DISTRIBUTED GEOGRAPHIC IDENTITY ENCODING IN VISION-LANGUAGE MODELS: FROM LINEAR PROBING TO SPARSE AUTOENCODER FEATURE ABLATION, AND WHY POST-HOC DEBIASING FAILS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) are increasingly deployed on tasks — credit assessment, job-ad targeting, content moderation — where their representation of geographic identity has direct fairness consequences. Building on linear-probe recovery of macroeconomic signals from satellite imagery and street-view wealth estimation, we show that decoder-only VLMs encode **country identity as a structural property of the residual stream**. A ridge probe at Gemma-4 layer 6 recovers GDP PPP at 5-fold cross-validation $R^2 = 0.33$ but collapses under leave-one-country-out to $R^2 = -0.05$ (gap $\Delta = 0.38$); this LOCO-collapse signature recurs across four architectures, with the direction **emerging at layer 1** and not inherited from the vision tower. The subspace resists every post-hoc intervention we tested — input PGD, sparse-autoencoder feature ablation, universal adversarial perturbation (UAP), ROME weight edits, LoRA fine-tuning, and closed-form linear erasure (INLP, LEACE) — in a characteristic way each: geometric erasure leaves behavior intact (PGD: 84.2% geometric erasure, -7.4% behavioral shift; likewise ROME, LoRA, and closed-form LEACE), the feature dictionary offers no concentrated structure to ablate (SAE), and the only behavioral movers (UAP; INLP on one architecture, matched by a random control) act by collapsing predictions rather than removing the signal. The signal is nonetheless consequential: a single UAP direction gives Regress-To-Mean-specific gap reduction of $5.06\times$ $[3.61, 6.70]$, transfers cross-domain to FairFace face stimuli ($3.24\times$; no shared linear wealth axis, $r = -0.08$), and closes the mortgage approval gap by 49.5%. We argue country-identity is structurally encoded, distributed, and behaviorally consequential; closing the output gap is not evidence the signal was removed, so post-hoc debiasing of VLMs is unreliable. We propose the Variable Dependence Spectrum for deployment-time triage and identify training-time concept regularization as the remaining intervention surface.

SECTION 1: INTRODUCTION

Vision-Language Models (VLMs) are increasingly asked to make decisions that mimic high-stakes human judgement — estimating the wealth of a city from a street view, screening job applicants from profile photos, ranking loan applicants from property images, moderating content from visual context. In each case the model's representation of **geographic identity** — the country, region, or economic context of the input — directly shapes the output, and any systematic bias in that representation propagates into decisions about real people. Linear probes recover macro-economic signals from VLM hidden states with high accuracy and can steer behavior, suggesting to a practitioner that the relevant concept is **localized** and therefore removable by a targeted intervention. **These representations are instead distributed across the residual stream — resistant to sparse autoencoder feature ablation and every other post-hoc intervention we tested — which makes post-hoc debiasing of model output distributions unreliable.**

Across four VLM families, nine countries, and six post-hoc intervention classes, we find that the "macro-economic concept" a linear probe recovers is better understood as a **country-identity direction**: a structural feature of the residual stream that memorizes *which country* the model is seeing rather than an abstraction that generalizes across economies. The direction is born at layer 1, shares geometry with neighboring macroeconomic concepts, and distributes its encoding across a long tail of SAE features and multiple residual-stream pathways. Every post-hoc intervention we tested —

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.

input PGD, SAE feature ablation, universal adversarial perturbation (UAP), ROME-style rank-one weight edit, LoRA fine-tuning, and closed-form linear erasure — meets, or provably cannot meet, its geometric objective and fails behaviorally, or succeeds only via distributional collapse. We refer to this **geometry-behavior gap**, in its fairness-specific form (protected-attribute decisions audited by output parity), as the *defense gap* (Anonymous, 2026a); our companion work documents the same pattern for causally necessary directions in language models (Anonymous, 2026b).

The country-identity signal is nonetheless **behaviorally consequential**. A single UAP direction produces a Regress-To-Mean-specific gap reduction of $5.06\times$ $[3.61, 6.70]$ on GDP prediction, transfers across visual domains from city images to FairFace face stimuli at $3.24\times$ specificity without a shared linear wealth axis ($r = -0.08$), and closes the mortgage approval gap between low- and high-GDP cities by 49.5%. The intervention works — in the narrow sense that output disparities shrink — but works by collapsing predictions toward the unconditional mean, not by removing the country-identity knowledge the model encodes.

Positioning. Three literatures touch this phenomenon. Interpretability work recovers socioeconomic signals from VLM hidden states via linear probes and documents a geometry–behavior dissociation in safety and simple visual attributes (Authors anonymized (under review), 2026; Golovanevsky et al., 2025; Deng et al., 2026); concept-erasure theory (Belrose et al., 2023; Saillenfest & Lemberger, 2025) guarantees removal only against linear classifiers and is silent on nonlinear re-encoding downstream; and a separate line documents geographic and cultural bias in VLMs behaviorally (Huang et al., 2025; Jackson et al., 2026) but does not causally intervene on the encoded signal. We contribute the validation split (leave-one-country-out, long standard in satellite poverty mapping (Jean et al., 2016)) that separates memorization from abstraction, the first systematic six-intervention characterization of a protected-attribute concept in VLMs, and the output-vs-decision audit distinction the defense gap exposes.

CONTRIBUTIONS

1. **LOCO collapse as a memorization diagnostic.** We apply leave-one-country-out cross-validation — a standard long established in poverty prediction from satellite imagery (Jean et al., 2016) but not adopted in VLM hidden-state probing — as a deployment-relevant test for whether a probing direction abstracts across economies or memorizes which country is on screen. The KFold-to-LOCO gap ($\Delta = 0.38$ on Gemma-4 L6) is structural, replicated across four VLM architectures at the universal birth layer L1, identified by two orthogonal methods.

2. **Six converging evidence lines for distributed encoding.** Each of six post-hoc intervention classes fails, defeated by a different redundancy carrier of the same concept: L1 birth (input PGD), extreme SAE distributedness (feature ablation), attention compensation when MLP is ablated (ROME), activation-level rather than weight-level causation (LoRA vs UAP), and subspace redundancy beyond the erased directions (closed-form erasure). The pattern is a structural robustness property of the residual stream.

3. **Cross-domain transfer of the UAP perturbation (adversarial, not shared encoding).** A UAP direction from city images gives clean RTM-specific gap reduction on FairFace faces with no shared visual features ($3.24\times$, 64% of city magnitude), yet the city wealth direction does *not* project onto faces as a linear readout ($r = -0.08$): what transfers is the perturbation's behavioral effect, not a shared wealth axis — positive evidence that country-identity is distributed and context-dependent.

4. **Variable Dependence Spectrum and the defense gap.** We propose VDS as deployment-time triage and document the **defense gap**: output-distribution audits cannot distinguish genuine debiasing from distributional collapse, so closing the output gap is not evidence the signal was removed. Training-time concept regularization is the remaining intervention surface.

Sections 4–7 develop the diagnostic, the geometry, the six-intervention resistance, and the fairness consequences; limitations and robustness checks are in Section 9 and the Appendix.

SECTION 2: BACKGROUND AND RELATED WORK

Three literatures touch our phenomenon; none occupies the intersection we target: probing one abstract socioeconomic concept, intervening on it with several geometric classes under a unified protocol, and tracing whether the intervention propagates to deployment-relevant decisions.

2.1 GEOMETRY–BEHAVIOR DISSOCIATION

Safety-refusal VLMs exhibit concept distributedness with single-direction behavioral governance Authors anonymized (under review) (2026); steering vectors shift simple visual attributes where parametric priors and visual evidence compete Golovanevsky et al. (2025); concurrent causal probing reports abstract concepts globally distributed while entities localize, with compensatory activation surges Deng et al. (2026). These works establish the dissociation for safety or simple attributes; none extends it to an abstract socioeconomic concept whose geometric erasure is traced through to decisions.

2.2 GEOGRAPHIC AND CULTURAL BIAS

Geolocation benchmarks quantify *where* VLMs err across regions Huang et al. (2025); Yang et al. (2026); Wasi et al. (2026); Geng et al. (2026), while Culturescope Yu et al. (2025) and CultureVLM Liu et al. (2025) mechanize or benchmark cultural bias without causal intervention. Closest to our probing result, Jackson et al. (2026) shows nationality encoded in text-only LMs and explicitly leaves erasure as future work; we take that step in the visual modality, with a LOCO-memorization diagnostic and six intervention classes (Sections 6–7).

