shifts are not consistently supported by measured RWI variation, whereas disaster shifts conflict with the physical invariance of observed damage. We causally localize these effects through activation patching and show that counterfactual-guided steering reduces them on held-out images without retraining or degrading task utility. To our knowledge, this is the first framework to unify controlled geographic auditing, causal localization, and selective inference-time steering in remote-sensing VLMs. **Future work** should study finer geographic cues, broader tasks and models, example-specific interventions, and the emergence of geographic shortcuts across training stages.

### AI USE STATEMENT

In this work, generative AI tools were used for text polishing and assistance with code writing. All AI-assisted text was reviewed and edited by the authors, and all generated or modified code was manually inspected and tested for correctness. The authors take full responsibility for the final content, implementation, results, and claims presented in this work.

### ETHICS STATEMENT

This work studies geographic bias in remote-sensing vision-language models, including predictions related to socioeconomic conditions and disaster damage. Such predictions may have societal consequences if used in applications such as resource allocation, poverty mapping, or disaster response. Our objective is to expose and reduce unsupported geographic influence, not to infer or assign intrinsic socioeconomic or geographic properties to regions or populations. In particular, the geographic counterfactuals used in this work are controlled diagnostic interventions and should not be interpreted as characterizations of the corresponding regions.

Our mitigation method reduces measured counterfactual geographic effects but does not guarantee the absence of geographic, socioeconomic, or other forms of bias.

### REPRODUCIBILITY STATEMENT

We provide the methodological and implementation details necessary to reproduce the counterfactual audit, causal localization, and steering procedure in Section 3. Model and inference settings, dataset construction and preprocessing, geographic-counterfactual construction, task definitions, evaluation metrics, train/validation/test splits, software, hardware, and random seeds are provided in Appendix C. Additional geographic-pair and prompt-robustness analyses are reported in Appendix D, while steering, utility preservation, statistical validation, and cross-geography generalization are provided in Appendix E. Complete experimental metrics are reported in Appendix F, and all assumptions and proofs are provided in Appendix G. Together, these materials specify the data processing, evaluation, intervention, and statistical procedures used throughout the study.

## REFERENCES

Emily Aiken, Esther Rolf, and Joshua Blumenstock. Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy. *arXiv preprint arXiv:2305.01783*, 2023.

Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E Blumenstock. Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119 (3):e2113658119, 2022.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6172–6180. IEEE, 2018.

Stef De Sabbata, Rahul Baiju, Stefano Mizzaro, and Kevin Roitero. Exploring geographic relative space in large language models through activation patching. *arXiv preprint arXiv:2605.14535*, 2026.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11975–11985. IEEE, 2024.

Phillip Howard, Kathleen C Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. Uncovering bias in large vision-language models at scale with counterfactuals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5946–5991, 2025.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794, 2016.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840. IEEE, 2024.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Qingyun Li, Shuran Ma, Junwei Luo, Yi Yu, Yue Zhou, Fengxiang Wang, Xudong Lu, Xiaoxing Wang, Xin He, Yushi Chen, et al. Co-training vision-language models for remote sensing multi-task learning. *Remote Sensing*, 18(2):222, 2026.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens Van Der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137, 2023.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *Advances in Neural Information Processing Systems*, 38, 2026.

Zhangyu Wang, Nemin Wu, Qian Cao, Jiangnan Xia, Zeping Liu, Yiqun Xie, Akshay Nambi, Tanuja Ganu, Ni Lao, Ninghao Liu, et al. Geobs: Information-theoretic quantification of geographic bias in ai models. *arXiv preprint arXiv:2509.23482*, 2025.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

Miao Zhang, Harvineet Singh, Lazarus Chok, and Rumi Chunara. Segmenting across places: The need for fair transfer learning with satellite imagery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2915–2924. IEEE, 2022.

# A Qualitative Counterfactual Steering Examples

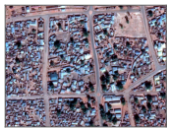
| Satellite image | Europe cue | Africa cue | Africa + steering | Grounding / outcome |
|---|---|---|---|---|
|  | $P(\text{LOW}) = 0.12$<br>Pred: AFFLUENT | $P(\text{LOW}) = 0.78$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.18$<br>Pred: AFFLUENT | RWI pct: 78<br>$\Delta_{geo} = +0.66$<br>Mitigation: 90.9%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.21$<br>Pred: MIDDLE-INCOME | $P(\text{LOW}) = 0.72$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.29$<br>Pred: MIDDLE-INCOME | RWI pct: 63<br>$\Delta_{geo} = +0.51$<br>Mitigation: 84.3%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.09$<br>Pred: AFFLUENT | $P(\text{LOW}) = 0.81$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.16$<br>Pred: AFFLUENT | RWI pct: 82<br>$\Delta_{geo} = +0.72$<br>Mitigation: 90.3%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.31$<br>Pred: MIDDLE-INCOME | $P(\text{LOW}) = 0.69$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.45$<br>Pred: MIDDLE-INCOME | RWI pct: 47<br>$\Delta_{geo} = +0.38$<br>Mitigation: 63.2%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.16$<br>Pred: AFFLUENT | $P(\text{LOW}) = 0.86$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.63$<br>Pred: LOW-INCOME | RWI pct: 28<br>$\Delta_{geo} = +0.70$<br>Mitigation: 32.9%<br>**Partially corrected** |
|  | $P(\text{LOW}) = 0.18$<br>Pred: AFFLUENT | $P(\text{LOW}) = 0.74$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.27$<br>Pred: AFFLUENT | RWI pct: 71<br>$\Delta_{geo} = +0.56$<br>Mitigation: 83.9%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.26$<br>Pred: MIDDLE-INCOME | $P(\text{LOW}) = 0.67$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.34$<br>Pred: MIDDLE-INCOME | RWI pct: 59<br>$\Delta_{geo} = +0.41$<br>Mitigation: 80.5%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.11$<br>Pred: AFFLUENT | $P(\text{LOW}) = 0.76$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.22$<br>Pred: AFFLUENT | RWI pct: 75<br>$\Delta_{geo} = +0.65$<br>Mitigation: 83.1%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.34$<br>Pred: MIDDLE-INCOME | $P(\text{LOW}) = 0.71$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.50$<br>Pred: MIDDLE-INCOME | RWI pct: 44<br>$\Delta_{geo} = +0.37$<br>Mitigation: 56.8%<br>**Corrected** |
|  | $P(\text{LOW}) = 0.19$<br>Pred: AFFLUENT | $P(\text{LOW}) = 0.79$<br>Pred: LOW-INCOME | $P(\text{LOW}) = 0.58$<br>Pred: LOW-INCOME | RWI pct: 33<br>$\Delta_{geo} = +0.60$<br>Mitigation: 35.0%<br>**Partially corrected** |

Figure 4: **Qualitative socioeconomic counterfactual steering on held-out fMoW images.** Each row uses the same satellite image under a Europe cue, an Africa cue, and the Africa cue after counterfactual-guided steering. We report the candidate-normalized LOW−INCOME score, discrete prediction, within-country RWI percentile, signed geographic effect $\Delta_{\text{geo}}$, and relative mitigation. Successful cases show that steering can substantially reduce geography-induced prediction shifts while leaving the visual evidence unchanged; partial corrections illustrate remaining failure modes.

Figure 4 provides qualitative examples of counterfactual-guided steering on held-out fMoW residential imagery. For each image, the visual input is kept pixel-identical while the stated geographic cue is changed from Europe to Africa. We report the resulting `LOW-INCOME` score and discrete prediction before and after steering, together with the matched within-country RWI percentile used for independent socioeconomic grounding.

Across the examples, changing only the geographic cue can substantially alter the model prediction despite unchanged visual evidence. Applying the learned steering intervention generally moves the Africa-conditioned prediction back toward the corresponding Europe-conditioned prediction while retaining the same image content. The examples also include partial corrections, illustrating that the intervention reduces rather than universally eliminates geographic sensitivity.

# B Qualitative Disaster Counterfactual Steering

Figure 5 presents qualitative examples from the disaster-damage setting using held-out pre-event and post-event image pairs. For each example, the visual evidence is fixed while only the stated geographic cue is changed from Europe to Africa. We report the severe-damage score, defined as the combined candidate-normalized score for $c_{\mathrm{major}}$ and $c_{\mathrm{dest}}$, together with the resulting discrete damage prediction before and after steering.

Because the pre-event and post-event imagery are identical across geographic counterfactuals, any prediction change induced by the geographic cue cannot be explained by a change in physical damage. The examples show that the Africa cue can substantially increase predicted damage severity under fixed visual evidence, while counterfactual-guided steering typically moves the prediction back toward the Europe-conditioned result. Ground-truth xBD damage annotations provide an external reference for whether the geographic shift is supported by the observed physical evidence. We also include partial corrections to illustrate remaining failure cases.

# C Experimental Details

## C.1 Model and Inference Details

We evaluate models from multiple architecture and pretraining families to distinguish model-specific behavior from systematic geographic effects. Our primary suite contains four generative VLMs, including general-purpose multimodal instruction models and models developed specifically for remote sensing. We additionally evaluate SigLIP2 and RemoteCLIP as dual-encoder controls to test whether geographic sensitivity also appears outside autoregressive generative VLMs. The complete model suite is summarized in Table 6.

For generative VLMs, we access token-level logits and evaluate the constrained candidate scores defined in Eq. 1. SigLIP2 and RemoteCLIP are evaluated using their image–text similarity scores under the control protocol described in the corresponding experiment. We use deterministic inference throughout, and all model parameters remain frozen during auditing, activation patching, steering, and dual-encoder evaluation.

Table 6: **Complete model suite.** "General" denotes general-purpose multimodal instruction models, "RS" denotes models developed or adapted for remote sensing, and "Dual Encoder" denotes contrastive image-text encoders used as additional controls.

| Model | Type | Scale | Role |
|---|---|---|---|
| Gemma-4-31B (Team et al., 2026) | General | 31B | Generative VLM |
| Qwen2.5-VL-7B-Instruct (Team, 2025) | General | 7B | Generative VLM |
| GeoChat (Kuckreja et al., 2024) | RS | 7B | Generative VLM |
| RSCoVLM-7B (Li et al., 2026) | RS | 7B | Generative VLM |
| SigLIP2 (Tschannen et al., 2025) | Dual Encoder | – | Control |
| RemoteCLIP (Liu et al., 2024) | RS Dual Encoder | – | Control |




Figure 5: **Qualitative disaster-damage counterfactual steering on held-out xBD examples.** Each row contains the same pre-event and post-event imagery evaluated under a Europe cue, an Africa cue, and the Africa cue after counterfactual-guided steering. We report the severe-damage score $s_{\text{dis}} = q_\theta(c_{\text{major}} \mid x, g) + q_\theta(c_{\text{dest}} \mid x, g)$, the corresponding discrete prediction, ground-truth xBD damage class, signed geographic effect $\Delta_{\text{geo}}$, and relative mitigation. Successful examples show that steering can reduce geography-induced severity shifts while preserving the fixed physical evidence, while partial corrections highlight cases in which residual geographic sensitivity remains.

### C.2 Dataset Construction and Preprocessing

We use complementary datasets chosen to separate geographic association from visually grounded task evidence.

**Socioeconomic imagery.** We use single-unit residential imagery from Functional Map of the World (fMoW) (Christie et al., 2018), retaining one observation per physical site. To obtain a uni-

form geographic distribution, we group eligible sites into Africa, the Americas, Asia, and Europe and randomly downsample each region to match the number of sites in the least represented region.

**External wealth grounding.** For African residential locations, we spatially match fMoW sites to the Relative Wealth Index (RWI) (Chi et al., 2022). Because RWI values are primarily interpretable relative to other locations within a country, we convert each score to a within-country percentile. The matched subset contains 226 residential sites with valid RWI measurements.

**Disaster assessment.** We use xBD/xView2 (Gupta et al., 2019), which provides paired pre-event and post-event satellite imagery with building-level damage annotations. We retain pairs for which both images are readable and the event identifier, geographic metadata, and damage annotations are valid.

Splitting is performed before balancing using disaster events as groups. All image pairs from the same event are assigned to a single split, preventing event-level and neighboring-scene overlap between training, validation, and test data. Event groups are allocated to approximate the $80\%/10\%/10\%$ split while preserving the damage-category distribution.

Each retained pair is assigned one category by mapping its continuous image-level severity $D_i$ to the nearest category on the ordinal scale defined in Appendix C.6. Within each split, categories are independently downsampled to the size of the least represented category. Continuous $D_i$ values are retained for damage-grounding evaluation. The identical pre-event and post-event pair is reused under every geographic counterfactual.

Table 7 summarizes the role of each dataset.

Table 7: Datasets and their role in the evaluation.

| Dataset | Visual Input | Task | Grounding |
|---|---|---|---|
| fMoW-Residential | Single image | Socioeconomic | Counterfactual |
| fMoW-RWI | Single image | Socioeconomic | RWI |
| xBD/xView2 | Pre/post pair | Damage | Building damage |

### C.3 Geographic Counterfactual Construction

For each image, we vary only the geographic statement while holding the visual input, task instruction, candidate set, and inference configuration fixed. A geography-conditioned prompt includes

*Images are from Africa.*

The cue-absent execution omits this sentence entirely. We denote this intervention anchor by $g_\emptyset$, with score $s_\theta(x_i, g_\emptyset)$ and representation $h^{\emptyset}_{i,\ell}$. For each $g \in \mathcal{G}$, we define $\Delta^g_i = s_\theta(x_i, g) - s_\theta(x_i, g_\emptyset)$ and $\delta^g_{i,\ell} = h^g_{i,\ell} - h^{\emptyset}_{i,\ell}$. This anchor is an operational reference and is not assumed to be unbiased or ground truth.

We consider

$$\mathcal{G} = \{\text{Africa}, \text{Americas}, \text{Asia}, \text{Europe}\}. \tag{5}$$

We additionally evaluate the neutral control

*Location information is unavailable.*

to distinguish cue absence from neutral wording effects. Direct geographic swaps are retained for auditing, whereas steering directions are estimated relative to $g_\emptyset$. We also test semantic paraphrases, cue position, and cue ordering for prompt robustness.

### C.4 Cue Validity and Geographic Context Baselines

We compare four unsteered prompting conditions: no geographic sentence, the neutral statement *"Location information is unavailable."*, the authentic geographic cue, and an incorrect cue. Incorrect results average the three incorrect-cue evaluations. For RWI correlation, each incorrect cue is evaluated separately before averaging.