2.3 CONCEPT-ERASURE THEORY AND PROBING METHODOLOGY

LEACE Belrose et al. (2023) guarantees removal from linear classifiers but not against nonlinear re-encoding; R-LACE Ravfogel et al. (2022) formalizes erasure as a constrained maximin game; LEOPARD Saillenfest & Lemberger (2025) argues single-axis projection is fundamentally insufficient. Superposition Elhage et al. (2022) and holographic-encoding theory Xu (2026) predict the redundancy we observe, and disentangled-safety work Pan et al. (2025); Wu et al. (2026) generalizes the picture: recognition distributed, execution localized. Our toolkit — input PGD, SAE feature ablation Pach et al. (2025), UAP Zhang et al. (2024), rank-one editing Meng et al. (2022), LoRA, and closed-form linear erasure Ravfogel et al. (2020); Belrose et al. (2023) — applies six classes under one protocol. Methodologically, prior probing and erasure work evaluates under i.i.d. splits with no held-out entity Ravfogel et al. (2020); Elazar et al. (2021); we adopt the leave-one-country-out standard that satellite poverty-mapping has held for a decade Jean et al. (2016); Murugaboopathy et al. (2025) — under which the "wealth encoding" reading of high probing R^2 dissolves into country-specific memorization (concurrent critique: Sahoo et al. (2026); full analysis in Appendix A.7).

SECTION 3: METHOD

We study four decoder-only Vision-Language Models — **Gemma-4-E4B-it** (primary, probed at layer 6), **Gemma-3-4b-IT**, **Qwen3-VL-4B**, and **MiniCPM-V-4** — all open-weights, $\sim$4B parameters, so cross-architecture agreement reflects shared representational structure rather than shared scale. Gemma-4 carries the full intervention study; the other three replicate the core diagnostic (Section 4). Inputs are 700 urban street-view images spanning 14 countries; macro-economic labels (GDP PPP, Pollution, Quality of Life) come from Numbeo/World Bank, with **9 of 14 countries having complete labels** ($n = 450$ image–label pairs). **GDP PPP** is the primary target.

For each input we run a single forward pass and pool residual-stream hidden states over **text-token positions** (image tokens excluded), the canonical mode used throughout probing and every intervention. We pool text tokens because they are the generation pathway: behavioral outputs (GDP estimates, country names) are decoded from text-token context, so the text-pooled residual stream is the substrate the LM reads out when behaving, and image tokens influence that readout only through attention. Image-token pooling is examined as a sensitivity boundary in Appendix A.4. On the pooled representations we fit a **ridge regression probe** and report two cross-validation regimes

that must not be conflated: **KFold** (5-fold random; same nine countries in train and test) and **LOCO** (leave-one-country-out). A **GDP prediction from hidden states via a linear probe** that scores well under KFold but collapses under LOCO has memorized *which country* it is seeing rather than abstracting an economic property — the KFold-to-LOCO gap is the paper's central diagnostic. At Gemma-4 L6 the probe extracts a **GDP direction** d_{gdp} and a second axis $d_{\text{pollution}}$ that are anti-correlated, $\cos(d_{\text{gdp}}, d_{\text{pollution}}) = -0.54$ (Section 5).

To test whether d_{gdp} is behaviorally removable we apply six post-hoc intervention classes spanning the available edit techniques: **input PGD** (30-step projected gradient descent on image embeddings, optimizing geometric erasure at L6), **SAE feature ablation** (ablate the top-k TopK features most aligned with d_{gdp}), **UAP** (a single universal perturbation δ, $\|\delta\|$=20.06), **ROME** (rank-one edit to the L1 MLP down-projection), and **LoRA** (rank-16 adapter, 200 steps), plus a multi-target variant (Appendix A.9). Each intervention is scored against a **geometric objective** (fraction of $\cos^2$ removed; selectivity over random controls) and a **behavioral objective** (GDP/country-output shift). The central finding of Sections 6–7 is that these objectives come apart: the geometric objective is reliably achievable, the behavioral objective is not.

Behavioral consequences are measured on **deployment-relevant binary decisions** — mortgage approval, ad targeting, property estimation on city stimuli, plus a cross-domain FairFace task (N=500, no shared visual features). We stratify responses into low-/high-GDP terciles and quantify intervention effect as **Regress-To-Mean (RTM) specificity** — gap reduction vs. random-δ controls of matched norm, with 95% bootstrap CIs (B=2000).

Because probing studies on geographic attributes are susceptible to artifact classes — LOCO point-estimate instability, regression-to-mean gradients, within-city paraphrase leakage — we apply a prospective **artifact-guard validation protocol** (bootstrap CIs, label-permutation p, bias-free RDM/Mantel on raw centroids) and pre-register a gate any "generalization" claim must pass (Appendix A.5).

SECTION 4: COUNTRY-IDENTITY MEMORIZATION

Setup. A ridge regression probe on Gemma-4-E4B-it hidden states at layer 6 recovers GDP PPP (World Bank) at 5-fold cross-validation $R^2 = 0.33 \pm 0.04$ (5 seeds, $n = 450$ images, 9 countries with full Numbeo macro labels) — at face value the standard "linear probes find the concept" result.

The probing R^2 is in-distribution only. The 5-fold split shares images from the same nine countries between train and test. Switching to leave-one-country-out (LOCO) cross-validation (Figure 1) — training on eight countries, predicting the ninth — collapses R^2 to -0.05 (Pearson $r = -0.005$). The gap $\Delta = 0.38$ between in-distribution and cross-country R^2 is not statistical noise but the diagnostic signature of cross-validation memorization: the direction has memorized *which country* is on screen rather than learning an axis that generalizes across economies.

Cross-architecture universality. LOCO collapse is not a Gemma-4 idiosyncrasy. Replicated on three additional VLMs at the universal birth layer L1 (Section 5), three of four architectures exhibit full collapse — Qwen3-VL gap 1.05, MiniCPM-V gap 0.80, Gemma-4 gap 0.38; Qwen3-VL shows the cleanest signature (highest KFold R^2, most negative LOCO). Gemma-3-4b is the sole partial-generalization outlier (LOCO $R^2 = +0.15$), but a B=2000 within-country image-bootstrap places its point estimate in the favorable tail of a wide CI $[-0.32, +0.20]$ that includes zero, so it is **apparent, not robust** (Appendix A.5). A 12B scale-up of the primary family (Gemma-4-12B-it) also fails the abstraction gate on all conditions — scale amplifies in-distribution readability without producing cross-country abstraction (Appendix).

Bias-free corroboration. Because a ridge probe can carry its own artifacts, we corroborate memorization with two probe-free methods (Appendix A.5): pairwise distances between *raw* (non-d_{gdp}-projected) country centroids show no GDP-axis alignment (Mantel test non-significant for 3/4 architectures), and a text-only probe (city-name input, no image) yields negative LOCO R^2 across all cross-arch models. Memorization is confirmed by three independent methods and is not an artifact of the visual pathway or of the probe.