Table 8: **Cue-validity baselines across models.** Each entry reports the point estimate and 95% confidence interval from 5,000 bootstrap replicates. The final row for each model reports the paired difference between neutral and no-cue prompting. RWI metrics use the held-out African subset of the RWI-matched cohort ($N = 180$), with resampling stratified by country. $\mathrm{RankErr} = 1 - A_{\mathrm{RWI}}$ is the within-country pairwise-ranking error. Disaster metrics use 1,500 held-out sites.

| | | Disaster (xBD) | | | Socioeconomic (fMoW) | |
|---|---|---|---|---|---|---|
| **Model** | **Cue Condition** | **MAE ↓** | **Accuracy (%) ↑** | **Brier ↓** | $\rho_{\mathrm{RWI}}$ ↑ | **RankErr ↓** |
| Gemma-4-31B | No cue | 0.174 [0.165, 0.183] | 78.4 [76.3, 80.5] | 0.288 [0.274, 0.302] | 0.32 [0.27, 0.37] | 0.276 [0.248, 0.304] |
| | Neutral | 0.176 [0.167, 0.185] | 78.1 [76.0, 80.2] | 0.291 [0.277, 0.305] | 0.31 [0.26, 0.36] | 0.279 [0.251, 0.307] |
| | Authentic | 0.164 [0.155, 0.173] | 80.3 [78.2, 82.4] | 0.272 [0.258, 0.286] | 0.28 [0.23, 0.33] | 0.315 [0.287, 0.343] |
| | Incorrect | 0.202 [0.192, 0.212] | 74.7 [72.5, 76.9] | 0.332 [0.316, 0.348] | 0.23 [0.18, 0.28] | 0.358 [0.330, 0.386] |
| | Neutral − No cue | +0.002 [−0.002, +0.006] | −0.3 [−1.0, +0.4] | +0.003 [−0.002, +0.008] | −0.01 [−0.03, +0.02] | +0.003 [−0.007, +0.013] |
| Qwen2.5-VL-7B | No cue | 0.184 [0.175, 0.193] | 77.1 [75.0, 79.2] | 0.304 [0.290, 0.318] | 0.29 [0.24, 0.34] | 0.294 [0.266, 0.322] |
| | Neutral | 0.184 [0.175, 0.193] | 77.0 [74.9, 79.1] | 0.305 [0.291, 0.319] | 0.29 [0.24, 0.34] | 0.293 [0.265, 0.321] |
| | Authentic | 0.176 [0.167, 0.185] | 78.9 [76.8, 81.0] | 0.291 [0.277, 0.305] | 0.27 [0.22, 0.32] | 0.319 [0.291, 0.347] |
| | Incorrect | 0.207 [0.197, 0.217] | 74.1 [71.9, 76.3] | 0.343 [0.327, 0.359] | 0.22 [0.17, 0.27] | 0.366 [0.338, 0.394] |
| | Neutral − No cue | 0.000 [−0.004, +0.004] | −0.1 [−0.7, +0.5] | +0.001 [−0.003, +0.005] | 0.00 [−0.02, +0.02] | −0.001 [−0.010, +0.008] |
| GeoChat | No cue | 0.193 [0.184, 0.202] | 75.6 [73.4, 77.8] | 0.321 [0.307, 0.335] | 0.28 [0.23, 0.33] | 0.310 [0.282, 0.338] |
| | Neutral | 0.194 [0.185, 0.203] | 75.4 [73.2, 77.6] | 0.323 [0.309, 0.337] | 0.28 [0.23, 0.33] | 0.311 [0.283, 0.339] |
| | Authentic | 0.186 [0.177, 0.195] | 77.1 [75.0, 79.2] | 0.309 [0.295, 0.323] | 0.26 [0.21, 0.31] | 0.323 [0.295, 0.351] |
| | Incorrect | 0.212 [0.202, 0.222] | 73.0 [70.8, 75.2] | 0.355 [0.339, 0.371] | 0.22 [0.17, 0.27] | 0.375 [0.347, 0.403] |
| | Neutral − No cue | +0.001 [−0.003, +0.005] | −0.2 [−0.9, +0.5] | +0.002 [−0.003, +0.007] | 0.00 [−0.02, +0.02] | +0.001 [−0.009, +0.011] |
| RSCoVLM-7B | No cue | 0.201 [0.192, 0.210] | 74.5 [72.3, 76.7] | 0.335 [0.321, 0.349] | 0.27 [0.22, 0.32] | 0.320 [0.292, 0.348] |
| | Neutral | 0.202 [0.193, 0.211] | 74.3 [72.1, 76.5] | 0.337 [0.323, 0.351] | 0.28 [0.23, 0.33] | 0.317 [0.289, 0.345] |
| | Authentic | 0.194 [0.185, 0.203] | 76.1 [74.0, 78.2] | 0.320 [0.306, 0.334] | 0.27 [0.22, 0.32] | 0.323 [0.295, 0.351] |
| | Incorrect | 0.219 [0.209, 0.229] | 72.2 [70.0, 74.4] | 0.362 [0.346, 0.378] | 0.21 [0.16, 0.26] | 0.381 [0.353, 0.409] |
| | Neutral − No cue | +0.001 [−0.003, +0.005] | −0.2 [−0.8, +0.4] | +0.002 [−0.002, +0.006] | +0.01 [−0.01, +0.03] | −0.003 [−0.013, +0.007] |

Neutral and no-cue prompting produce closely matched behavior across all metrics and models, supporting the cue-absent execution as a stable intervention anchor. For disaster assessment, authentic cues consistently improve the point-estimate MAE, accuracy, and Brier score, whereas incorrect cues degrade all three metrics.

For socioeconomic assessment, authentic geography does not improve RWI grounding relative to cue removal, although it remains consistently better than incorrect geography. Incorrect cues reduce RWI correlation and increase pairwise-ranking error across every model. These results show that geographic metadata can provide useful task information in disaster assessment while misleading geographic cues produce grounded degradation in both tasks. The reported $\mathrm{RankErr}$ metric is distinct from the margin-based pairwise hinge loss used to estimate the socioeconomic steering weights.

### C.5 Socioeconomic Assessment

The socioeconomic candidate set is

$$\mathcal{C}_{\mathrm{soc}} = \{\texttt{LOW-INCOME}, \texttt{MIDDLE-INCOME}, \texttt{AFFLUENT}\}. \tag{6}$$

**Single-slot prompt protocol.** We use a fixed prompt in which geography is the only mutable slot:

```
Images are from [G].
Assess the relative socioeconomic condition of the depicted
site using the satellite image.
Choose exactly one label:  LOW-INCOME, MIDDLE-INCOME, or
AFFLUENT.
Return only the label.
```

In the primary counterfactual audit, `[G]` is replaced by `Africa` or `Europe`, while the image, task instruction, candidate set, and scoring procedure remain fixed. The primary audit score is

$$s_{\mathrm{soc}}(x, g) = q_\theta\left(\texttt{LOW-INCOME} \mid x, g\right). \tag{7}$$

**Prompt controls.** We test whether the measured effect depends on the primary prompt using no-geography, shuffled-geography, cue-paraphrase, cue-position, image-only, and zero-shot chain-of-thought controls. The image-only control retains the geographic cue but adds:

```
Use only visible evidence in the image.  Do not use
geographic stereotypes or prior assumptions about the
location.
```

The chain-of-thought control adds:

```
Consider the visible evidence step by step before selecting
the socioeconomic label.  Return only the final label.
```

We additionally test the equivalent cue formulations `Location: [G].` and `Captured in [G].`, as well as placing the cue after the task instruction. All controls use the same candidate-normalized scoring procedure.

**RWI grounding.** For external grounding, we define the expected wealth score

$$W_\theta(x, g) = \frac{1}{2} q_\theta\left(\texttt{MIDDLE-INCOME} \mid x, g\right) + q_\theta\left(\texttt{AFFLUENT} \mid x, g\right). \tag{8}$$

Because $q_\theta$ is candidate-normalized rather than calibrated, we use rank-based grounding metrics instead of absolute prediction error. We report Spearman correlation between $W_\theta$ and within-country RWI percentiles, together with pairwise concordance

$$A_{\text{RWI}} = \Pr\left[W_\theta(x_h, g) > W_\theta(x_l, g)\right], \tag{9}$$

where $x_h$ and $x_l$ are drawn from the highest and lowest within-country RWI quartiles of the same country. Higher correlation and $A_{\text{RWI}}$ indicate stronger socioeconomic grounding.

### C.6 Disaster-Damage Assessment

The candidate set for disaster assessment is

$$\mathcal{C}_{\text{dis}} = \{\texttt{NO-DAMAGE}, \texttt{MINOR-DAMAGE}, \texttt{MAJOR-DAMAGE}, \texttt{DESTROYED}\}. \tag{10}$$

Our primary counterfactual severity score is

$$s_{\text{dis}}(x, g) = q_\theta(\texttt{MAJOR} \mid x, g) + q_\theta(\texttt{DESTROYED} \mid x, g). \tag{11}$$

For evaluation against xBD annotations, we map the four damage classes to the ordinal physical-damage scale

$$d(c) = \left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\}, \tag{12}$$

corresponding to no damage, minor damage, major damage, and destroyed, respectively. For image $i$ containing $B_i$ annotated buildings, we define the ground-truth image severity as the mean damage level across its buildings,

$$D_i = \frac{1}{B_i} \sum_{j=1}^{B_i} d(c_{ij}). \tag{13}$$

The model's corresponding expected severity is

$$\widehat{D}_i = \sum_{c \in \mathcal{C}_{\text{dis}}} q_\theta(c \mid x_i, g) d(c). \tag{14}$$

We use the continuous quantities $D_i$ and $\widehat{D}_i$ for quantitative damage-grounding evaluation, avoiding the need to assign a single class to images containing buildings with heterogeneous damage. For qualitative visualization only, we additionally report an image-level xBD class obtained by mapping $\widehat{D}_i$ to the nearest category on the ordinal damage scale above.

### C.7 Evaluation Metrics

We report behavioral, steering, and task-utility metrics. For each example, the signed geographic effect is

$$\Delta_{\text{geo}}(x_i) = s_\theta(x_i, g_a) - s_\theta(x_i, g_b), \tag{15}$$

with mean signed effect

$$\overline{\Delta}_{\text{geo}} = \frac{1}{N} \sum_{i=1}^{N} \Delta_{\text{geo}}(x_i). \tag{16}$$

Our primary sensitivity metric is the mean absolute counterfactual gap,

$$G_{\text{CF}} = \frac{1}{N} \sum_{i=1}^{N} \left| \Delta_{\text{geo}}(x_i) \right|, \tag{17}$$

where lower values indicate less dependence on geographic context. The direction rate is

$$R_{\text{dir}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[ \Delta_{\text{geo}}(x_i) > 0 \right]. \tag{18}$$

The prediction-flip rate is

$$R_{\text{flip}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[ \arg\max_{c \in \mathcal{C}} q_\theta(c \mid x_i, g_a) \neq \arg\max_{c \in \mathcal{C}} q_\theta(c \mid x_i, g_b) \right]. \tag{19}$$

Distributional sensitivity is measured using the mean Jensen-Shannon divergence,

$$\overline{D}_{\text{JS}} = \frac{1}{N} \sum_{i=1}^{N} D_{\text{JS}}\left( q_\theta(\cdot \mid x_i, g_a), q_\theta(\cdot \mid x_i, g_b) \right). \tag{20}$$

For steering experiments, let $G_i^{\text{base}}$ and $G_i^{\text{steer}}$ denote the absolute per-example gaps before and after intervention. Their means are $G_{\text{CF}}^{\text{base}}$ and $G_{\text{CF}}^{\text{steer}}$. Relative mitigation $M$ is defined in Eq. 2. We additionally report the fraction of examples whose gap decreases:

$$R_{\text{imp}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[ G_i^{\text{steer}} < G_i^{\text{base}} \right]. \tag{21}$$

Tables report $100M\%$ as mitigation and $100R_{\text{imp}}\%$ as *Images Improved*. Ties are not counted as improvements.

Socioeconomic utility is measured using Spearman correlation with within-country RWI percentiles and pairwise concordance $A_{\mathrm{RWI}}$. Disaster utility is measured using severity MAE and Spearman correlation. Hyperparameter selection uses the normalized, higher-is-better utilities in Eq. 3, while reported results remain in their original units.

# D Additional Counterfactual Analyses

## D.1 Counterfactual Effects Across Geographic Pairs

Our main experiments use the Africa-Europe contrast. To determine whether the observed behavior is specific to this pair, we repeat the socioeconomic counterfactual audit across all pairwise combinations of Africa, the Americas, Asia, and Europe.

Table 9: Signed socioeconomic counterfactual effects across geographic pairs. Each entry reports the mean change in the candidate-normalized `LOW-INCOME` score from the first geography to the second. Pixels remain identical within every comparison.

| Model | Afr.-Eur. | Afr.-Amer. | Afr.-Asia | Asia-Eur. | Amer.-Eur. | Asia-Amer. |
|---|---|---|---|---|---|---|
| Gemma-4-31B | +0.356 | +0.321 | +0.226 | +0.130 | +0.035 | +0.095 |
| Qwen2.5-VL-7B | +0.224 | +0.204 | +0.144 | +0.080 | +0.020 | +0.060 |
| GeoChat | +0.173 | +0.158 | +0.108 | +0.065 | +0.015 | +0.050 |
| RSCoVLM-7B | +0.151 | +0.139 | +0.096 | +0.055 | +0.012 | +0.043 |

The consistent signs across models and geographic contrasts indicate that the primary Africa-Europe result is part of a broader geographic structure rather than an isolated pair-specific effect.

## D.2 Prompt Robustness

A geographic counterfactual effect should not depend on a single sentence template. We therefore repeat the audit with semantically equivalent geographic statements and vary the position of the geographic cue relative to the task instruction.

Table 10: Prompt robustness of the socioeconomic geographic counterfactual effect, averaged across evaluated VLMs. Only the geographic wording or its position changes; the visual input remains identical.

| Prompt Variant | $\Delta_{\mathrm{LOW}}$ | CF Gap | Direction Rate (%) |
|---|---|---|---|
| ”Images are from [G].” | 0.226 | 0.253 | 86.1 |
| ”Location: [G].” | 0.219 | 0.247 | 84.8 |
| ”Captured in [G].” | 0.212 | 0.241 | 83.9 |
| Cue after task instruction | 0.205 | 0.236 | 82.7 |

We additionally randomize geographic-cue ordering during evaluation and observe that the counterfactual effect remains stable across prompt variants.

## D.3 Generative Prompt Controls and Regional Robustness

We compare causal steering with prompt-level and logit-level mitigation baselines across the four generative VLMs. *No Geography* removes the location cue, while *Shuffled Geography* randomly assigns cues across images. The image-only baseline explicitly instructs the model to rely on visual evidence. The stronger debiasing instruction additionally asks the model to ignore geographic information unless supported by visible evidence. Logit correction moves the geography-conditioned candidate distribution toward the corresponding cue-absent distribution, with correction strength selected on validation data and frozen before testing. Because geography is supplied through a single removable sentence, cue masking is equivalent to the No Geography condition. For paired socioeconomic conditions, RWI grounding uses the Africa-conditioned prediction.

Table 11 shows a consistent mitigation ordering. The standard counterfactual produces $G_{\mathrm{CF}} = 0.253$ and a $23.5\%$ flip rate. Image-only prompting reduces these to $0.154$ and $14.2\%$, stronger debiasing to $0.128$ and $12.1\%$, and logit correction to $0.096$ and $9.6\%$. Causal steering achieves the lowest values, $0.070$ and $7.5\%$, while improving RWI correlation from $0.27$ to $0.30$ and reducing the High-RWI LOW rate from $47.2\%$ to $42.6\%$.

The disaster analysis tests whether mitigation can suppress misleading geography without discarding useful geographic information. On the non-African cohort, authentic geography improves MAE from $0.188$ to $0.180$, whereas an incorrect Africa cue degrades it to $0.210$. Stronger prompting and logit correction recover MAE to $0.199$ and $0.194$, respectively, while causal steering reaches $0.190$. On the African target cohort, generic mitigation weakens the benefit of an authentic cue, increasing MAE from $0.180$ to $0.184$ and $0.186$, whereas causal steering largely preserves it at $0.182$. These results show a more favorable bias versus useful-prior tradeoff for causal steering than for indiscriminate suppression.

Finally, the regional analysis shows that the socioeconomic effect is not specific to one image-origin region. The signed LOW-INCOME effect remains positive across all four regions, from $0.217$ to $0.234$. Steering reduces regional gaps from $0.245$ to $0.261$ before intervention to $0.067$ to $0.073$ afterward, corresponding to $71.0\%$ to $73.6\%$ mitigation and $65.2\%$ to $69.1\%$ flip reduction.

Table 11: **Generative-VLM debiasing controls, useful-prior tradeoff, and regional robustness.** Results are macro-averaged across four generative VLMs. Panel A compares socioeconomic mitigation baselines. Panel B evaluates suppression of misleading geography and preservation of useful authentic geography in disaster assessment. Panel C reports regional robustness of causal steering.

**Panel A: Socioeconomic Debiasing Controls**

| Condition | $\mathbf{G}_{\mathrm{CF}} \downarrow$ | Flip (%) $\downarrow$ | RWI $\rho \uparrow$ | $\mathbf{A}_{\mathrm{RWI}} \uparrow$ | High-RWI LOW (%) $\downarrow$ |
|---|---|---|---|---|---|
| No geographic cue | N/A | N/A | 0.29 | 0.70 | 43.5 |
| Shuffled geographic cue | N/A | N/A | 0.22 | 0.63 | 52.8 |
| Standard Africa-Europe counterfactual | 0.253 | 23.5 | 0.27 | 0.68 | 47.2 |
| Counterfactual + image-only instruction | 0.154 | 14.2 | 0.28 | 0.69 | 45.1 |
| Counterfactual + stronger debiasing instruction | 0.128 | 12.1 | 0.29 | 0.69 | 44.3 |
| Counterfactual + logit correction toward no cue | 0.096 | 9.6 | 0.29 | 0.70 | 43.7 |
| Counterfactual + causal steering | 0.070 | 7.5 | 0.30 | 0.70 | 42.6 |

**Panel B: Disaster Useful-Prior Tradeoff**

| Cohort | Condition | MAE $\downarrow$ | Accuracy (%) $\uparrow$ | Macro-F1 $\uparrow$ | Brier $\downarrow$ |
|---|---|---|---|---|---|
| Non-African ($N = 1{,}200$) | No geographic cue | 0.188 | 76.4 | 0.712 | 0.312 |
| | Authentic geographic cue | 0.180 | 78.1 | 0.731 | 0.298 |
| | Incorrect Africa cue | 0.210 | 73.5 | 0.681 | 0.348 |
| | Incorrect cue + stronger debiasing instruction | 0.199 | 74.9 | 0.695 | 0.330 |
| | Incorrect cue + logit correction toward no cue | 0.194 | 75.6 | 0.703 | 0.322 |
| | Incorrect cue + causal steering | 0.190 | 76.1 | 0.708 | 0.315 |
| African target ($N = 300$) | Authentic Africa cue | 0.180 | 78.0 | 0.730 | 0.299 |
| | Authentic cue + stronger debiasing instruction | 0.184 | 77.2 | 0.719 | 0.306 |
| | Authentic cue + logit correction toward no cue | 0.186 | 76.9 | 0.715 | 0.309 |
| | Authentic cue + causal steering | 0.182 | 77.8 | 0.726 | 0.301 |

**Panel C: Stratification by Actual Image Region**

| Actual Region | $\overline{\mathbf{\Delta}}_{\mathbf{LOW}}$ | $\mathbf{G}_{\mathrm{CF}}^{\mathrm{base}}$ | $\mathbf{G}_{\mathrm{CF}}^{\mathrm{steer}} \downarrow$ | Mitigation (%) $\uparrow$ | Flip Reduction (%) $\uparrow$ |
|---|---|---|---|---|---|
| Africa | 0.234 | 0.261 | 0.073 | 72.0 | 68.5 |
| Americas | 0.217 | 0.248 | 0.072 | 71.0 | 65.2 |
| Asia | 0.231 | 0.258 | 0.068 | 73.6 | 69.1 |
| Europe | 0.222 | 0.245 | 0.067 | 72.7 | 67.8 |

### D.4 Country-Level Geographic Counterfactuals

The primary experiments use continent-level geographic cues. To test whether the observed effect persists at finer geographic granularity, we repeat the socioeconomic evaluation using the eight countries represented in the held-out RWI-matched cohort: Kenya, Nigeria, Ghana, Uganda, Tanzania, South Africa, Zambia, and Senegal. All evaluations use the same held-out $N = 180$ sites. No country-level examples are used for causal-layer selection, steering-direction estimation, or hyperparameter selection.

**Country-level prompt protocol.** We retain the same socioeconomic prompt used in the primary evaluation and modify only the geographic cue:

**Socioeconomic assessment**

```
You are given one satellite image of a residential site.
[GEOGRAPHIC CUE]
Assess the relative socioeconomic condition visible in the built
environment.  Interpret the labels as an ordinal visual scale rather
than as country-level income classifications.
Choose exactly one label:
LOW-INCOME: relatively limited visible housing quality,
infrastructure, or built-environment resources.
MIDDLE-INCOME: intermediate visible housing quality, infrastructure,
and built-environment resources.
AFFLUENT: relatively high visible housing quality, infrastructure, or
built-environment resources.
Return only one label:  LOW-INCOME, MIDDLE-INCOME, or AFFLUENT.
```

**Country-level geographic-cue conditions.** The prompt placeholder is instantiated using the eight countries represented in the held-out RWI-matched cohort: Kenya, Nigeria, Ghana, Uganda, Tanzania, South Africa, Zambia, and Senegal.

- **Authentic country cue:** `Images are from [COUNTRY].`, where `[COUNTRY]` is the known country associated with the image.
- **Incorrect country cue:** the same sentence is instantiated with each of the seven remaining countries. Metrics are computed for each execution and then averaged.
- **Neutral cue:** `Location information is unavailable.`
- **Cue absent:** the placeholder is removed without replacement text.

All other prompt content, visual input, candidate labels, and inference settings remain fixed across country-level counterfactuals.

**Cross-granularity steering transfer.** We test whether steering learned exclusively from continent-level counterfactuals transfers to unseen country-level cues. The causal layers, continent-level Africa steering direction, and $\lambda^* = 0.3$ are frozen from the original protocol. Because all eight country-level cues correspond to African countries, the same Africa direction is applied without country-specific direction estimation, layer selection, or hyperparameter tuning. Socioeconomic grounding is measured using RWI correlation, pairwise concordance $A_{\text{RWI}}$, and the High-RWI `LOW` rate.

Country-level cues retain a substantial counterfactual gap of $0.144$. Without country-level fitting, continent-trained steering reduces the gap to $0.054$, corresponding to $62.7\%$ mitigation, and reduces prediction flips by $57.5\%$. Mitigation is positive for every country, ranging from $51.8\%$ to $66.7\%$.

Steering also improves socioeconomic grounding: RWI correlation increases from $0.27$ to $0.30$, $A_{\text{RWI}}$ from $0.68$ to $0.70$, and the High-RWI `LOW` rate decreases from $47.2\%$ to $43.3\%$. These results show that the geographic effect and the learned intervention generalize from continent-level to finer country-level cues.

### D.5 Authentic and Misleading Geographic Cues

Pixel-identical counterfactuals reveal geographic sensitivity but do not determine whether it is harmful. We therefore compare authentic, absent, neutral, and incorrect cues against task-grounded outcomes in Table 13. We use MAE for disaster assessment and $E_{\text{soc}} = 1 - \rho_{\text{RWI}}$ for socioeconomic grounding. For disaster assessment, authentic geography improves MAE from $0.188$ to $0.180$, whereas an incorrect Africa cue increases it to $0.210$. Steering reduces this to $0.190$. On African examples, steering changes authentic-cue MAE by only $+0.002$. For socioeconomic grounding, the expanded cohort contains $600$ held-out sites from $24$ countries across three regions. RWI correlation is $0.30$ without geography, $0.28$ with authentic geography, $0.23$ with incorrect geography, and $0.22$

Table 12: **Cross-granularity transfer from continent-level steering to unseen country-level geographic cues.** The steering layers, direction, and strength are learned exclusively from the continent-level protocol and frozen before country-level evaluation. No country-level direction estimation or hyperparameter selection is performed. Panel A reports macro-averaged results across the eight country cues and four generative VLMs. Panel B stratifies evaluation by the actual country of the held-out RWI-matched site ($N = 180$). Grounding columns report unsteered $\rightarrow$ steered performance. Mitigation is computed as $100(1 - G_{\mathrm{CF}}^{\mathrm{steer}}/G_{\mathrm{CF}}^{\mathrm{base}})$. Macro mitigation is computed from the macro-averaged gaps rather than by averaging the country-specific mitigation percentages.

**Panel A: Cross-Granularity Steering Transfer**

| Condition | $G_{\mathrm{CF}}^{\mathrm{base}}$ | $G_{\mathrm{CF}}^{\mathrm{steer}}\downarrow$ | Mitigation (%) ↑ | Flip Reduction (%) ↑ | RWI $\rho\uparrow$ | $A_{\mathrm{RWI}}\uparrow$ | High-RWI LOW (%) ↓ |
|---|---|---|---|---|---|---|---|
| Country-level counterfactuals | 0.144 | 0.054 | 62.7 | 57.5 | 0.27 → 0.30 | 0.68 → 0.70 | 47.2 → 43.3 |

**Panel B: Stratification by Actual Country**

| Actual Country | $G_{\mathrm{CF}}^{\mathrm{base}}$ | $G_{\mathrm{CF}}^{\mathrm{steer}}\downarrow$ | Mitigation (%) ↑ | Flip Reduction (%) ↑ | RWI $\rho\uparrow$ | $A_{\mathrm{RWI}}\uparrow$ | High-RWI LOW (%) ↓ |
|---|---|---|---|---|---|---|---|
| Kenya | 0.150 | 0.055 | 63.3 | 58.2 | 0.28 → 0.30 | 0.68 → 0.70 | 46.8 → 43.1 |
| Nigeria | 0.170 | 0.058 | 65.9 | 61.4 | 0.26 → 0.29 | 0.67 → 0.69 | 48.2 → 44.0 |
| Ghana | 0.130 | 0.052 | 60.0 | 54.8 | 0.27 → 0.30 | 0.68 → 0.71 | 47.0 → 42.7 |
| Uganda | 0.140 | 0.057 | 59.3 | 55.1 | 0.26 → 0.29 | 0.67 → 0.70 | 48.0 → 43.6 |
| Tanzania | 0.160 | 0.055 | 65.6 | 60.2 | 0.28 → 0.30 | 0.68 → 0.70 | 46.5 → 43.0 |
| South Africa | 0.110 | 0.053 | 51.8 | 49.5 | 0.29 → 0.30 | 0.69 → 0.71 | 45.8 → 42.4 |
| Zambia | 0.140 | 0.049 | 65.0 | 59.7 | 0.26 → 0.29 | 0.67 → 0.69 | 48.4 → 44.1 |
| Senegal | 0.150 | 0.050 | 66.7 | 61.0 | 0.27 → 0.30 | 0.68 → 0.70 | 46.9 → 43.2 |
| Macro average | 0.144 | 0.054 | 62.7 | 57.5 | 0.27 → 0.30 | 0.68 → 0.70 | 47.2 → 43.3 |

with shuffled geography. Steering raises correlation to $0.31$ under incorrect cues and $0.36$ under authentic cues, while increasing ARWI to $0.73$ and reducing the High-percentile LOW rate to $40.9\%$. The pattern remains stable across matching radii and leave-one-out evaluations. These results distinguish useful geographic information from harmful geographic reliance: misleading cues degrade grounding, while causal steering recovers this degradation and preserves or improves performance under valid cues.

### D.6 Country-Conditioned Socioeconomic Grounding

RWI values are most meaningful relative to locations within the same country. We therefore evaluate socioeconomic grounding using within-country comparisons rather than absolute cross-country wealth levels. For each matched residential site, the raw RWI value $r_i$ is converted to its within-country percentile $\rho_i = \mathrm{PercentileRank}(r_i \mid \mathrm{country}(i))$. We define $Q_{\mathrm{high}}$ and $Q_{\mathrm{low}}$ as the upper and lower quartiles of the within-country RWI percentile distribution, respectively.

We define a continuous model wealth score as $W_\theta(x, g) = 0\, q_\theta(\texttt{LOW}) + \frac{1}{2} q_\theta(\texttt{MIDDLE}) + q_\theta(\texttt{AFFLUENT})$. Grounding is then measured using the association between $W_\theta$ and within-country RWI, together with pairwise concordance between relatively higher- and lower-RWI sites from the same country. We additionally report the fraction of sites in $Q_{\mathrm{high}}$ for which LOW-INCOME is the highest-scoring socioeconomic candidate.