Conclusion. The direction extracted by linear probing is a country-identity direction: it generalizes within the training distribution but fails across countries. The probing $R^2 = 0.33$ on Gemma-4

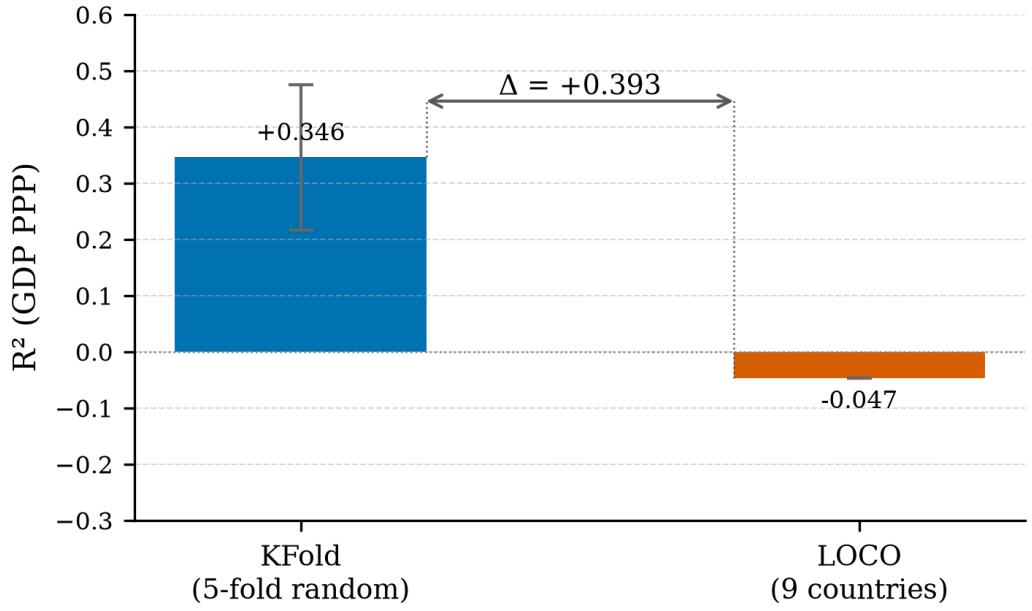

Figure 1: Cross-validation generalization gap for GDP-PPP probing at Gemma-4 layer 6. Five-fold random CV recovers R-squared = 0.33 (in-distribution, mean over 5 seeds), while leave-one-country-out (LOCO) collapses to R-squared = -0.05 (Pearson r = -0.005); the gap of 0.38 is diagnostic of country-specific memorization rather than wealth abstraction.

should not be read as evidence that VLMs encode wealth — it is evidence that they encode country identity, and that this identity subspace is the dominant carrier of any macro-economic signal linear probes recover.

Architecture	Layer	KFold R^2	LOCO R^2	Gap Δ	Verdict
Gemma-4-E4B-it	L6	0.33 ± 0.04	-0.05	0.38	collapse
Qwen3-VL-4B	L1	0.73	-0.32	1.05	collapse
MiniCPM-V-4	L1	0.45	-0.35	0.80	collapse
Gemma-3-4b	L1	0.73	$+0.15$	0.57	partial

Gap Δ is computed at full precision from the artifacts before rounding, so subtracting the rounded KFold and LOCO columns can differ from Δ by up to 0.01 (e.g., Gemma-3: $0.727 - 0.153 = 0.574 \rightarrow 0.57$*).*

SECTION 5: GEOMETRIC STRUCTURE OF THE IDENTITY SUBSPACE

Two concept directions, not one. The country-identity subspace carries a second macroeconomic axis: a ridge probe for perceived Pollution on the same Gemma-4 L6 hidden states yields a direction $d_{\text{pollution}}$ of comparable accuracy, **anti-correlated** with d_{gdp} — $\cos(d_{\text{gdp}}, d_{\text{pollution}}) = -0.54$ at L6. "Wealthy" and "polluted" are opposing poles of a shared axis, not independent concepts.

Universal birth layer. Two orthogonal methods — (i) per-layer ridge probing sweeps across all four VLMs, and (ii) cross-concept cosine anti-correlation at every depth — converge on the same answer: country-identity geometry is **born at layer 1** (Figure 2), not a late-layer abstraction. It is not inherited from the vision encoder: probing GDP on the pre-LM representation (connector output

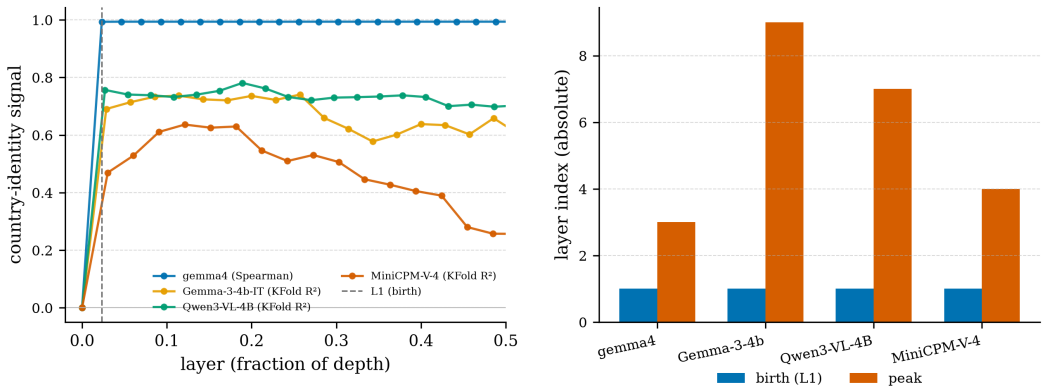

Figure 2: Universal L1 birth of the country-identity signal across four VLM architectures. Left: country-identity signal vs fractional layer depth. Right: absolute birth layer (green) vs peak layer (red) per architecture - birth is consistently at L1 while the peak varies.

at image-token positions, before any LM layer) recovers worse than the mean, and the L1/L6 wealth directions are orthogonal to the pre-LM visual direction, so the transformer constructs country-identity de novo (Appendix, Figure 5).

Cross-architecture conservation. The $\cos(d_{\text{gdp}}, d_{\text{pollution}})$ anti-correlation at L1 is conserved across VLM families — Gemma-3 -0.55, Qwen3-VL -0.39, MiniCPM-V -0.45 (tight band $[-0.39, -0.55]$). Combined with the LOCO collapse of Section 4, country-identity encoding is a **universal structural feature** of current decoder-only VLMs (per-architecture replication in Appendix).

Interpretability probing reveals two geometric facts about the country-identity concept direction — early L1 birth persisting across the residual stream, and multiple macroeconomic axes sharing one entangled axis — predict that point-level geometric interventions should have **bounded behavioral impact**, motivating the six-intervention study of Section 6.

Section 6: Distributed Encoding — Resistance to Intervention

Section 5 predicts that point-level interventions on the country-identity direction should have bounded behavioral impact, because the concept is born at L1 and shares geometry with neighbors. We test six post-hoc intervention classes — input PGD, SAE feature ablation, UAP, ROME-style rank-one weight edit, LoRA fine-tuning, and closed-form linear erasure (tercile mean-difference, single-direction nullspace, INLP, LEACE) — plus a multi-concept variant. Every class meets, or provably cannot meet, its geometric objective; none removes the concept behaviorally (closed-form protocol: Appendix A.13).

PGD — geometric win, behavioral tie. A 30-step projected gradient descent attack on the input image embeddings at Gemma-4 L6 achieves 84.2% geometric erasure of d_{gdp} with $15.7\times$ selectivity over random directions, yet downstream GDP prediction shifts by only -1636 USD (a -7.4% relative change).

SAE — extreme distributedness. A TopK sparse autoencoder on Gemma-4 L6 reconstructs d_{gdp} at FVE $= 0.556$, but the top-30 features together carry only 1.6% of the $\cos^2$ mass; we tested the top 17 features for cross-country generalization and 0 generalize — every top feature is itself a country-memorization feature. The signal is **extremely distributed** across the feature dictionary (Figure 3).

UAP — behavioral effect via distributional collapse. A single universal adversarial perturbation δ ($\|\delta\| = 20.06$) aligned with d_{gdp} produces a behavioral effect $5.06\times$ $[3.61, 6.70]$ larger than random-δ controls of matched norm, with a clean Regress-To-Mean signature (low-tercile estimates shift up, high-tercile down). The effect is real and large, but UAP achieves gap reduction by **collapsing**

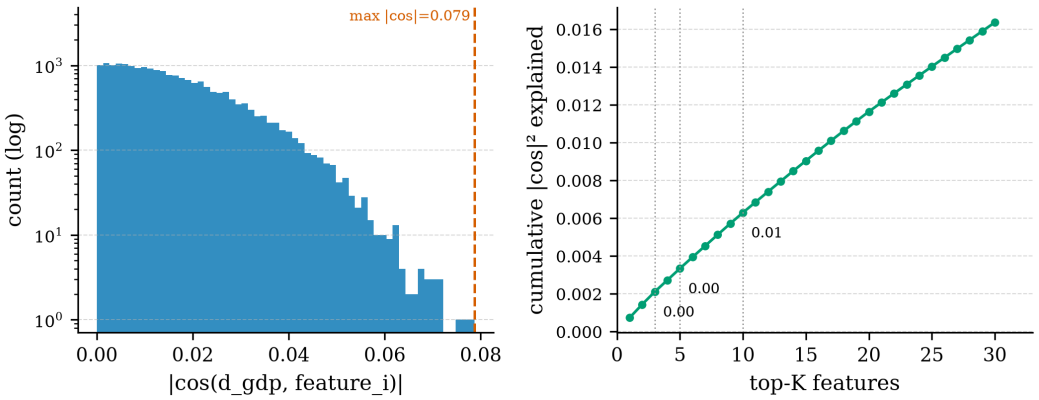

Figure 3: Extreme distribution of the country-identity direction across the Gemma-4 layer-6 feature dictionary (TopK sparse autoencoder). Left: histogram of decoder-column alignment over all 20480 features (log count). Right: cumulative cosine-squared mass captured by the top-K features reaches only 1.6% at K=30 - the signal is too thinly spread for sparse localization.

predictions toward the mean, not by redistributing mass across identity-preserving alternatives: it works as a debiasing intervention through distributional saturation, not fair feature reweighting (Section 7).