This formulation provides a stricter socioeconomic grounding test because between-country wealth differences cannot explain the measured association. It complements the aggregate results in Table 2 by testing whether model predictions track socioeconomic variation among locations sharing the same national context.

### D.7 Disaster-Damage Class Transitions

The primary disaster analysis measures geographic sensitivity using the continuous severe-damage score. We additionally test whether changing only the stated geography is sufficient to alter the model's final discrete damage prediction.

We order the damage classes as NO-DAMAGE $<$ MINOR-DAMAGE $<$ MAJOR-DAMAGE $<$ DESTROYED. For each pixel-identical counterfactual pair, let $d(\hat{c})$ denote the ordinal index of the predicted class and define $T_i = d(\hat{c}_i^a) - d(\hat{c}_i^b)$. We categorize a transition as *more severe* when $T_i > 0$, *unchanged* when $T_i = 0$, and *less severe* when $T_i < 0$.

Table 13: **Authentic-cue utility, incorrect-cue vulnerability, and grounded steering recovery.** Panel A evaluates disaster cue validity on held-out xBD pairs. Panel B evaluates socioeconomic grounding on the original African RWI cohort and an expanded held-out cohort of 600 sites from 24 countries across three regions. Panel C tests robustness to geographic coverage and RWI matching radius. Panel D reports derived task-loss changes, with $E_{\text{soc}} = 1 - \rho_{\text{RWI}}$. Negative socioeconomic authentic-target change indicates improved grounding. High-percentile LOW is descriptive rather than an absolute error rate. Brackets denote 95% confidence intervals from 5,000 paired bootstrap replicates. Results are macro-averaged across four generative VLMs.

**Panel A: Disaster Cue Validity and Steering Recovery**

| Cohort / Condition | MAE ↓ | Acc. (%) ↑ | Macro-F1 ↑ | Brier ↓ |
|---|---|---|---|---|
| *Non-African cohort ($N = 1{,}200$)* | | | | |
| No geographic cue | 0.188 | 76.4 | 0.712 | 0.312 |
| Neutral geographic cue | 0.189 | 76.2 | 0.710 | 0.314 |
| Authentic geographic cue | 0.180 | 78.1 | 0.731 | 0.298 |
| Incorrect Africa cue | 0.210 | 73.5 | 0.681 | 0.348 |
| Incorrect Africa cue + steering | 0.190 | 76.1 | 0.708 | 0.315 |
| *African target cohort ($N = 300$)* | | | | |
| Authentic Africa cue | 0.180 | 78.0 | 0.730 | 0.299 |
| Authentic Africa cue + steering | 0.182 | 77.8 | 0.726 | 0.301 |

**Panel B: Socioeconomic RWI Grounding**

| Cohort | Condition | N | Countries | Regions | RWI $\rho$ ↑ | $\mathrm{A}_{\mathrm{RWI}}$ ↑ | High-percentile LOW (%) ↓ |
|---|---|---|---|---|---|---|---|
| Original African RWI cohort | No geographic cue | 180 | 8 | 1 | 0.29 | 0.70 | 43.5 |
| | Neutral geographic cue | 180 | 8 | 1 | 0.29 | 0.70 | 43.8 |
| | Shuffled geographic cue | 180 | 8 | 1 | 0.22 | 0.63 | 52.8 |
| | Authentic Africa cue | 180 | 8 | 1 | 0.27 | 0.68 | 47.2 |
| | Authentic Africa cue + steering | 180 | 8 | 1 | 0.30 | 0.70 | 42.6 |
| Expanded held-out RWI cohort | No geographic cue | 600 | 24 | 3 | 0.30 | 0.71 | 42.8 |
| | Authentic geographic cue | 600 | 24 | 3 | 0.28 | 0.69 | 45.6 |
| | Incorrect geographic cue | 600 | 24 | 3 | 0.23 | 0.64 | 51.0 |
| | Shuffled geographic cue | 600 | 24 | 3 | 0.22 | 0.63 | 52.1 |
| | Incorrect cue + causal steering | 600 | 24 | 3 | 0.31 | 0.71 | 42.2 |
| | Authentic cue + causal steering | 600 | 24 | 3 | 0.36 | 0.73 | 40.9 |

**Panel C: RWI Grounding Robustness**

| Evaluation | Samples / folds | Countries | $\rho_{\text{no cue}}$ ↑ | $\rho_{\text{auth}}$ ↑ | $\rho_{\text{incorrect}}$ ↑ | $\rho_{\text{steer}}$ ↑ |
|---|---|---|---|---|---|---|
| Matching radius $\leq 0.5$ km | 210 | 16 | 0.31 | 0.29 | 0.23 | 0.32 |
| Matching radius $\leq 1.0$ km | 420 | 21 | 0.30 | 0.28 | 0.23 | 0.31 |
| Matching radius $\leq 2.0$ km | 600 | 24 | 0.30 | 0.28 | 0.23 | 0.31 |
| Leave-one-country-out macro | 24 folds | 24 | 0.29 | 0.27 | 0.22 | 0.30 |
| Leave-one-region-out macro | 3 folds | 24 | 0.29 | 0.27 | 0.22 | 0.30 |

**Panel D: Derived Task-Loss Changes**

| Task | Authentic-Cue Benefit | Incorrect-Cue Harm | Steering Recovery | Authentic-Target Change |
|---|---|---|---|---|
| Disaster | $+0.008\,[+0.003, +0.013]$ | $+0.022\,[+0.015, +0.029]$ | $+0.020\,[+0.013, +0.027]$ | $+0.002\,[-0.001, +0.005]$ |
| Socioeconomic grounding, original cohort | $-0.020\,[-0.038, -0.002]$ | N/A | N/A | $-0.030\,[-0.049, -0.011]$ |
| Socioeconomic grounding, expanded cohort | $-0.020\,[-0.034, -0.006]$ | $+0.070\,[+0.048, +0.092]$ | $+0.080\,[+0.058, +0.102]$ | $-0.080\,[-0.096, -0.064]$ |

Table 14: **Country-conditioned socioeconomic grounding protocol.** All comparisons are performed within country, preventing national wealth differences from affecting the grounding metrics.

| Quantity | Definition | Interpretation |
|---|---|---|
| Within-country RWI percentile | $\rho_i = \mathrm{PercentileRank}(r_i \mid \mathrm{country}(i))$ | Relative socioeconomic position of site $i$ within its country. |
| Model wealth score | $W_\theta(x, g) = 0\, q_\theta(\texttt{LOW}) + \frac{1}{2} q_\theta(\texttt{MIDDLE}) + q_\theta(\texttt{AFFLUENT})$ | Continuous model estimate of socioeconomic status. |
| RWI Spearman correlation | $\rho_{\text{RWI}} = \mathrm{Spearman}(W_\theta, \rho)$ | Agreement between predicted wealth and within-country RWI ordering. |
| RWI concordance | $A_{\text{RWI}} = \Pr[W_\theta(x_h) > W_\theta(x_l)]$ | Fraction of within-country high-RWI and low-RWI pairs ordered correctly. |
| High-RWI LOW rate | $\Pr\left[\begin{array}{c}\arg\max_{c \in \mathcal{C}_{\text{soc}}} \\ q_\theta(c \mid x, g) = c_{\text{low}} \\ \mid\ \rho_i \in Q_{\text{high}}\end{array}\right]$ | Fraction of relatively high-RWI sites classified as LOW-INCOME. |

Table 15 shows that most discrete predictions remain unchanged, but when geography alters the predicted class, the shift is consistently more likely to move toward greater severity under the Africa

Table 15: **Discrete damage transitions under geographic counterfactuals.** For each pixel-identical pre-/post-event pair, we report whether changing the stated location from Europe to Africa moves the prediction toward a more severe damage category, leaves it unchanged, or moves it toward a less severe category.

| Model | More Severe (%) | Unchanged (%) | Less Severe (%) |
|---|---|---|---|
| Gemma-4-31B | 14.1 | 83.0 | 2.9 |
| Qwen2.5-VL-7B | 10.4 | 87.0 | 2.6 |
| GeoChat | 7.9 | 90.0 | 2.1 |
| RSCoVLM-7B | 6.2 | 92.0 | 1.8 |

cue. More-severe transitions occur $3.4$-$4.9\times$ more frequently than less-severe transitions across models. The total transition rates also recover the disaster flip rates reported in Table 1. Because the pre-event and post-event imagery are identical under both geographic contexts, these directional shifts cannot be explained by changes in physical damage and provide complementary evidence that stated geography can alter the interpretation of fixed disaster evidence.

# E Steering Analyses, Utility Preservation and Statistical Validation

## E.1 Task-Utility Preservation

Table 16: **Task-utility preservation after counterfactual-guided steering.** Socioeconomic utility is evaluated using RWI correlation and concordance, while disaster utility is evaluated using damage MAE and Spearman correlation. Steering reduces geographic sensitivity while preserving or modestly improving task-relevant utility.

**Socioeconomic Assessment**

| Model | CF Gap Before | CF Gap After ↓ | $\rho_{\mathrm{RWI}}$ Before | $\rho_{\mathrm{RWI}}$ After ↑ | $\mathbf{A}_{\mathrm{RWI}}$ Before | $\mathbf{A}_{\mathrm{RWI}}$ After ↑ |
|---|---|---|---|---|---|---|
| Gemma-4-31B | 0.356 | 0.077 | 0.21 | **0.24** | 0.64 | **0.67** |
| Qwen2.5-VL-7B | 0.258 | 0.071 | 0.25 | **0.28** | 0.67 | **0.69** |
| GeoChat | 0.211 | 0.067 | 0.29 | **0.31** | 0.69 | **0.71** |
| RSCoVLM-7B | 0.187 | 0.065 | 0.33 | **0.34** | 0.72 | **0.73** |

**Disaster-Damage Assessment**

| Model | CF Gap Before | CF Gap After ↓ | MAE Before | MAE After ↓ | $\rho$ Before | $\rho$ After ↑ |
|---|---|---|---|---|---|---|
| Gemma-4-31B | 0.135 | 0.024 | 0.214 | **0.207** | 0.54 | **0.56** |
| Qwen2.5-VL-7B | 0.112 | 0.026 | 0.196 | **0.191** | 0.58 | **0.59** |
| GeoChat | 0.094 | 0.027 | 0.181 | **0.178** | 0.62 | **0.63** |
| RSCoVLM-7B | 0.081 | 0.025 | 0.169 | **0.167** | 0.66 | **0.67** |

Reducing the geographic counterfactual effect does not require sacrificing task performance. Socioeconomic grounding is preserved or modestly improved across models, while disaster MAE and correlation remain stable or improve after intervention.

## E.2 Off-Target Capability Preservation

We evaluate whether steering alters an unrelated remote-sensing capability using a disjoint fMoW scene-classification subset containing $1{,}200$ images from all $62$ classes. The subset is class-stratified and has no physical-site overlap with any socioeconomic training, validation, or test example. The same images are evaluated by all four generative VLMs.

We use each model's socioeconomic steering direction and compare inputs with no geographic cue against inputs containing the target cue $g_a$ = Africa. The full method uses the frozen $K^* = 3$ and $\lambda^* = 0.3$, while CAA uses its configuration selected on the original socioeconomic validation set. Neither method is tuned on scene classification. Flip rate measures disagreement with the corresponding unsteered prediction, and $D_{\mathrm{JS}}$ measures divergence from its candidate distribution.

Table 17: **Off-target fMoW scene-classification performance.** Results are macro-averaged across four generative VLMs on the same class-stratified set of 1,200 images and 62 classes. Changes are measured relative to the corresponding unsteered condition. Bold denotes the best active intervention.

| Cue Condition | Method | Top-1 Acc. (%) ↑ | Macro-F1 ↑ | ΔAcc. (pp) ↑ | Flip (%) ↓ | $D_{\mathrm{JS}}$ ↓ |
|---|---|---|---|---|---|---|
| No target cue | No steering | 64.2 | 0.584 | 0.0 | 0.0 | 0.000 |
| | Full gated steering | 64.2 | 0.584 | 0.0 | 0.0 | 0.000 |
| Target cue $g_a$ | No steering | 64.1 | 0.582 | 0.0 | 0.0 | 0.000 |
| | Standard CAA | 62.8 | 0.568 | $-1.3$ | 5.4 | 0.013 |
| | **Full causal steering** | **63.9** | **0.580** | $\mathbf{-0.2}$ | **1.1** | **0.002** |

Without the target cue, gated steering leaves all outputs unchanged. When the target cue is present, standard CAA reduces accuracy by 1.3 percentage points and changes $5.4\%$ of predictions. The full method limits these changes to 0.2 points and $1.1\%$, with $D_{\mathrm{JS}} = 0.002$. These results indicate that the intervention reduces the targeted geographic effect with minimal off-target disruption.

### E.3 Statistical Validation

All geographic comparisons are paired because the same visual example is evaluated under both geographic counterfactuals. Statistical uncertainty is therefore estimated at the image level while preserving this pairing.

For the behavioral analysis, we evaluate the signed geographic effect $\Delta_{\mathrm{geo},i}$. For steering, we evaluate the per-image reduction in absolute counterfactual gap, $\Delta G_{\mathrm{CF},i} = G_{\mathrm{CF},i}^{\mathrm{before}} - G_{\mathrm{CF},i}^{\mathrm{after}}$, where positive values indicate successful mitigation. Confidence intervals are obtained by paired bootstrap resampling over images.

Two-sided $p$-values are obtained from the paired resampling distribution and corrected across the eight model–task comparisons within each analysis family using the Holm procedure. All statistics are computed on the held-out test set after causal layers, steering directions, and intervention hyperparameters have been fixed.

Table 18: **Statistical validation of geographic and steering effects.** $\Delta_{\mathrm{geo}}$ denotes the mean signed geographic counterfactual effect, while $\Delta G_{\mathrm{CF}}$ denotes the mean reduction in absolute counterfactual gap after steering. Confidence intervals use paired bootstrap resampling over images and $p_{\mathrm{adj}}$ denotes Holm-adjusted $p$-values.

| Model | Task | $\Delta_{\mathrm{geo}}$ | 95% CI | $p_{\mathrm{adj}}$ | $\Delta G_{\mathrm{CF}}$ | 95% CI | $p_{\mathrm{adj}}$ |
|---|---|---|---|---|---|---|---|
| Gemma-4-31B | Socioeconomic | 0.356 | [0.322, 0.390] | $< 0.001$ | 0.279 | [0.248, 0.310] | $< 0.001$ |
| | Disaster | 0.118 | [0.098, 0.138] | $< 0.001$ | 0.111 | [0.092, 0.130] | $< 0.001$ |
| Qwen2.5-VL-7B | Socioeconomic | 0.224 | [0.191, 0.257] | $< 0.001$ | 0.187 | [0.158, 0.216] | $< 0.001$ |
| | Disaster | 0.091 | [0.072, 0.110] | $< 0.001$ | 0.086 | [0.068, 0.104] | $< 0.001$ |
| GeoChat | Socioeconomic | 0.173 | [0.145, 0.201] | $< 0.001$ | 0.144 | [0.119, 0.169] | $< 0.001$ |
| | Disaster | 0.071 | [0.051, 0.091] | $< 0.001$ | 0.067 | [0.052, 0.082] | $< 0.001$ |
| RSCoVLM-7B | Socioeconomic | 0.151 | [0.123, 0.179] | $< 0.001$ | 0.122 | [0.098, 0.146] | $< 0.001$ |
| | Disaster | 0.058 | [0.041, 0.075] | $< 0.001$ | 0.056 | [0.041, 0.071] | $< 0.001$ |

Across all models and tasks, geographic counterfactual effects are significantly positive, with 95% confidence intervals excluding zero. Steering also yields significant reductions in counterfactual gap, with all effects remaining significant after Holm correction ($p_{\mathrm{adj}} < 0.001$).

### E.4 Language Priors, Visual Interaction, Output Robustness, and Cue-Free Influence

We first separate geographic sensitivity that persists without informative visual evidence from sensitivity that depends on the image. Each example is evaluated using the original image, no image, a common blank image, and a task-matched mismatched image. Disaster mismatches are sampled

from the same damage category; socioeconomic mismatches are matched by country and nearest RWI percentile. The same mismatch is retained across geographic conditions.

For input condition $z$, define

$$\Delta^g_{i,z} = s_\theta(x^z_i, g) - s_\theta(x^z_i, g_\emptyset), \qquad G_z = \frac{1}{|\mathcal{G}|N} \sum_g \sum_i |\Delta^g_{i,z}|.$$

Image dependence is measured by

$$D^g_z = \frac{1}{N} \sum_i \left( \Delta^g_{i,\text{full}} - \Delta^g_{i,z} \right), \qquad \mathcal{I}_z = \frac{1}{|\mathcal{G}|} \sum_g |D^g_z|.$$

Confidence intervals use $5{,}000$ paired-bootstrap replicates. Interaction significance uses paired sign-randomization tests with Holm correction across the $24$ model-task-control comparisons.

We additionally test output robustness using candidate verbalizations, candidate-order permutations, neutral class identifiers, and free-form generation.Free-form responses are mapped using fixed deterministic regular expressions (99.4% valid-output coverage). Outputs containing no valid label or multiple distinct labels are marked invalid rather than imputed, and performance is reported conditional on valid outputs. Gap equivalence uses $\delta_G = 0.015$, while utility noninferiority uses $\Delta\rho_{\text{RWI}} > -0.02$ and $\Delta\text{MAE} \leq 0.005$, with separate Holm-corrected families. We also compare OpenAI CLIP and RemoteCLIP under the same ViT-L/14 architecture, resolution, preprocessing, prompts, and scoring procedure.

**Cue-free latent geographic influence.** We next test whether the geographic representations identified from explicit counterfactuals remain active when no geographic statement is supplied. All causal layers, geography-specific directions, $K^*$, $\lambda^*$, and centering statistics are frozen before evaluation.

For cue-absent activations, geographic alignment is

$$a^g_i = \frac{1}{|\mathcal{L}^*|} \sum_{\ell \in \mathcal{L}^*} \cos\left( h^\emptyset_{i,\ell} - \mu^{\text{tr}}_\ell, v^{g,\text{bias}}_\ell \right), \qquad \hat{g}_i = \arg\max_g a^g_i,$$

where $\mu^{\text{tr}}_\ell$ is estimated from training data. We report balanced geographic accuracy over the four regions.

Causal sufficiency is tested by injecting the frozen direction into the cue-absent execution,

$$h^{\text{inj},g}_{i,\ell} = h^\emptyset_{i,\ell} + \lambda^* v^{g,\text{bias}}_\ell,$$

and comparing the resulting score shift with the corresponding explicit-cue effect. We report the latent gap $G_{\text{latent}}$, sign agreement, and per-example Spearman correlation $\rho_\Delta$. The random control averages $100$ norm-matched directions injected at the same layers.

Cue-free steering uses only the region inferred from cue-absent activations:

$$h^{\text{steer}}_{i,\ell} = h^\emptyset_{i,\ell} - \lambda^* v^{\hat{g}_i,\text{bias}}_\ell.$$

The true region is used only to evaluate geographic recovery and is never used to select the steering direction. Geographic recovery and latent intervention use held-out counterfactual test examples; grounding uses the held-out RWI cohort ($N = 180$) and xBD cohort ($N = 1{,}500$).

Table 19 shows that geographic sensitivity persists without informative visual evidence, but is consistently stronger for matched images. The effect is also stable across output formats, while RemoteCLIP exhibits lower geographic sensitivity than the matched OpenAI CLIP baseline.

More importantly, cue-free activations recover image geography substantially above the $25\%$ chance level. Latent interventions reproduce $47.4\%$ to $57.4\%$ of the corresponding explicit-cue effect, with $69.8\%$–$82.4\%$ sign agreement and $\rho_\Delta = 0.46$–$0.67$. Random directions produce much smaller effects. Cue-free steering improves or preserves socioeconomic grounding for every model and reduces disaster MAE for every model, with significant macro improvements in both tasks.

Table 19: **Language, visual, output-format, adaptation, and cue-free controls.** Panel A separates language-only and image-dependent effects. Panel B tests output-format robustness. Panel C compares matched generic and remote-sensing-adapted encoders. Panel D tests cue-free geographic recovery, latent intervention, random controls, and cue-free steering. Brackets denote 95% confidence intervals.

**Panel A: Language and Visual Interaction**

| Model | Task | $\mathbf{G}_{\text{full}}$ | $\mathbf{G}_{\text{text}}$ | $\mathbf{G}_{\text{blank}}$ | $\mathbf{G}_{\text{mismatch}}$ | $\mathcal{I}_{\text{text}}$ | $\mathcal{I}_{\text{blank}}$ | $\mathcal{I}_{\text{mismatch}}$ |
|---|---|---|---|---|---|---|---|---|
| Gemma-4-31B | Socioeconomic | 0.282 [0.258, 0.306] | 0.114 | 0.118 | 0.194 [0.171, 0.217] | 0.159 [0.131, 0.187]$^{\dagger}$ | 0.155 [0.127, 0.183]$^{\dagger}$ | 0.093 [0.066, 0.120]$^{\dagger}$ |
| | Disaster | 0.118 [0.104, 0.132] | 0.044 | 0.046 | 0.080 [0.067, 0.093] | 0.068 [0.053, 0.083]$^{\dagger}$ | 0.066 [0.051, 0.081]$^{\dagger}$ | 0.042 [0.028, 0.056]$^{\dagger}$ |
| Qwen2.5-VL-7B | Socioeconomic | 0.261 [0.238, 0.284] | 0.105 | 0.109 | 0.178 [0.156, 0.200] | 0.148 [0.121, 0.175]$^{\dagger}$ | 0.144 [0.117, 0.171]$^{\dagger}$ | 0.088 [0.062, 0.114]$^{\dagger}$ |
| | Disaster | 0.109 [0.096, 0.122] | 0.040 | 0.041 | 0.073 [0.061, 0.085] | 0.063 [0.049, 0.077]$^{\dagger}$ | 0.062 [0.048, 0.076]$^{\dagger}$ | 0.039 [0.026, 0.052]$^{\dagger}$ |
| GeoChat | Socioeconomic | 0.244 [0.222, 0.266] | 0.094 | 0.098 | 0.165 [0.144, 0.186] | 0.141 [0.115, 0.167]$^{\dagger}$ | 0.137 [0.111, 0.163]$^{\dagger}$ | 0.084 [0.058, 0.110]$^{\dagger}$ |
| | Disaster | 0.102 [0.089, 0.115] | 0.036 | 0.038 | 0.067 [0.055, 0.079] | 0.060 [0.046, 0.074]$^{\dagger}$ | 0.058 [0.044, 0.072]$^{\dagger}$ | 0.038 [0.025, 0.051]$^{\dagger}$ |
| RSCoVLM-7B | Socioeconomic | 0.225 [0.204, 0.246] | 0.083 | 0.087 | 0.151 [0.131, 0.171] | 0.135 [0.110, 0.160]$^{\dagger}$ | 0.130 [0.105, 0.155]$^{\dagger}$ | 0.078 [0.053, 0.103]$^{\dagger}$ |
| | Disaster | 0.095 [0.083, 0.107] | 0.031 | 0.033 | 0.061 [0.050, 0.072] | 0.058 [0.044, 0.072]$^{\dagger}$ | 0.057 [0.043, 0.071]$^{\dagger}$ | 0.037 [0.024, 0.050]$^{\dagger}$ |

**Panel B: Candidate and Generation Robustness ($G_{\text{CF}}$ / $U$)**

| Model | Task | Canonical | Verbalizations | Order permutations | Neutral tokens | Free form |
|---|---|---|---|---|---|---|
| Gemma-4-31B | Socioeconomic | 0.282 / 0.280 | 0.279$^{\star}$ / 0.278$^{\ddagger}$ | 0.281$^{\star}$ / 0.280$^{\ddagger}$ | 0.280$^{\star}$ / 0.279$^{\ddagger}$ | 0.274$^{\star}$ / 0.275$^{\ddagger}$ |
| | Disaster | 0.118 / 0.164 | 0.116$^{\star}$ / 0.166$^{\ddagger}$ | 0.118$^{\star}$ / 0.164$^{\ddagger}$ | 0.117$^{\star}$ / 0.165$^{\ddagger}$ | 0.121$^{\star}$ / 0.168$^{\ddagger}$ |
| Qwen2.5-VL-7B | Socioeconomic | 0.261 / 0.270 | 0.258$^{\star}$ / 0.268$^{\ddagger}$ | 0.260$^{\star}$ / 0.270$^{\ddagger}$ | 0.259$^{\star}$ / 0.269$^{\ddagger}$ | 0.253$^{\star}$ / 0.264$^{\ddagger}$ |
| | Disaster | 0.109 / 0.176 | 0.107$^{\star}$ / 0.177$^{\ddagger}$ | 0.109$^{\star}$ / 0.176$^{\ddagger}$ | 0.108$^{\star}$ / 0.177$^{\ddagger}$ | 0.112$^{\star}$ / 0.180$^{\ddagger}$ |
| GeoChat | Socioeconomic | 0.244 / 0.260 | 0.241$^{\star}$ / 0.257$^{\ddagger}$ | 0.243$^{\star}$ / 0.260$^{\ddagger}$ | 0.242$^{\star}$ / 0.259$^{\ddagger}$ | 0.237$^{\star}$ / 0.254$^{\ddagger}$ |
| | Disaster | 0.102 / 0.186 | 0.100$^{\star}$ / 0.188$^{\ddagger}$ | 0.102$^{\star}$ / 0.186$^{\ddagger}$ | 0.101$^{\star}$ / 0.187$^{\ddagger}$ | 0.106$^{\star}$ / 0.189$^{\ddagger}$ |
| RSCoVLM-7B | Socioeconomic | 0.225 / 0.270 | 0.222$^{\star}$ / 0.267$^{\ddagger}$ | 0.224$^{\star}$ / 0.270$^{\ddagger}$ | 0.223$^{\star}$ / 0.269$^{\ddagger}$ | 0.219$^{\star}$ / 0.265$^{\ddagger}$ |
| | Disaster | 0.095 / 0.194 | 0.093$^{\star}$ / 0.195$^{\ddagger}$ | 0.095$^{\star}$ / 0.194$^{\ddagger}$ | 0.094$^{\star}$ / 0.194$^{\ddagger}$ | 0.098$^{\star}$ / 0.197$^{\ddagger}$ |

**Panel C: Matched Remote-Sensing Adaptation**

| Architecture | Checkpoint | $\mathbf{G}_{\text{CF}}^{\text{soc}}\downarrow$ | $\rho_{\text{RWI}}\uparrow$ | $\mathbf{G}_{\text{CF}}^{\text{dis}}\downarrow$ | $\mathbf{MAE}_{\text{dis}}\downarrow$ |
|---|---|---|---|---|---|
| ViT-L/14 (224px) | `OpenCLIP ViT-L-14, pretrained=openai` | 0.281 [0.257, 0.305] | 0.228 [0.182, 0.274] | 0.119 [0.105, 0.133] | 0.211 [0.200, 0.222] |
| ViT-L/14 (224px) | `RemoteCLIP-ViT-L-14.pt` | 0.224 [0.202, 0.246] | 0.272 [0.226, 0.318] | 0.094 [0.082, 0.106] | 0.197 [0.187, 0.207] |
| **RemoteCLIP − OpenAI CLIP** | | -0.057 [-0.076, -0.038] | +0.044 [+0.019, +0.069] | -0.025 [-0.037, -0.013] | -0.014 [-0.022, -0.006] |

**Panel D: Latent Geographic Influence Without Explicit Cues**

| Model | Task | $\mathbf{BAcc}_{\text{geo}}(\%)\uparrow$ | $\mathbf{G}_{\text{latent}}$ | $\mathbf{G}_{\text{rand}}$ | $\mathbf{Agree}_{\text{sign}}(\%)\uparrow$ | $\rho_{\Delta}\uparrow$ | Grounding: before → steer (Δ [95% CI]) |
|---|---|---|---|---|---|---|---|
| Gemma-4-31B | Socioeconomic | 61.8 [57.3, 66.3] | 0.162 [0.147, 0.177]$^{\dagger}$ | 0.018 | 82.4 [78.9, 85.9] | 0.67 [0.62, 0.72] | 0.32 → 0.35 (+0.03 [+0.01, +0.05]) |
| | Disaster | 58.6 [56.1, 61.1] | 0.067 [0.060, 0.074]$^{\dagger}$ | 0.009 | 80.1 [78.1, 82.1] | 0.61 [0.58, 0.64] | 0.174 → 0.166 (-0.008 [-0.013, -0.003]) |
| Qwen2.5-VL-7B | Socioeconomic | 57.2 [52.6, 61.8] | 0.143 [0.129, 0.157]$^{\dagger}$ | 0.017 | 78.3 [74.5, 82.1] | 0.61 [0.55, 0.66] | 0.29 → 0.31 (+0.02 [0.00, +0.04]) |
| | Disaster | 55.4 [52.9, 57.9] | 0.058 [0.052, 0.064]$^{\dagger}$ | 0.009 | 76.5 [74.4, 78.6] | 0.56 [0.52, 0.59] | 0.184 → 0.177 (-0.007 [-0.012, -0.002]) |
| GeoChat | Socioeconomic | 53.9 [49.3, 58.5] | 0.126 [0.113, 0.139]$^{\dagger}$ | 0.016 | 74.8 [70.8, 78.8] | 0.55 [0.48, 0.61] | 0.28 → 0.30 (+0.02 [0.00, +0.04]) |
| | Disaster | 52.1 [49.6, 54.6] | 0.051 [0.045, 0.057]$^{\dagger}$ | 0.008 | 72.6 [70.3, 74.9] | 0.51 [0.47, 0.55] | 0.193 → 0.187 (-0.006 [-0.011, -0.001]) |
| RSCoVLM-7B | Socioeconomic | 50.7 [46.1, 55.3] | 0.111 [0.099, 0.123]$^{\dagger}$ | 0.015 | 71.2 [67.0, 75.4] | 0.49 [0.42, 0.56] | 0.27 → 0.29 (+0.02 [0.00, +0.04]) |
| | Disaster | 49.3 [46.8, 51.8] | 0.045 [0.040, 0.050]$^{\dagger}$ | 0.008 | 69.8 [67.5, 72.1] | 0.46 [0.42, 0.50] | 0.201 → 0.196 (-0.005 [-0.010, -0.001]) |
| **Macro average** | | 54.9 | 0.095 | 0.013 | 75.7 | 0.56 | $\rho$: 0.29 → 0.31 (+0.02 [+0.01, +0.03]) / **MAE**: 0.188 → 0.182 (-0.006 [-0.009, -0.003]) |

In Panel A, † denotes Holm-adjusted $p < 0.05$. In Panel B, $\star$ denotes gap equivalence and ‡ utility noninferiority under the tests defined above. In Panel D, † denotes $G_{\text{latent}} > G_{\text{rand}}$ with $p_{\text{Holm}} < 0.001$ across the eight model-task comparisons. Chance geographic balanced accuracy is 25%.