Weight-level edits — ROME null, LoRA bypass. A rank-one ROME edit to the L1 MLP down-projection that zeroes the d_{gdp} component is mathematically exact, but applied as a bf16 weight edit the change (2.6×10^{-4} relative Frobenius) falls below the bfloat16 floor and GDP/country outputs stay byte-identical ($0/45$). A corrected fp32 runtime-hook re-run confirms the edit takes effect at the representation level (projection shifts -11.0%) yet behavior stays flat (0.0 pp country-acc, -67 USD mean) — **representation-active but behavior-inert**, because distributed downstream computation compensates for the edited pathway; zeroing the top-K L1 neurons even *strengthens* the cos signature via attention compensation. A rank-16 LoRA adapter achieves 92.9% geometric erasure (higher than PGD) but only -65 USD behavioral shift — $\sim 40\times$ weaker than UAP, confirming the concept's causal effect is carried by **runtime activations, not frozen weights**.

Closed-form erasure — the missing baseline. Closed-form linear erasers fitted on the same split at both readout layers complete the spectrum. LEACE attains *perfect* linear erasure (fresh-probe $R^2 = -0.004$, 9-way accuracy at chance) yet moves the behavioral GDP gap only -16.2%. INLP on Qwen3-VL appears to succeed (-24.5%), but a matched-rank random-directions control reproduces it (-20.4%) — collapse, not debiasing — while on Gemma-4 the same removal yields 99% refusals: collapse or refusal, never removal (Appendix A.13).

Synthesis. Across six intervention classes, no intervention removes the concept: four erase geometry without moving behavior (closed-form included), SAE cannot localize the signal to erase, and the behavioral movers — UAP, INLP on Qwen3-VL — act only through distributional collapse, the latter fully reproduced by a matched random control. This is a **structural robustness property** of the residual stream, replicated across architectures (Appendix): the country-identity signal has many redundant carriers — L1 birth, cross-concept entanglement, extreme SAE distributedness, attention compensation, and activation-level causation.

Section 7: Downstream Fairness — Behaviorally Consequential

Sections 4–6 established that country-identity is geometrically robust and resists erasure. We now ask whether this signal has **measurable effects on behavior that matters for deployment**, and whether the UAP intervention that moves GDP predictions via distributional collapse also moves fairness-relevant outputs.

Baseline harm. Prompted to make binary mortgage-approval decisions on city images, Gemma-4 exhibits a baseline approval gap of 0.314 between low- and high-GDP cities — the same signal

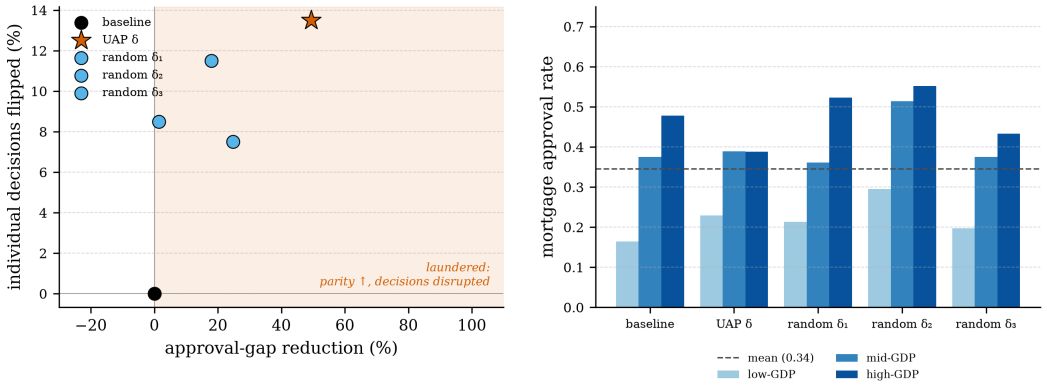

Figure 4: Fairness laundering: the city-trained UAP closes the mortgage wealth-bias gap (left, parity improves) precisely by flipping individual decisions (right, per-decision disruption rises). Right panel: low- and high-GDP approval rates converge toward the unconditional mean under UAP (distributional collapse), confirming the gap closure is saturation, not debiasing.

probing reads at L6 directly drives binary decisions about who gets credit. The gap replicates across architectures (Appendix A.8).

The universal adversarial perturbation as a fairness intervention in vision models. Applying the UAP δ (Section 6) at inference reduces the mortgage approval gap by 49.5% $[17.6\%, 98.8\%]$ and the ad-showing gap by 90.4% $[28.8\%, 168.4\%]$, flipping 13.5% $[9.0\%, 18.5\%]$ of individual decisions. At face value UAP behaves as a working debiasing intervention.

Cross-domain transfer (adversarial, not shared encoding). The city-trained δ also moves a different visual domain: on FairFace face stimuli (N=500, no shared visual features with cities). SES is operationalized as a model-rated score, not a dataset label: the VLM rates each face's perceived socioeconomic status on a 1–10 scale, and the SES gap is the mean-rating difference between the top and bottom baseline terciles of that score. On this readout δ achieves SES gap reduction with RTM specificity $3.24\times$ ($\approx$64% of the city-domain magnitude). Yet this is transfer of the perturbation's *behavioral* effect, not a shared wealth axis — a direct projection test finds the city wealth direction does *not* recover perceived-SES of faces as a linear readout ($r = -0.08$). Behavioral-but-not-geometric transfer is itself evidence for distributed, context-dependent encoding: a universal axis would transfer representationally.

Mechanism caveat: fairness laundering. Section 6 showed UAP reduces the GDP gap through distributional collapse toward the mean, not fair feature reweighting. The projection test sharpens this: since no shared linear wealth direction connects cities to faces, UAP cannot be removing a common axis — collapse is the only mechanism consistent with the transfer. UAP closes the mortgage gap by pushing low- and high-GDP predictions toward the unconditional mean, not by making the model "fairer" per image. A defender auditing only output parity would call the intervention successful; one auditing per image would not — **closing a parity gap is not sufficient evidence that the underlying signal was removed** (*fairness laundering*).

Section 8: Discussion

The results leave three practical questions.

Variable Dependence Spectrum as predictive triage. Not every macroeconomic concept resists intervention equally: LOCO R^2 forms a spectrum, from visually-grounded variables (Pollution LOCO $R^2 \approx +0.32$) that generalize partially to country-identity-bound variables (Quality of Life LOCO $R^2 \approx +0.01$) that collapse fully. We propose the **Variable Dependence Spectrum (VDS)** as deployment-time triage: a four-type taxonomy over the peak-layer probe profile in which Type A (visually direct) is partially erasable and Type D (country-identity-bound) resists every post-hoc

class tested — full thresholds and the intermediate types in Appendix A.12 — converting a series of negative erasure results into an actionable deployment criterion.

The Gemma-3 outlier is apparent, not robust. As established in Section 4, the Gemma-3-4b partial-generalization signal sits in the favorable tail of a bootstrap CI $[-0.32, +0.20]$ that includes zero, and the 12B scale-up fails the gate on all conditions — so no architecture in our study exhibits *robust* cross-country wealth abstraction; whether any VLM can is open.

The defense gap. Fairness laundering (Section 7) generalizes: the gap is not a missing tool but a **missing audit axis** — output-distribution audits cannot distinguish genuine debiasing from saturation masking, so closing the output gap is not evidence the signal was removed.