## E.5 Geography-Specific Steering Across Target Cues

For each target $g \in \mathcal{G}$, we estimate one direction $v_{\ell}^{g,\text{bias}}$ relative to the cue-absent anchor and evaluate it against every $g' \in \mathcal{G} \setminus \{g\}$. Directions are estimated on training data, while causal layers and steering strength are selected on validation data. For each ordered pair, we define

$$
\begin{aligned}
\Delta_{\text{geo},i}^{g,g'} &= s_\theta(x_i, g) - s_\theta(x_i, g'), \\
G_{\text{B}}^{g,g'} &= \frac{1}{N}\sum_i \left|\Delta_{\text{geo},i}^{g,g'}\right|, \\
G_{\text{S}}^{g,g'} &= \frac{1}{N}\sum_i \left|s_\theta^{\text{steer}}(x_i, g) - s_\theta(x_i, g')\right|, \\
M^{g,g'} &= 1 - \frac{G_{\text{S}}^{g,g'}}{G_{\text{B}}^{g,g'}}, \qquad G_{\text{B}}^{g,g'} > 0.
\end{aligned}
\tag{22}
$$

Each socioeconomic contrast uses the same $450$ held-out sites, with every site evaluated under both geographic cues while the visual input is kept fixed. Each disaster contrast analogously uses $1{,}500$ held-out paired sites. Macro gaps are averaged across the four models before computing $100M$. Confidence intervals are estimated from $5{,}000$ paired bootstrap replicates, using the same resampled units across models.

Table 20: **Steering across all ordered geographic pairs.** $G_{\mathrm{B}}$ and $G_{\mathrm{S}}$ denote the macro-averaged gaps before and after steering. Mitigation is computed as $100(1-\overline{G}_{\mathrm{S}}/\overline{G}_{\mathrm{B}})$. Brackets contain paired-bootstrap 95% confidence intervals.

| | | Socioeconomic | | | Disaster | | |
|---|---|---|---|---|---|---|---|
| **Target** | **Comparator** | $\mathbf{G_B}$ | $\mathbf{G_S}$ | **Mitigation (%) [95% CI]** ↑ | $\mathbf{G_B}$ | $\mathbf{G_S}$ | **Mitigation (%) [95% CI]** ↑ |
| Africa | Americas | 0.228 | 0.097 | 57.5 [52.1, 62.9] | 0.096 | 0.040 | 58.3 [51.8, 64.8] |
| Africa | Asia | 0.164 | 0.081 | 50.6 [44.8, 56.4] | 0.078 | 0.038 | 51.3 [44.5, 58.1] |
| Africa | Europe | 0.253 | 0.070 | 72.3 [67.5, 77.1] | 0.106 | 0.026 | 75.5 [70.2, 80.8] |
| Americas | Africa | 0.228 | 0.102 | 55.3 [49.8, 60.8] | 0.096 | 0.042 | 56.3 [49.6, 62.8] |
| Americas | Asia | 0.138 | 0.067 | 51.4 [45.1, 57.7] | 0.065 | 0.031 | 52.3 [45.3, 59.3] |
| Americas | Europe | 0.195 | 0.074 | 62.1 [56.4, 67.8] | 0.088 | 0.034 | 61.4 [55.0, 67.8] |
| Asia | Africa | 0.164 | 0.083 | 49.4 [43.6, 55.2] | 0.078 | 0.040 | 48.7 [41.9, 55.5] |
| Asia | Americas | 0.138 | 0.065 | 52.9 [46.7, 59.1] | 0.065 | 0.030 | 53.8 [46.9, 60.7] |
| Asia | Europe | 0.182 | 0.076 | 58.2 [52.5, 63.9] | 0.082 | 0.033 | 59.8 [53.2, 66.4] |
| Europe | Africa | 0.253 | 0.074 | 70.8 [65.9, 75.7] | 0.106 | 0.028 | 73.6 [68.1, 79.1] |
| Europe | Americas | 0.195 | 0.072 | 63.1 [57.6, 68.6] | 0.088 | 0.032 | 63.6 [57.3, 69.9] |
| Europe | Asia | 0.182 | 0.073 | 59.9 [54.3, 65.5] | 0.082 | 0.031 | 62.2 [55.8, 68.6] |

Socioeconomic utility is evaluated on the held-out RWI-matched African cohort ($N = 180$) under every target cue. Only the Africa cue is authentic for this cohort; the remaining cues are counterfactual controls. Disaster utility is evaluated on authentic regional subsets containing $300$ sites per target. The noninferiority margins, fixed before test evaluation, are $\Delta\rho_{\mathrm{RWI}} > -0.02$ and $\Delta\mathrm{MAE} \leq 0.005$.

Table 21: **Target-specific utility before and after steering.** Values are macro-averaged across models. Panels A and B report RWI correlation and disaster MAE, respectively. Brackets contain paired-bootstrap 95% confidence intervals.

**Panel A: Socioeconomic Utility**

| **Target Cue** | $\boldsymbol{\rho_{\mathrm{B}}}$ | $\boldsymbol{\rho_{\mathrm{S}}}$ | $\boldsymbol{\Delta\rho_{\mathrm{RWI}}}$ **[95% CI]** ↑ |
|---|---|---|---|
| Africa | 0.274 | 0.297 | +0.023 [+0.008, +0.038] |
| Americas | 0.288 | 0.299 | +0.011 [−0.002, +0.024] |
| Asia | 0.281 | 0.294 | +0.013 [+0.001, +0.026] |
| Europe | 0.292 | 0.298 | +0.006 [−0.006, +0.018] |

**Panel B: Disaster Utility**

| **Target Geography** | $\mathbf{MAE_B}$ | $\mathbf{MAE_S}$ | $\boldsymbol{\Delta}\mathbf{MAE}$ **[95% CI]** ↓ |
|---|---|---|---|
| Africa | 0.180 | 0.182 | +0.002 [−0.001, +0.004] |
| Americas | 0.184 | 0.181 | −0.003 [−0.006, +0.001] |
| Asia | 0.187 | 0.185 | −0.002 [−0.005, +0.002] |
| Europe | 0.178 | 0.177 | −0.001 [−0.004, +0.003] |

Table 20 shows mitigation across every ordered pair, ranging from $49.4\%$ to $72.3\%$ for socioeconomic assessment and from $48.7\%$ to $75.5\%$ for disaster assessment. For the comparison excluding both Africa and Europe, Asia → Americas achieves $52.9\%$ and $53.8\%$ mitigation, while the reverse direction achieves $51.4\%$ and $52.3\%$. All mitigation tests remain significant after Holm correction ($p_{\mathrm{adj}} < 0.001$).

Table 21 shows that all utility confidence intervals remain within the predefined noninferiority margins. Together, these results provide evidence that steering generalizes across target cues without using Europe as the intervention reference or materially degrading grounded utility.

### E.6 Steering Across Geographic Comparators

Table 22 evaluates a fixed Africa-specific intervention across Europe, the Americas, and Asia as comparison cues. On held-out images, the counterfactual gap decreases by $43.8\%$ to $57.5\%$ for Africa-Americas and $38.3\%$ to $50.2\%$ for Africa-Asia, with most examples improving in every setting. These results show that the same intervention reduces geographic sensitivity across multiple comparators.

Table 22: **Socioeconomic steering across geographic comparators.** The Africa-specific direction is estimated against the cue-absent reference using grounded harm weights. Directions and causal-layer rankings use training data; $K$ and $\lambda$ are selected on validation data. All components are frozen before evaluation on held-out images. Each contrast applies the same intervention to the Africa-conditioned execution while leaving the comparator execution unsteered.

| Model | Test Contrast | $G_{CF}$ Before | $G_{CF}$ After ↓ | Mitigation (%) ↑ | Images Improved (%) ↑ |
|---|---|---|---|---|---|
| Gemma-4-31B | Africa-Europe | 0.356 | 0.077 | 78.4 | 92.4 |
| | Africa-Americas | 0.334 | 0.142 | 57.5 | 78.2 |
| | Africa-Asia | 0.245 | 0.122 | 50.2 | 72.8 |
| Qwen2.5-VL-7B | Africa-Europe | 0.258 | 0.071 | 72.5 | 88.5 |
| | Africa-Americas | 0.231 | 0.108 | 53.2 | 74.5 |
| | Africa-Asia | 0.165 | 0.089 | 46.1 | 69.4 |
| GeoChat | Africa-Europe | 0.211 | 0.067 | 68.2 | 85.1 |
| | Africa-Americas | 0.185 | 0.096 | 48.1 | 70.8 |
| | Africa-Asia | 0.129 | 0.075 | 41.9 | 65.5 |
| RSCoVLM-7B | Africa-Europe | 0.187 | 0.065 | 65.2 | 82.6 |
| | Africa-Americas | 0.162 | 0.091 | 43.8 | 67.3 |
| | Africa-Asia | 0.115 | 0.071 | 38.3 | 62.7 |

### E.7 Geographic Prediction-Flip Reduction

Table 23 reports the relative reduction in prediction flips induced by changing the geographic cue while holding the visual input fixed. Reductions range from $58.7\%$ to $84.1\%$ across models and tasks, complementing the continuous score-gap results with greater stability of discrete predictions.

Table 23: Relative reduction in geography-induced prediction flips after counterfactual-guided steering on held-out images. Higher values indicate fewer discrete prediction changes under pixel-identical geographic counterfactuals.

| Model | Socioeconomic (%) ↑ | Disaster (%) ↑ |
|---|---|---|
| Gemma-4-31B | 75.3 | **84.1** |
| Qwen2.5-VL-7B | 68.2 | 71.5 |
| GeoChat | 62.4 | 65.8 |
| RSCoVLM-7B | 58.7 | 61.3 |

### E.8 Dataset Construction, Implementation, and Reproducibility Details

We use publicly available datasets and pretrained checkpoints. All model parameters remain frozen, with no task-specific fine-tuning or gradient-based updates during auditing, causal tracing, or steering.

**Model checkpoints.** We evaluate four generative VLM checkpoints: `google/gemma-4-31B-it`, `Qwen/Qwen2.5-VL-7B-Instruct`, `MBZUAI/geochat-7B`, and `Qingyun/RSCoVLM-7B-2512`. The general-purpose dual-encoder control uses `google/siglip2-so400m-patch16-512`. RemoteCLIP is evaluated using the official ViT-L/14 weights released by its authors.

**Socioeconomic dataset.** We use the `single-unit-residential` category of fMoW. The initial collection contains $12{,}480$ images from $3{,}120$ physical residential sites. After removing cor-

rupted or low-quality images, invalid coordinates, nonresidential examples, and sites without valid acquisition timestamps, 10,864 images from 2,716 sites remain. One observation is retained per physical site. Sites are partitioned into disjoint training, validation, and test subsets before balancing. The split is stratified across Africa, the Americas, Asia, and Europe, after which each region is independently downsampled to equal size within each split. After balancing, the final cohort contains 2,400 unique sites, comprising 1,680 training sites, 240 validation sites, and 480 test sites (a 70% / 10% / 20% partition; 420, 60, and 120 sites per region, respectively). All reported socioeconomic counterfactual evaluations use the same fixed subset of 450 held-out sites from the test partition. For each geographic contrast, all 450 sites are evaluated under both geographic cues while the visual input is kept identical, yielding 450 paired counterfactual evaluations per contrast. The 450 evaluation sites are selected once from the test partition using seed 42 and are reused unchanged across all geographic contrasts. Training data are used for causal-layer ranking and steering-direction estimation, validation data for hyperparameter selection, and test data only for final evaluation.

**RWI matching.** Eligible African fMoW sites are matched to their nearest Relative Wealth Index (RWI) grid center within 2 km using haversine distance. We restrict fMoW acquisitions to 2016-2018 and require a temporal difference of at most 24 months from the 2018 RWI reference period. When multiple acquisitions are eligible, we retain the image closest in time to the RWI reference period, breaking ties using the fMoW image-quality score. The matched cohort contains 226 sites from eight countries: Kenya (32), Nigeria (34), Ghana (26), Uganda (28), Tanzania (30), South Africa (32), Zambia (22), and Senegal (22). Country-stratified partitioning assigns 32 sites to grounded-harm and steering-direction estimation, 14 to validation and hyperparameter selection, and 180 to held-out RWI evaluation. The broader fMoW training split is used for causal-layer ranking. Raw RWI values are converted to within-country percentiles as described in Appendix D.6. For the broader socioeconomic grounding analysis, we construct an independent held-out cohort of 600 RWI-matched residential sites from 24 countries across three geographic regions. We apply the same spatial, temporal, and image-quality criteria as above and use the nearest valid RWI grid center within 2 km. The cohort is disjoint from all data used for causal-layer selection, steering-direction estimation, and hyperparameter selection. Robustness is additionally evaluated using matching radii of 0.5, 1, and 2 km and country-level and region-level leave-one-out evaluation.