Training-time is the remaining surface. The convergence of six negative post-hoc results — each defeated by a different redundancy carrier — points at **training-time concept regularization** as the remaining intervention surface. We do not test it here (controlled re-pre-training is out of scope for a 9-country probing study); the claim is a hypothesis pointed at by six converging nulls. Caveat: our VLM intervention pipeline lacks a behavioral positive control — a trivially steerable attribute through the identical stack — so we cannot fully separate "distributed encoding defeats post-hoc intervention" from "the pipeline is globally incapable on this model class"; our companion work closes this alternative for code models (Anonymous, 2026b), but it remains open for VLMs.

SECTION 9: LIMITATIONS

- **Probing scale.** Results rest on nine countries with complete macro labels ($n = 450$); generality to a larger or more diverse country set is open.
- **Scene selection.** Images are street-view crops of one urban center per country, so the signal is specific to dense-urban statistics, and one city per country cannot capture within-country regional variation.
- **Intervention depth is single-architecture.** Five of the six intervention classes are studied on Gemma-4 only; cross-architecture replication of every class is a natural next step (PGD/SAE/UAP/ROME replicate on Qwen3-VL, Appendix A.6).
- **LM-only forward hook.** Extraction uses an LM-only hook ($\cos = 0.9998$ with the full model), so the vision-encoder pathway is unmodeled; a healing mechanism could in principle reside there, and full-model verification is a committed follow-up.
- **Pooling-channel boundary.** Geometric readouts are pooled over text-token positions only; the image-token channel is a parallel carrier our erasure percentages do not audit, and image-token pooling yields higher R^2 under OCR masking (Appendix A.4) — so removal may be misestimated; which pooling standard best supports intervention is open.
- **No VLM behavioral positive control** bounds the intervention-resistance claims (Section 8).

Negative and degenerate-prompt results (property-value, city-prompt RTM specificity, FairFace income bracket) are reported in Appendix A.1–A.3.

SECTION 10: CONCLUSION

Country-identity in current decoder-only VLMs is **structurally encoded, distributed, cross-architectural, and behaviorally consequential**. A ridge probe recovers GDP PPP at KFold $R^2 = 0.33$ but LOCO collapses to -0.05 (gap $\Delta = 0.38$), replicated across four architectures at the universal birth layer L1; six post-hoc intervention classes each meet, or provably cannot meet, their geometric objective — none removes the concept behaviorally; and the signal drives a 0.314 mortgage-approval gap that UAP closes by 49.5% — through distributional collapse, not concept removal.

Three implications follow. First, **VDS triage**: compute a variable's LOCO generalization before relying on any post-hoc intervention. Second, the **defense gap** is an audit-axis problem — output parity cannot distinguish debiasing from saturation masking. Third, **training-time concept regularization** is the most plausible remaining intervention surface against distributed representational bias. The path forward runs through training-time intervention and per-decision audit, not further refinement of post-hoc geometric edits.

REPRODUCIBILITY STATEMENT

To support reproducibility, the anonymous supplementary package accompanying this submission contains the probing and intervention code, the per-experiment frozen configurations, the aggregated result artifacts (per-architecture probing tables, intervention metrics, bootstrap CIs), per-image city identifiers, the city-level macro-economic analysis dataset, and a self-contained runnable example that reproduces the headline probing result on a CPU in seconds. The 700 street-view images are derived from public sources and, together with the face-image datasets and model weights, will be released upon publication; the 9-country Numbeo/World Bank macro-economic labels are publicly available. The artifact-guard protocol (Section 3.6, Appendix A.5) is specified prospectively, and the Gemma-3 robustness gate and the 12B scale-up test follow frozen pre-registrations. All LOCO point estimates are accompanied by B=2000 within-country image-bootstrap 95% CIs and label-permutation p-values. The four VLMs are open-weights, and the cross-architecture replications (Appendix A.6, A.8) use pipelines identical to the primary Gemma-4 study. We release per-image city identifiers, which are needed to reproduce the tercile and LOCO structure, but not the images themselves, so the package does not enable the downstream harms we document.

AI USE STATEMENT

We used large-language-model assistants throughout this project. In line with the ICLR 2027 AI policy, we disclose the following uses. Research ideation and execution: assistants helped formulate experimental designs, prototype and implement analysis code, and interpret results; all experimental claims, statistical choices, and conclusions were designed, verified, and approved by the author. Drafting sections: assistants produced initial prose drafts from verified experimental artifacts, which the author then edited and iterated. Retrieval and discovery: literature-search support; every cited reference was independently verified against its source paper. Writing polish: language editing and formatting. AI-assisted code was reviewed by the author, exercised against unit and integration checks, and its outputs validated through re-runs with fixed seeds on frozen datasets. We did not use generative AI to generate synthetic datasets or to prove or derive mathematical claims. All reported quantitative values were obtained from executed experiments and frozen result artifacts; the author independently verified the analysis pipeline and the values reported in the manuscript. The author takes full responsibility for the manuscript, the claims, the experiments, and the artifacts.

APPENDIX: NEGATIVE AND DEGENERATE-PROMPT RESULTS

This appendix documents three negative or degenerate-prompt findings that are referenced from Sections 6–8 but excluded from the primary claims. We report them in full so that the failure modes of the country-identity intervention paradigm are visible rather than hidden.

A.1 PROPERTY PROMPT — GAP CI INCLUDES ZERO

The "property value estimation" prompt was the third city-domain prompt we tested alongside the primary mortgage and ad prompts. After tercile stratification by baseline response, the property prompt yields a UAP gap reduction point estimate of 29.7% (95% bootstrap CI $[-17.8\%, 69.9\%]$, $B = 2000$). **The CI includes zero**, meaning the point estimate is not statistically distinguishable from no effect at the 95% level. The failure is largely mechanical: many model responses land in a "Mid" category that collapses the binary low/high coding, leaving $n_{\text{low}} = 20$ and $n_{\text{high}} = 9$ in some conditions — too few for a stable bootstrap. We exclude the property prompt from headline claims and report it here as a degenerate-prompt caveat: not every binary decision prompt the model can be asked yields enough variance for fairness analysis.

A.2 CITY-PROMPT RTM SPECIFICITY — CIS INCLUDE 1

Section 7 reported RTM specificity for the city-domain mortgage and ad prompts as point estimates (mortgage $1.20\times$, ad $1.48\times$). The 95% bootstrap CIs on these ratios are wide and **both include 1**: mortgage RTM CI $[0.38, 2.59]$, ad RTM CI $[0.78, 2.75]$. At the city-prompt fairness level, UAP's RTM specificity cannot be reliably distinguished from random-δ controls at 95% confidence. This

is consistent with the distributional-collapse mechanism (Section 6.3): random δ of matched norm also produces mild RTM on noisy decision tasks, so the UAP effect is not unique at this behavioral granularity. The GDP prediction level (Section 6.3) and the FairFace SES level (Section 7) both retain tight, well-above-threshold CIs, so the specificity claim holds at those granularities but not on the noisier city fairness prompts.

A.3 Income bracket prompt — degenerate tercile split

The "income bracket" prompt on FairFace (Section 7 cross-domain transfer setup) produces a degenerate tercile split: most predictions cluster at the "C" bracket, collapsing the tercile boundaries to $(3.0, 3.0)$ and leaving only 107 low-tercile samples versus 253 high-tercile samples with no mid cell. RTM specificity is $1.30\times$, well below the $1.5\times$ threshold for even partial transfer. The bracket format itself does not yield enough output variance for clean tercile analysis on Gemma-4-E4B-it. We exclude income bracket from primary cross-domain claims and report it here as a prompt-format caveat: fine-grained estimation prompts that the model collapses to a single category cannot be analyzed via tercile-based RTM.

A.4 Pooling contract for hidden-state extraction

Referenced from Section 3.2. All probing and intervention results in Sections 4–7 use a single pooling convention: the residual stream at the target layer is mean-pooled over **text-token positions**, with the dedicated image-token positions (`image_token_id` = 258880 for Gemma-4-E4B-it) excluded:

```
states      <- forward_pass(model, input, output_hidden_states=True)
                 # states[L] = hidden state at layer L, shape [1, T, D]
text_mask   <- (token_ids != IMAGE_TOKEN_ID)      # [T]
h_layer     <- states[L][0][text_mask]            # [T_text, D]
h_pooled    <- h_layer.mean(dim=0)                # [D]
```

Extraction prompt and probe regularization. The extraction forward pass uses a task-neutral prompt — *"Rate the visual appeal of this image on a scale from 0 to 10"* — rather than the GDP question, so the pooled representation is not primed by the probe target; GDP-specific prompts appear only in behavioral evaluation. The probe is ridge regression with the regularization strength fixed at $\alpha = 1.0$ (with per-fold standardization) in both the KFold and LOCO regimes — no hyperparameter selection is performed on evaluation folds; the fresh-probe recovery of Appendix A.13 is the one exception, using a four-point α grid selected on the selection split only.