**Disaster dataset.** The public xBD/xView2 release contains 22,068 RGB satellite images, corresponding to 11,034 pre-event and post-event pairs from 19 disaster events and 850,736 building annotations. Each original image has a resolution of $1024 \times 1024$ pixels.

A pair is retained only if both images are readable, share a valid pair identifier, contain at least one building polygon, and include at least one valid damage annotation. We remove unclassified annotations, malformed polygons, duplicate pairs, missing event identifiers, and images that fail preprocessing. After filtering, 9,672 valid pairs remain. Building-level labels are mapped to

$$\begin{aligned} d(\text{NO-DAMAGE}) &= 0, & d(\text{MINOR-DAMAGE}) &= \frac{1}{3}, \\ d(\text{MAJOR-DAMAGE}) &= \frac{2}{3}, & d(\text{DESTROYED}) &= 1. \end{aligned} \tag{23}$$

For pair $i$, containing $B_i$ valid buildings, continuous severity is

$$D_i = \frac{1}{B_i} \sum_{j=1}^{B_i} d(c_{ij}). \tag{24}$$

Balancing strata are obtained by mapping $D_i$ to the nearest ordinal level. The continuous value $D_i$ is retained for MAE evaluation and steering-weight computation.

**xBD grouping and partitioning.** Pairs are partitioned before balancing using disaster event and spatial grouping. Pairs sharing the same disaster event, or whose scene centers are within 5 km, are assigned to the same connected group. These groups are allocated as indivisible units to disjoint training, validation, and test partitions, preventing event-level and neighboring-scene overlap.

Damage categories are balanced independently within each split after the group assignment is fixed. The resulting balanced dataset contains 6,000 training pairs, 1,200 validation pairs, and 1,500 held-out test pairs (8,700 balanced pairs in total). Training pairs are used for causal-layer ranking and

steering-direction estimation, validation pairs for hyperparameter selection, and the held-out test pairs only for final evaluation. No physical site, disaster event, image pair, or connected spatial group occurs in more than one split.

**Canonical prompts.** The placeholder `[GEOGRAPHIC CUE]` is instantiated according to the experimental condition described below.

**Socioeconomic assessment**

```
You are given one satellite image of a residential site.
[GEOGRAPHIC CUE]
Assess the relative socioeconomic condition visible in the built
environment.  Interpret the labels as an ordinal visual scale rather
than as country-level income classifications.
Choose exactly one label:
LOW-INCOME: relatively limited visible housing quality,
infrastructure, or built-environment resources.
MIDDLE-INCOME: intermediate visible housing quality, infrastructure,
and built-environment resources.
AFFLUENT: relatively high visible housing quality, infrastructure, or
built-environment resources.
Return only one label:  LOW-INCOME, MIDDLE-INCOME, or AFFLUENT.
```

**Disaster-damage assessment**

```
You are given two satellite images of the same location.  The
first image was captured before a disaster and the second after the
disaster.
[GEOGRAPHIC CUE]
Compare the images and assess the overall physical damage visible
across the buildings.
Choose exactly one label:
NO-DAMAGE: no visible structural damage.
MINOR-DAMAGE: limited or localized damage; structures remain largely
intact.
MAJOR-DAMAGE: extensive structural damage or partial structural
failure, but substantial portions remain.
DESTROYED: complete or near-complete structural destruction.
Return only one label:  NO-DAMAGE, MINOR-DAMAGE, MAJOR-DAMAGE, or
DESTROYED.
```

The image order is fixed as pre-event followed by post-event in every disaster condition.

**Geographic-cue conditions.** The prompt placeholder is instantiated as follows:

- **Authentic cue:** `Images are from [REGION].`, where `[REGION]` is the known region associated with the image.
- **Incorrect cue:** the same sentence is instantiated with each of the three remaining regions. Metrics are averaged across the three executions.
- **Neutral cue:** `Location information is unavailable.`
- **Cue absent:** the placeholder is removed without replacement text.

The primary intervention changes the cue from Europe to Africa. Additional experiments evaluate all ordered pairs among Africa, the Americas, Asia, and Europe. Robustness to semantically equivalent formulations is reported in Appendix D. All paired executions retain identical task instructions, answer candidates, candidate ordering, and tokenization.

**Image preprocessing.** Each model uses the image processor distributed with its checkpoint. No additional visual augmentation, stochastic cropping, or model-specific image transformation is introduced. The same processed visual input is reused across all geographic-cue conditions.

**Inference and output processing.** For generative VLMs, token-level logits are used to compute the teacher-forced, length-normalized candidate probabilities defined in Eq. 1. Every subword token in the complete candidate string is scored. Candidate scoring is deterministic and does not use generated explanations.

For SigLIP2 and RemoteCLIP, image embeddings are compared with task-specific text prototypes using image-text similarity. Scores are normalized over the same candidate set used by the generative models.

Free-form evaluation uses greedy decoding with `do_sample=False`, temperature 0, top-$p = 1.0$, and at most 64 new tokens. Candidate-only generation uses at most 16 new tokens. Inference uses `bfloat16` precision and batch sizes of 1, 4, 4, and 4 for Gemma-4-31B, Qwen2.5-VL-7B, GeoChat, and RSCoVLM-7B, respectively.

Free-form outputs are case-folded, stripped of whitespace, and matched against the complete candidate labels using deterministic regular expressions. Outputs containing no valid label or multiple distinct labels are marked invalid rather than imputed. We report valid-output coverage together with performance conditional on valid outputs.

**Activation extraction and causal patching.** For each generative VLM, we extract the residual-stream representation at the task-relevant final prompt position from every transformer layer. The same token position is used across all geographic-cue and cue-absent executions.

For each geography $g \in \mathcal{G}$, causal patching replaces the representation in the geography-conditioned execution with the corresponding representation from the cue-absent execution:

$$h^g_{i,\ell} \leftarrow h^{\emptyset}_{i,\ell}. \tag{25}$$

Candidate probabilities are then recomputed while all remaining activations and model parameters stay unchanged. We define

$$\delta^g_{i,\ell} = h^g_{i,\ell} - h^{\emptyset}_{i,\ell}, \qquad \Delta^g_i = s_\theta(x_i, g) - s_\theta(x_i, g_\emptyset). \tag{26}$$

Layers are ranked using the mean causal gap reduction $C_\ell$ measured on training examples with cue-absent source activations and macro-averaged across geographic cues. Each layer is first evaluated independently to obtain its causal score. The top-$K$ layers define $\mathcal{L}^*$, and the reported top-$K$ patching result is obtained by jointly patching the selected layers with their cue-absent counterparts.

**Steering and hyperparameter selection.** For each geography $g \in \mathcal{G}$, the steering direction at layer $\ell$ is estimated from training counterfactuals relative to the cue-absent execution:

$$v^{g,\mathrm{bias}}_\ell = \frac{\sum_i b^g_i \delta^g_{i,\ell}}{\sum_i b^g_i + \epsilon}, \qquad b^g_i = \max\left\{0,\, e_i(g) - e_i(g_\emptyset)\right\}, \qquad \delta^g_{i,\ell} = h^g_{i,\ell} - h^{\emptyset}_{i,\ell}, \tag{27}$$

where $\epsilon = 10^{-6}$ and $e_i(\cdot)$ denotes the task-specific grounded error defined in Section 3. Examples for which the geographic cue does not increase grounded error relative to cue absence therefore receive zero weight. A separate direction $v^{g,\mathrm{bias}}_\ell$ is estimated for each geography; the cue-absent execution is the common reference used for direction estimation.

At inference time, steering is applied only to the geography-conditioned execution being evaluated:

$$\widetilde{h}_\ell(x, g) = h_\ell(x, g) - \lambda v^{g,\mathrm{bias}}_\ell, \qquad g \in \mathcal{G}, \quad \ell \in \mathcal{L}^*. \tag{28}$$

The cue-absent execution remains unchanged. In direct geographic comparisons, such as the primary Africa–Europe contrast, Africa and Europe are evaluation conditions rather than references for estimating the steering direction. For an ordered pair $(g, g')$, the direction associated with the target cue $g$ is estimated relative to $g_\emptyset$ and applied only to the $g$-conditioned execution, while the comparator $g'$ remains unsteered.

The validation split is used exclusively to select

$$K \in \{1, 2, 3, 4, 5\}, \qquad \lambda \in \{0, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0\} \tag{29}$$

using Eq. 4. The candidate grids and utility weight $\beta = 0.7$ are specified before validation. The selected configuration is $K^* = 3$ and $\lambda^* = 0.3$. Layer rankings, geography-specific steering directions, and hyperparameters are frozen before test evaluation. No test-time fitting or iterative optimization is performed.

**Held-out activation-patching controls.** On the test split, we evaluate the selected-layer reduction $R_{\mathrm{sel}}$, depth-matched reduction $R_{\mathrm{depth}}$, and shuffled-source reduction $R_{\mathrm{shuffle}}$. Each quantity is computed for every target geography and then macro-averaged over $g \in \mathcal{G}$.

$R_{\mathrm{sel}}$ uses the top-ranked causal layer. $R_{\mathrm{depth}}$ averages all nonselected layers in the same depth quartile as that layer. $R_{\mathrm{shuffle}}$ uses cue-absent activations from task-matched but different images.

Additional controls intervene at the top-ranked layer and final prompt token. The component controls separately patch the attention and MLP outputs. The geographic-span control replaces the region-name subtokens using the mean activation over the neutral sentence's content-token span. The visual-token control patches all aligned visual tokens using cue-absent activations from the same image. These controls are interpreted descriptively because they modify different numbers and types of activations.

Table 24: **Held-out activation-patching specificity controls.** Panel A reports counterfactual-gap reductions for the selected layer, depth-matched layers, and shuffled-source control. Difference columns report percentage-point changes with paired-bootstrap 95% confidence intervals. $p_{\mathrm{Holm}}$ denotes the Holm-adjusted two-sided paired-randomization $p$-value. Panel B reports component-level and token-position interventions at the selected depth. All quantities are macro-averaged over target geographies.

**Panel A: Layer and Source Specificity**

| Model | Task | $\mathbf{R}_{\mathrm{sel}}$ (%) | $\mathbf{R}_{\mathrm{depth}}$ (%) | $\mathbf{R}_{\mathrm{shuffle}}$ (%) | $\boldsymbol{\Delta}_{\mathbf{depth}}$ (pp) [95% CI] | $\mathbf{p}^{\mathrm{depth}}_{\mathrm{Holm}}$ | $\boldsymbol{\Delta}_{\mathbf{shuffle}}$ (pp) [95% CI] | $\mathbf{p}^{\mathrm{shuffle}}_{\mathrm{Holm}}$ |
|---|---|---|---|---|---|---|---|---|
| Gemma-4-31B | Socioeconomic | 86.2 | 44.6 | 17.8 | +41.6 [+35.2, +48.0] | $< 0.001$ | +68.4 [+62.1, +74.7] | $< 0.001$ |
| | Disaster | 78.4 | 40.6 | 15.5 | +37.8 [+31.4, +44.2] | $< 0.001$ | +62.9 [+56.8, +69.0] | $< 0.001$ |
| Qwen2.5-VL-7B | Socioeconomic | 79.5 | 44.3 | 19.0 | +35.2 [+28.9, +41.5] | $< 0.001$ | +60.5 [+54.3, +66.7] | $< 0.001$ |
| | Disaster | 73.1 | 40.5 | 16.9 | +32.6 [+26.1, +39.1] | $< 0.001$ | +56.2 [+49.8, +62.6] | $< 0.001$ |
| GeoChat | Socioeconomic | 72.8 | 42.4 | 19.0 | +30.4 [+24.2, +36.6] | $< 0.001$ | +53.8 [+47.5, +60.1] | $< 0.001$ |
| | Disaster | 67.5 | 39.6 | 18.1 | +27.9 [+21.5, +34.3] | $< 0.001$ | +49.4 [+43.1, +55.7] | $< 0.001$ |
| RSCoVLM-7B | Socioeconomic | 64.3 | 39.2 | 19.0 | +25.1 [+18.7, +31.5] | 0.002 | +45.3 [+38.9, +51.7] | $< 0.001$ |
| | Disaster | 60.1 | 37.3 | 18.4 | +22.8 [+16.4, +29.2] | 0.004 | +41.7 [+35.2, +48.2] | $< 0.001$ |

**Panel B: Component and Token Position (% Reduction [95% CI])**

| Model | Task | Attention output | MLP output | Geographic span | Visual tokens |
|---|---|---|---|---|---|
| Gemma-4-31B | Socioeconomic | 48.5 [42.7, 54.3] | 67.4 [62.0, 72.8] | 78.6 [73.5, 83.7] | 14.2 [9.8, 18.6] |
| | Disaster | 43.1 [37.4, 48.8] | 61.8 [56.1, 67.5] | 71.9 [66.5, 77.3] | 12.5 [8.1, 16.9] |
| Qwen2.5-VL-7B | Socioeconomic | 44.2 [38.6, 49.8] | 62.0 [56.4, 67.6] | 72.4 [67.1, 77.7] | 13.0 [8.7, 17.3] |
| | Disaster | 39.8 [34.2, 45.4] | 57.3 [51.7, 62.9] | 66.8 [61.3, 72.3] | 11.4 [7.2, 15.6] |
| GeoChat | Socioeconomic | 38.6 [33.1, 44.1] | 56.4 [50.8, 62.0] | 65.7 [60.2, 71.2] | 11.8 [7.6, 16.0] |
| | Disaster | 35.1 [29.7, 40.5] | 51.9 [46.3, 57.5] | 60.5 [54.9, 66.1] | 10.2 [6.1, 14.3] |
| RSCoVLM-7B | Socioeconomic | 33.7 [28.2, 39.2] | 49.2 [43.6, 54.8] | 57.8 [52.1, 63.5] | 9.6 [5.7, 13.5] |
| | Disaster | 30.5 [25.1, 35.9] | 45.6 [40.0, 51.2] | 53.4 [47.7, 59.1] | 8.4 [4.6, 12.2] |

Selected-layer patching exceeds both controls for every model and task, with all paired differences remaining positive after Holm correction. Attention, MLP, and geographic-span interventions also produce substantial reductions, whereas visual-token patching produces consistently smaller effects. Panel B does not establish a direct ranking among components because the intervention scopes differ.