This convention is fixed for **provenance consistency**, not because image tokens carry no signal. Every result in Sections 4–7 — probing, all six intervention classes, and every cross-condition comparison (original versus OCR-masked inputs, pre- versus post-edit) — is computed on the text-token-pooled representation, so a difference in R^2 between conditions reflects the intended effect rather than a pooling-mode switch.

Sensitivity (image-token pooling). A parallel probe that pools over image-token positions instead yields *higher* R^2 under OCR masking (GDP PPP at L6: `image_tokens` 0.387 / 0.363 vs `text_tokens` 0.303 / 0.288 for blur / inpaint; values from `r2_original_vs_masked.json`, experiment `exp-20260616-concept-validation`). This is consistent with the causally active signal propagating through image tokens via attention, and it means pooling choice is material. We retain `text_tokens` as the canonical mode throughout — so that intervention ΔR^2 values trace to one pooling definition — and flag the question of which pooling standard best supports intervention as open.

A.5 Methodological lessons: three probing artifacts

Section 3.6 introduced an artifact-guard protocol. Here we document, prospectively, the three artifact classes it targets, each instantiated on a claim in this paper. We report these so that the failure modes of the probing paradigm are visible, and so that the protocol — not any single result — is the transferable methodological contribution.

A.5.1 SEED-INSTABILITY / POINT-ESTIMATE FRAGILITY

LOCO cross-validation is deterministic (there is no random seed in the leave-one-country-out split); the relevant uncertainty is image-level resampling, which we capture with a $B{=}2000$ within-country bootstrap. Applied to the Gemma-3 L1 outlier of Section 4 (point LOCO $R^2 = +0.15$), the bootstrap yields 95% CI $[-0.32, +0.20]$ (mean -0.04): the $+0.15$ sits in the favorable tail of an interval that crosses zero. A label-permutation test ($p = 0.018$) confirms wealth-specific signal is present, but the magnitude is not robust. The pre-registered gate (point est $> 0.10 \wedge$ perm $p < 0.10 \wedge$ CI lower $> 0 \wedge$ memorizer-margin) **FAILS** for Gemma-3, so we do not claim cross-country abstraction. *Lesson:* a single LOCO point estimate is not generalization evidence; report the bootstrap CI.

A.5.2 REGRESSION-TO-MEAN GRADIENT

Regressing per-country LOCO error on the minimum $|\Delta \log \text{GDP}|$ to the nearest training country yields slope > 0, $r = 0.75$–0.95, permutation $p < 0.02$ across all four architectures — superficially the signature of wealth abstraction (nearby-wealth countries transfer). But this is a ridge-shrinkage artifact: predictions are pulled toward the training mean, so GDP-extreme countries incur both larger error *and* larger nearest-neighbor distance. We falsify the abstraction reading with the bias-free RDM/Mantel test: raw country-centroid distances show no wealth-axis ($r \leq 0.22$, NS, 3/4 architectures). *Lesson:* a probe-in-the-loop gradient is not abstraction evidence; use a probe-free geometric test.

A.5.3 PARAPHRASE-LEAKAGE INFLATION

A text-only probe over N paraphrases of one city yields KFold $R^2 = 0.92$–0.98, which a scrambled/fake-city-name control reproduces almost exactly ($\Delta \approx 0$): the paraphrases are within-city near-duplicates that leak the label, and any per-entity clustering (real or fake name) suffices. The leakage-free metric is LOCO (held-out country = a genuinely novel city token), under which text R^2 is negative across all three cross-arch models. *Lesson:* KFold on paraphrase-structured text inputs must not be reported as a generalization metric; use LOCO and replace the fake-name control with a label-permutation test.

A.6 CROSS-ARCHITECTURE REPLICATION OF DISTRIBUTED ENCODING

A natural concern is whether the bounded-asymmetry pattern is specific to Gemma-4's weights. We replicated the connector-output feature-space PGD attack on two further VLM families: Gemma-3-4b-IT (L9 target) and Qwen3-VL-4B-Instruct (L7 target, DeepStack architecture). All three architectures erase d_{gdp} to a comparable degree — 82.3% (Gemma-4), 91.8% (Gemma-3), and 91.5% (Qwen3-VL) — and a Kruskal-Wallis test confirms that high geometric erasure is present in every architecture ($H{=}37.47$, $p{=}7.3{\times}10^{-9}$). Erasure percentages here are per-image means (directly comparable across architectures); the aggregate projection drop of the same Gemma-4 run, 84.2%, is the figure quoted in Section 6. Distributed-encoding resistance to geometric erasure is **not a Gemma-4 weight artifact**.

The behavioral dissociation, however, reveals an architecture-dependent redundancy carrier. At matched high geometric erasure, behavioral GDP shift ranges from -0.8% (Qwen3-VL) to 4.8% (Gemma-3) to 7.4% (Gemma-4). Qwen3-VL's DeepStack architecture re-injects visual features at multiple hidden layers (L5, L11, L17) via `deepstack_visual_embeds`, overwriting the connector-output perturbation during generation — we confirmed this by injecting 16σ feature-space noise at image positions, which produces byte-identical output. This is a **distinct redundancy mechanism** from the distributed encoding of the Gemma family, yet it converges on the same geometry-behavior dissociation: multiple architectural redundancies, one outcome — post-hoc geometric erasure cannot remove the concept.

All intervention classes (Qwen3-VL). The PGD replication above establishes the pattern. We extended the replication to **every remaining intervention class on Qwen3-VL-4B-Instruct**, the DeepStack-architecture second model. Table A.1 summarizes the outcome: across all four geometric/weight intervention classes, the behavioral objective fails on both architectures despite the geometric objective being met (or the edit falling below numerical precision).

Table A.1: Bounded asymmetry across intervention class and architecture (geometric success → behavioral shift).

Method	Gemma-4 (geometric → behavioral)	Qwen3-VL-4B (geometric → behavioral)
PGD (connector-output, single-concept)	84.2% erasure → -7.4%	91.5% → -0.8%
UAP (universal δ)	RTM $5.06\times$ (distributional collapse)	96.4% erasure, sel. $56.6\times$ → median Δ 0
SAE (TopK feature decomp.)	top-30 = 1.6%, 0/17 generalize	top-30 = 2.0%, 1/27 gen. (FVE 0.635)
ROME (rank-one weight edit)	ΔW 2.6×10^{-4}, 0/45 byte-diff	ΔW 1.6×10^{-5}, GDP -44 USD (NULL)
Closed-form (LEACE / INLP, layer-output hook)	LEACE fresh-probe $R^2 = -0.004$ (perfect) → gap -16.2%; INLP → 99% refusals	INLP $-24.5\% \approx$ random ctrl -20.4% (collapse)

UAP (universal δ). A single universal perturbation optimized on Qwen3-VL ($\|\delta\|$=19.57, 300 steps, L_∞ connector space) achieves 96.4% geometric erasure at L7 with $56.6\times$ selectivity over random-δ controls — yet the median per-image behavioral GDP shift is zero (the mean -754 USD figure is a mean-of-ratios artifact from low-GDP outliers). The universal-attack class, like per-image PGD, is bounded.

SAE feature decomposition. A TopK SAE (ef8_k16) trained on Qwen3-VL L7 reconstructs the country-identity direction at FVE 0.635, but the top-30 features carry only 2.0% of $|\cos|^2$ mass and 26/27 (96%) are country-memorization features under feature-level LOCO. The extreme feature-level distributedness seen on Gemma-4 (1.6%, 0/17) replicates on Qwen3-VL.

ROME rank-one weight edit. A rank-one edit to the Qwen3-VL L1 MLP down-projection is mathematically exact (projection residual $\sim 10^{-3}$), but its weight change has relative Frobenius norm 1.6×10^{-5} — sixteen times smaller than Gemma-4 and far below the bfloat16 precision floor (7.8×10^{-3}) — so it vanishes on the forward pass. Post-edit projection drop is $\leq 2.6\%$, and behavioral GDP shifts by only -44 USD ($\approx 0.15\%$, economically null). The weight-edit class is ineffective cross-architecture.

DeepStack architectural defense. Why does Qwen3-VL show even *stronger* behavioral robustness than Gemma-4 (-0.8% vs -7.4% for PGD; median 0 for UAP)? Qwen3-VL uses the DeepStack architecture, which re-injects high-resolution visual features into the residual stream at multiple late layers (L5, L11, L17) via `deepstack_visual_embeds` Meng et al. (2024). These injections act as an **error-correction mechanism**: a perturbation applied at the connector output (the only gradient-accessible attack surface for vision) is overwritten by a fresh, unperturbed dose of visual information as the signal ascends the stack. We verified this directly — injecting 16σ feature-space noise at image-token positions produces byte-identical generation output, because DeepStack's mid-stack injections bypass the corrupted connector signal entirely.

This is a **distinct redundancy mechanism** from the within-MLP distributed encoding of the Gemma family (Sections 6.2, 6.4), yet it converges on the same geometry-behavior dissociation. Architectural diversity (Gemma's distributed encoding vs Qwen3-VL's DeepStack injection) yields a single outcome: post-hoc geometric erasure cannot remove the country-identity concept. Independent corroboration comes from adversarial-robustness benchmarking of the Qwen-VL family Torre et al. (2026), which measured elevated attack-resistance without identifying the mechanism — our work names it.

A.7 Probing methodology: accuracy is not evidence of understanding

A complementary methodological point motivates the LOCO diagnostic we rely on throughout (Section 4). Linear probing is the standard tool for claiming that a model "encodes" a concept, and influential probing methodology — control tasks with selectivity, minimum-description-length probing

— is explicitly designed to prevent the *probe itself* from overfitting. These defenses are orthogonal to the failure mode our LOCO result exposes: a probe of low complexity can still read high accuracy from a representation that has memorized *entity identity* (which country, which speaker) rather than the target concept. Among the four concept-erasure and probing-methodology works we build on most directly — the linear-erasure family (INLP Ravfogel et al. (2020), R-LACE Ravfogel et al. (2022), LEACE Belrose et al. (2023)) and amnesic probing Elazar et al. (2021) — every one evaluates under i.i.d. random splits with no held-out-entity generalization test (INLP's own Section 7 explicitly acknowledges this i.i.d. assumption), leaving entity memorization unchecked. Our critique extends a prior step in the same critical tradition: amnesic probing Elazar et al. (2021) established that probing the *presence* of information does not establish its causal *use*; we expose the orthogonal failure that probing accuracy does not establish *generalization* across entities. The two are independent weaknesses of a single tool. Concurrently, Sahoo et al. (2026) shows that linear probes detect task format rather than reasoning mode in LM hidden states — an independent cross-domain confirmation that high probing accuracy need not reflect the target property, the same failure mode we document for socioeconomic attributes. Notably, a sibling field has held the standard we adopt for a decade: the influential satellite-poverty study of Jean et al. (2016) (*Science*) validated predictions with leave-one-country-out splits from inception, current multimodal poverty-mapping Murugaboopathy et al. (2025) continues to require out-of-country and out-of-time generalization, and the erasure literature itself acknowledges the i.i.d. assumption our LOCO test relaxes (Ravfogel et al. (2020), Section 7). Our contribution is not a new probe but the adoption of this standard where it matters — and the observation that, once adopted, the "wealth encoding" reading of a high probing R^2 dissolves into country-specific memorization.

A.8 Cross-architecture replication of fairness harm

Cross-architecture fairness harm. The baseline mortgage-approval gap is not Gemma-4-specific. We evaluated the same binary mortgage prompt on Gemma-3-4b-IT and Qwen3-VL-4B-Instruct (200 images, identical GDP-tercile boundaries). All three architectures favor high-GDP cities by a comparable margin: approval gap $+0.314$ (Gemma-4), $+0.390$ (Gemma-3), $+0.387$ (Qwen3-VL), i.e. high-GDP neighborhoods receive 1.8–2.9$\times$ more approvals for low-GDP neighborhoods across two model families (Google and Alibaba). The country-identity-driven wealth bias in binary decisions is a **universal VLM property**, not an artifact of one model's weights.

Cross-architecture cross-domain transfer of the perturbation. The behavioral transfer also replicates on a second architecture. Applying the Qwen3-VL UAP δ (from Appendix A.6) to FairFace face stimuli (N=500, 5 races) yields **partial behavioral transfer**: RTM specificity $2.72\times$ on competence and $2.10\times$ on income-bracket, both with a clean low-up/high-down signature, while the per-image effect is bounded (median shift 0, 10–27% of predictions changed) — consistent with Qwen3-VL's stronger per-image DeepStack defense (Appendix A.6). As in the single-architecture case, this is adversarial generalization of the perturbation across model families, not a shared representational subspace.

A.9 Multi-concept extension — gradient interference at cos = -0.54

We tested whether the cross-concept entanglement (Section 5, $\cos(d_{\text{gdp}}, d_{\text{pollution}}) = -0.54$) induces interference when both directions are targeted simultaneously. This extension runs a separate protocol — the Ridge-extracted cross-concept directions of Section 5 on an 8-country subset (240 images), with erasure reported as the per-image median of $1-\text{proj}_{\text{adv}}/\text{proj}_{\text{clean}}$ — so its erasure levels are not comparable to the aggregate erasure of the main attack (Section 6, Table A.1); the quantity of interest is the single-to-multi ratio at a fixed budget. A multi-target PGD attack against both d_{gdp} and $d_{\text{pollution}}$ yields 31% gradient interference between the objectives and only 11.7% geometric erasure per direction — roughly half the 22.1% attained by single-target PGD on either direction alone. Anti-correlated geometry makes the two concepts partially substitutable in the optimization landscape: gradient steps that erase d_{gdp} partially restore $d_{\text{pollution}}$ and vice versa. Distributed encoding extends across concept space, not just across features within one concept.

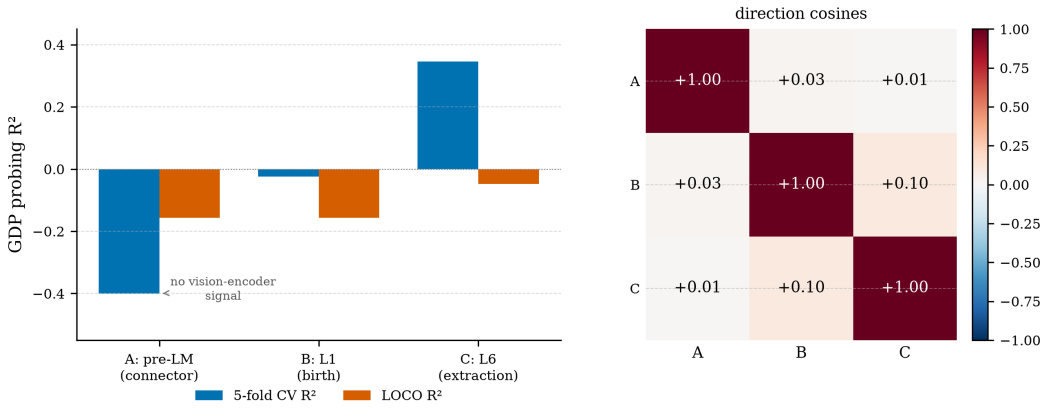

Figure 5: L1-origin ablation: the country-identity GDP signal is not inherited from the vision encoder. Probing the pre-LM representation recovers worse than the mean (5-fold R-squared = -0.40), while the signal emerges by L1 and peaks at L6. Right: the L1/L6 wealth directions are orthogonal to the pre-LM visual direction. Country-identity is constructed by transformer computation, not handed to it by the vision tower.