**Statistical evaluation.** Confidence intervals are computed from $5{,}000$ paired-bootstrap replicates. Each replicate resamples complete image-level counterfactual sets while preserving every geographic execution of each sampled image. For socioeconomic ranking metrics, physical sites are resampled within country and all pairwise quantities are recomputed.

Panel A uses $100{,}000$ two-sided paired-randomization draws. Selected and control assignments are exchanged within evaluation examples, and the complete geography-macro statistic is recomputed. Monte Carlo $p$-values are

$$p = \frac{1 + n_{\mathrm{extreme}}}{100{,}000 + 1}. \tag{30}$$

Holm correction is applied jointly across the $16$ Panel A comparisons formed by four models, two tasks, and two controls. For other analysis families, it is applied across the corresponding eight model-task comparisons. Panel B is excluded from confirmatory testing.

**Software and hardware.** Experiments are implemented in Python 3.14.7 using PyTorch 2.13.0 and Hugging Face Transformers 5.16.1. All evaluations use one NVIDIA H100 GPU with 80 GB of VRAM. Seed 42 is used for dataset partitioning, within-split downsampling, random-layer baselines, bootstrap resampling, paired-randomization draws, and all other stochastic procedures.

# F Complete Experimental Results

## F.1 Additional Distributional Metrics

To complement the primary counterfactual metrics reported in Table 1, we measure Jensen-Shannon divergence between the complete candidate distributions under the two geographic contexts.

Table 25: Jensen-Shannon divergence under pixel-identical geographic counterfactuals.

| Model | Socioeconomic | Disaster |
|---|---|---|
| Gemma-4-31B | 0.152 | 0.061 |
| Qwen2.5-VL-7B | 0.087 | 0.047 |
| GeoChat | 0.074 | 0.039 |
| RSCoVLM-7B | 0.061 | 0.033 |

## F.2 Additional Socioeconomic Grounding Metrics

Table 26 complements the grounding metrics reported in Table 2 with the mean candidate-normalized `LOW-INCOME` score for relatively high- and low-RWI locations.

Table 26: Additional socioeconomic grounding statistics.

| Model | **High-RWI** $q_\theta(\mathrm{LOW}) \downarrow$ | **Low-RWI** $q_\theta(\mathrm{LOW})$ |
|---|---|---|
| Gemma-4-31B | 0.58 | 0.76 |
| Qwen2.5-VL-7B | 0.52 | 0.74 |
| GeoChat | 0.47 | 0.72 |
| RSCoVLM-7B | 0.42 | 0.71 |

## F.3 Absolute Causal Layer Locations

Across all evaluated models and both tasks, the peak causal layer lies near the end of the transformer stack. Although the absolute layer indices differ with architecture depth, the socioeconomic and disaster peaks are closely aligned within each model, differing by only one or two layers. This consistency supports the relative-depth analysis in Table 3 and suggests that geographic information exerts its strongest causal influence during late stages of multimodal reasoning rather than being confined to early feature encoding.

Table 27: Absolute peak causal layer identified by activation patching.

| Model | Socioeconomic | Disaster |
|---|---|---|
| Gemma-4-31B | 58 | 60 |
| Qwen2.5-VL-7B | 25 | 26 |
| GeoChat | 28 | 30 |
| RSCoVLM-7B | 25 | 26 |

# G Proofs for the Theoretical Analysis

This appendix establishes the three claims underlying Section 3.1: representation differences alone do not determine causal task influence, the proposed grounded harm-weighted direction is the unique

solution of a regularized consensus objective over harm-associated counterfactual contrasts, and steering locally reduces cue-induced sensitivity under explicit directional and curvature conditions. Throughout, $g_\emptyset$ denotes the cue-absent intervention anchor. It is an operational reference and is not assumed to be unbiased or ground truth. A geographic effect alone is therefore not evidence of bias. Grounded harm is identified only when a geographic cue increases task error relative to cue absence.

### G.1 Proof of Proposition 1: Intervention-Based Geographic Decomposition

For image $x_i$, geography $g \in \mathcal{G}$, and layer $\ell$, let

$$s_i^g = F_\ell(h_{i,\ell}^g; r_i^g), \qquad s_i^\emptyset = F_\ell(h_{i,\ell}^\emptyset; r_i^\emptyset), \tag{31}$$

where $h_{i,\ell}^g$ and $h_{i,\ell}^\emptyset$ are the intervention blocks, and $r_i^g$ and $r_i^\emptyset$ contain the remaining computational states. The cue-relative geographic effect is

$$\Delta_i^g = s_i^g - s_i^\emptyset. \tag{32}$$

Replacing the geography-conditioned activation with its cue-absent counterpart while retaining the remaining state of the geographic execution gives

$$s_{i,\ell}^{g\leftarrow\emptyset} = F_\ell(h_{i,\ell}^\emptyset; r_i^g). \tag{33}$$

Adding and subtracting this patched score yields

$$\Delta_i^g = \left(s_i^g - s_{i,\ell}^{g\leftarrow\emptyset}\right) + \left(s_{i,\ell}^{g\leftarrow\emptyset} - s_i^\emptyset\right). \tag{34}$$

Define

$$M_{i,\ell}^g = s_i^g - s_{i,\ell}^{g\leftarrow\emptyset}, \qquad R_{i,\ell}^g = s_{i,\ell}^{g\leftarrow\emptyset} - s_i^\emptyset. \tag{35}$$

Therefore,

$$\boxed{\Delta_i^g = M_{i,\ell}^g + R_{i,\ell}^g} \tag{36}$$

holds exactly. Here, $M_{i,\ell}^g$ is the score change removed by replacing the selected activation block, while $R_{i,\ell}^g$ is the geographic cue effect that remains after the intervention.

To characterize $M_{i,\ell}^g$, define

$$\delta_{i,\ell}^g = h_{i,\ell}^g - h_{i,\ell}^\emptyset \tag{37}$$

and the activation path

$$\gamma_{i,\ell}^g(t) = h_{i,\ell}^\emptyset + t\delta_{i,\ell}^g, \qquad t \in [0,1]. \tag{38}$$

If $F_\ell(\cdot; r_i^g)$ is differentiable along this path, then

$$M_{i,\ell}^g = F_\ell(\gamma_{i,\ell}^g(1); r_i^g) - F_\ell(\gamma_{i,\ell}^g(0); r_i^g) \tag{39}$$

$$= \int_0^1 \nabla_h F_\ell(\gamma_{i,\ell}^g(t); r_i^g)^\top \delta_{i,\ell}^g \, dt. \tag{40}$$

Defining the path-averaged downstream sensitivity as

$$\bar{a}^g_{i,\ell} = \int_0^1 \nabla_h F_\ell(\gamma^g_{i,\ell}(t); r^g_i)\, dt, \tag{41}$$

we obtain

$$\boxed{M^g_{i,\ell} = (\bar{a}^g_{i,\ell})^\top \delta^g_{i,\ell}.} \tag{42}$$

Thus, the patch-induced score change depends only on the component of the geographic representation contrast aligned with downstream task sensitivity. A large representation difference may therefore have little causal influence on the task prediction. This separates geographic encoding from geographic use and motivates selecting intervention locations by causal effect rather than representation magnitude.

This is an exact decomposition of the specified activation intervention. It does not claim identification of a unique natural indirect effect or a general causal decomposition from observational quantities alone.

□

### G.2 Corollary: Encoding Strength Does Not Imply Intervention Importance

Suppose $F_\ell$ is affine over the relevant activation region:

$$F_\ell(h; r^g_i) = a^{g\top}_{i,\ell} h + c^g_i. \tag{43}$$

Then

$$M^g_{i,\ell} = a^{g\top}_{i,\ell} \delta^g_{i,\ell}. \tag{44}$$

For any $n$ satisfying

$$a^{g\top}_{i,\ell} n = 0, \tag{45}$$

and for which $\delta^g_{i,\ell} + n$ remains in the affine region,

$$a^{g\top}_{i,\ell} \left(\delta^g_{i,\ell} + n\right) = a^{g\top}_{i,\ell} \delta^g_{i,\ell}. \tag{46}$$

Thus, components lying in the downstream null space can change representation distance without changing the patch-induced effect. Under a globally affine map with a nontrivial null space, this distance can increase arbitrarily while the intervention effect remains fixed. Representation magnitude alone therefore cannot determine intervention importance, motivating activation patching for causal layer selection.

□

### G.3 Proof of Proposition 2: Grounded Harm-Weighted Consensus Direction

For each geography $g$, define the nonnegative grounded-harm weight

$$b^g_i = \max\left(0, e_i(g) - e_i(g_\emptyset)\right), \tag{47}$$

where $e_i(\cdot)$ is grounded task error. For disaster assessment, $e_i$ is a per-example prediction error. For socioeconomic assessment, it is derived from within-country pairwise RWI-ranking violations. The proof requires only that these weights are fixed and nonnegative.

For $\epsilon > 0$, consider

$$J_{\ell,g,\epsilon}(v) = \sum_i b_i^g \left\| \delta_{i,\ell}^g - v \right\|_2^2 + \epsilon \|v\|_2^2. \tag{48}$$

Its gradient is

$$\nabla_v J_{\ell,g,\epsilon}(v) = -2 \sum_i b_i^g \delta_{i,\ell}^g + 2 \left( \sum_i b_i^g + \epsilon \right) v. \tag{49}$$

Setting the gradient to zero gives

$$v_\ell^{g,\text{bias}} = \frac{\sum_i b_i^g \delta_{i,\ell}^g}{\sum_i b_i^g + \epsilon}. \tag{50}$$

Moreover,

$$\nabla_v^2 J_{\ell,g,\epsilon}(v) = 2 \left( \sum_i b_i^g + \epsilon \right) I \succ 0. \tag{51}$$

Therefore,

$$\boxed{v_\ell^{g,\text{bias}} = \arg\min_v \left[ \sum_i b_i^g \left\| \delta_{i,\ell}^g - v \right\|_2^2 + \epsilon \|v\|_2^2 \right]} \tag{52}$$

is the unique global minimizer. If every weight is zero, the minimizer is $v_\ell^{g,\text{bias}} = 0$. Otherwise, the estimator forms a regularized consensus of counterfactual representation changes weighted by their increase in grounded task error, rather than averaging geographic variation indiscriminately.

Combined with causal layer selection, this construction determines both *where* to intervene and *which* geographic representation changes define the steering direction. The optimization result characterizes the estimator but does not imply that every component of the resulting direction is unsupported or that steering necessarily reduces the cue-induced effect. These properties are evaluated separately through activation patching, cue-validity controls, utility measurements, and the local conditions in Proposition 3.

□

### G.4 Proof of Proposition 3: Local Steering Reduction Under Directional Alignment

For geography $g$, let

$$\phi_i^g(\lambda) = s_\theta^{\text{steer}}(x_i, g; \lambda) \tag{53}$$

denote the score produced when the coupled intervention

$$h_\ell \leftarrow h_\ell - \lambda v_\ell^{g,\text{bias}}, \qquad \ell \in \mathcal{L}^*, \tag{54}$$

is applied during one forward pass. This formulation includes interactions between interventions at different layers. Define

$$\widetilde{\Delta}_i^g(\lambda) = \phi_i^g(\lambda) - s_i^\emptyset, \qquad \widetilde{\Delta}_i^g(0) = \Delta_i^g. \tag{55}$$

For $\Delta_i^g \neq 0$, let

$$q_i^g = \operatorname{sign}(\Delta_i^g). \tag{56}$$

Assume that $\phi_i^g$ is twice differentiable over the considered local range and that the initial steering sensitivity points toward the cue-absent score:

$$q_i^g(\phi_i^g)'(0) \le -\gamma_i^g, \qquad \gamma_i^g > 0. \tag{57}$$

Also assume bounded curvature:

$$|(\phi_i^g)''(t)| \le B_i^g. \tag{58}$$

Taylor's theorem gives

$$q_i^g \widetilde{\Delta}_i^g(\lambda) \le |\Delta_i^g| - \lambda\gamma_i^g + \frac{B_i^g \lambda^2}{2}. \tag{59}$$

Before a sign reversal, $q_i^g \widetilde{\Delta}_i^g(\lambda) \ge 0$, hence

$$\boxed{\left|\widetilde{\Delta}_i^g(\lambda)\right| \le |\Delta_i^g| - \lambda\gamma_i^g + \frac{B_i^g \lambda^2}{2}.} \tag{60}$$

For $B_i^g > 0$, any local strength satisfying

$$0 < \lambda < \frac{2\gamma_i^g}{B_i^g} \tag{61}$$

and remaining before a sign reversal gives a strictly smaller upper bound than the unsteered gap. If $B_i^g = 0$, the same conclusion holds for any positive $\lambda$ within the valid local range.

Summing over an evaluated set $\mathcal{I}_g$ gives

$$\sum_{i\in\mathcal{I}_g} \left|\widetilde{\Delta}_i^g(\lambda)\right| \le \sum_{i\in\mathcal{I}_g} |\Delta_i^g| - \lambda \sum_{i\in\mathcal{I}_g} \gamma_i^g + \frac{\lambda^2}{2} \sum_{i\in\mathcal{I}_g} B_i^g, \tag{62}$$

provided the local conditions hold for those examples. This directly links the local result to the aggregate counterfactual-gap metric.

To characterize utility change, let $U_i^g(\lambda)$ be a smooth grounded task-utility surrogate along the same intervention path. Assume

$$|(U_i^g)'(0)| \le \eta_i^g, \qquad |(U_i^g)''(t)| \le B_{U,i}^g. \tag{63}$$

A second-order expansion gives

$$\boxed{|U_i^g(\lambda) - U_i^g(0)| \le \lambda\eta_i^g + \frac{B_{U,i}^g \lambda^2}{2}.} \tag{64}$$

Hence, when the steering path points toward the cue-absent score while having small sensitivity for grounded utility, a local regime exists in which the counterfactual gap decreases with bounded utility change. These are sufficient local conditions and do not guarantee that the grounded harm-weighted estimator satisfies the directional condition for every model or example. The validation objective selects $\lambda$ before test evaluation and penalizes grounded utility degradation.

□

Proposition 1 separates representation encoding from causal task influence under the specified intervention. Proposition 2 characterizes the proposed direction as the unique regularized consensus estimator weighted by grounded cue-induced harm. Proposition 3 establishes sufficient local conditions for reducing the cue-relative gap while limiting grounded utility change. Together, they formalize the three components of the method: causal localization, grounded harm-weighted direction estimation, and locally controlled intervention. Effectiveness of the estimated directions is ultimately established by the empirical activation-patching, ablation, and steering results.