A.10 Scale does not produce abstraction (12B)

To check whether scale alone builds a cross-country wealth axis, we extended the robust-abstraction gate (Section 3.6) to a fifth architecture: Gemma-4-12B-it (Gemma4UnifiedForConditionalGeneration, 48 layers, ≈3× Gemma-4-E4B). The gate fails on all six conditions — LOCO $R^2 = -0.148$ at the peak-readability layer (L19), bootstrap 95% CI $[-0.44, -0.06]$, permutation $p = 0.176$, raw-centroid RDM/Mantel flat, text-only LOCO ≤ 0 — with no layer among the 48 exceeding LOCO $R^2 > 0.10$. Scale amplifies *in-distribution readability* (KFold $R^2 = 0.449$ exceeds Gemma-4-E4B's 0.33) yet LOCO collapses harder (-0.148 vs -0.047): a larger model memorizes country identity more cleanly and still does not abstract, literalizing the claim that the extracted direction is a country-identity direction rather than a wealth axis.

A.11 L1 birth is transformer-computed, not a vision-encoder artifact

A natural concern is that the L1-readable country-identity direction is inherited from the connector-projected vision-encoder output (patch embeddings) rather than computed by the transformer. We rule this out by ablation (Figure 5): probing GDP on the pre-LM representation (connector output at image-token positions, before any LM layer) recovers worse than the mean (5-fold $R^2 = -0.40$), whereas L6 recovers it, and the L1/L6 wealth directions are orthogonal ($\cos \approx 0.01$–0.03) to the pre-LM visual direction. Country-identity is therefore constructed in the residual stream, not handed to it by the vision tower; "born at L1" is a mechanistic claim about transformer emergence.

A.12 Variable Dependence Spectrum: full taxonomy

Section 8 compresses the VDS taxonomy to its endpoints. The full classification is applied at each variable's peak-KFold layer:

Type	Name	Criterion
A	visually direct	KFold $R^2 \geq 0.4$, LOCO $R^2 \geq 0.3$, gap < 0.2
B	partial generalization	KFold $R^2 \geq 0.4$, LOCO $R^2 \in [0.15, 0.3)$
C	weak readability	KFold $R^2 < 0.4$

Type	Name	Criterion
D	country-identity-bound	KFold $R^2 \geq 0.4$, LOCO $R^2 < 0.15$, gap ≥ 0.3

Under this taxonomy Pollution is Type A (KFold 0.50, LOCO $+0.32$), Quality of Life is Type D (KFold 0.44, LOCO $+0.01$), and GDP PPP and Crime sit at the C/D boundary (KFold just below the 0.4 cutoff with fully collapsed LOCO). Type B is the intermediate band: no variable in our study falls there, so its erasability is untested — we recommend per-decision audit rather than extrapolation from either endpoint. Type C variables are too weakly readable for the diagnostic to apply.

A.13 Closed-form linear erasure: protocol and per-cell results

Protocol (pre-registered). Four closed-form erasers — mean-difference projection (top/bottom GDP-tercile contrast), single-direction nullspace (GDP ridge direction fitted on the selection split), INLP (iterated nullspace projection of a 9-way country logistic probe, $K \leq 35$, stop at majority accuracy), and LEACE (closed-form oblique projection $P = I - \Sigma^{-1}A(A^\top\Sigma^{-1}A)^{-1}A^\top$ with covariance shrinkage $10^{-3}\,\mathrm{tr}(\Sigma)/d$) — plus a matched-rank random-directions control are fitted on text-token pooled states of the 238-image selection split and applied online during generation via a float32 forward hook at the readout layer of each architecture (Gemma-4 L6, Qwen3-VL L7). Behavioral endpoints on the 200 held-out images: the GDP-estimation gap between clean top/bottom terciles and the mortgage approval-rate gap over the same terciles. The verdict rule was frozen before data collection: a method "works" only if the paired-bootstrap gap reduction is significant at α=0.01, exceeds the TOST equivalence bound (SESOI = $0.25\times$ clean gap), and passes capability gates (parse rate ≥ 0.80, clean-stop, dead-hook perturbation check); cells failing the parse floor are classed as capability-destroyed and excluded from the pooled estimate.

Per-cell outcomes (relative GDP-gap change, bootstrap B=2000):

Cell	Verdict	Δgap
Gemma-4 × mean-diff / single-dir	inert	-2.2% / -3.4%
Gemma-4 × LEACE	inert (perfect erasure)	-16.2%
Gemma-4 × INLP	capability-destroyed (99% refusals)	—
Gemma-4 × random dirs	capability-destroyed	—
Qwen3-VL × mean-diff	inert	-0.7%
Qwen3-VL × INLP	apparent success	-24.5%
Qwen3-VL × random dirs	control reproduces it	-20.4%
Qwen3-VL × LEACE / single-dir	capability-destroyed	—

Geometrically, LEACE removes all linear recoverability on both architectures (fresh-probe $R^2 = -0.004$, 9-way accuracy 0.08 at chance); INLP reaches majority accuracy with residual GDP R^2 of 0.12/0.33; the rank-one baselines leave the concept fully recoverable (redundancy). A text-token-only sensitivity variant does not rescue the destroyed cells. The INLP-on-Gemma-4 refusals are coherent ("I cannot estimate the GDP per capita … based solely on this photograph") while mortgage decisions remain parseable — the model declines to estimate wealth rather than losing the ability to answer. Finally, the projection-drop metric used for the PGD/UAP classes is uninformative here: pooled states are nearly orthogonal to d_{gdp} (mean $|\langle h, d\rangle| \approx 10^{-4}\|h\|$), so projection ratios are ratios of near-zeros; recoverability by a fresh probe is the appropriate geometric readout for closed-form erasure. Pooled over valid cells the class moves the GDP gap by -9.4% (99% CI $[-12.3\%, -7.1\%]$), equivalent within the pre-registered bound, with MDES 3.3%.

REFERENCES

Authors anonymized (under review). The encoding-behavior dissociation: How distributed safety representations are channeled through a single-direction refusal gate in vision-language models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=GGFjPyO3mZ. OpenReview GGFjPyO3mZ. Under review (double-blind).

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Zehao Deng, Tianjie Ju, Zheng Wu, Liangbo He, Jun Lan, Huijia Zhu, Weiqiang Wang, and Zhuosheng Zhang. Causal probing for internal visual representations in multimodal large language models, 2026.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics (TACL)*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. Argues that probing the presence of information does not establish causal use.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, et al. Toy models of superposition. Anthropic, transformer-circuits.pub, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

X. Geng, Y. Xiao, Y. Zhang, H. Wang, X. Liu, R. Min, T. Fang, and Y. R. Fung. Geobrowse: A geolocation benchmark for agentic tool use with expert-annotated reasoning traces, 2026.

Michael Golovanevsky et al. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of EMNLP*, 2025.

Jentse Huang, Jingyuan Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. Ai sees your location, but with a bias toward the wealthy world. In *Proceedings of EMNLP*, 2025.

P. Jackson, R. Li, and E. Edelstein. Nationality encoding in language model hidden states: Probing culturally differentiated representations in persona-conditioned academic text, 2026. Establishes nationality encoding in text-generating language models; leaves erasure as future work.

Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. doi: 10.1126/science.aaf7894. Leave-one-country-out validation across five African countries.

Sheng Liu, Yuhao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Huan Chen, Xing Xie, and Jianxin Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*, 2022.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for large multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Multi-layer visual-token stacking (LMM + ViT).

Satiyabooshan Murugaboopathy, Connor T. Jerzak, and Adel Daoud. Platonic representations for poverty mapping: Unified vision-language codes or agent-induced novelty?, 2025. Multimodal poverty mapping requiring out-of-country and out-of-time generalization (R^2 0.77 vs 0.63), continuing the leave-one-region-out standard of Jean et al. (2016).

Maciej Pach, Sridhama Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Proceedings of NeurIPS*, 2025.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, et al. The hidden dimensions of llm alignment: A multi-dimensional analysis of orthogonal safety directions. In *Proceedings of ICML*, 2025.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 2020. Method INLP.

Shauli Ravfogel, Valentina Vyas, Ryan Cotterell, and Yoav Goldberg. Linear adversarial concept erasure. In *Proceedings of ICML*, 2022.

S. Sahoo, V. Jain, A. Chadha, and D. Chaudhary. Linear probes detect task format, not reasoning mode in language model hidden states, 2026. Concurrent work; argues probing accuracy can reflect format artifacts rather than the target property.

Alexandre Saillenfest and Peter Lemberger. Nonlinear concept erasure: A density matching approach. In *Proceedings of ECAI*, 2025. Method name LEOPARD.

Anonymous. A surrogate-validity field guide for representation-level interventions, 2026a. Companion paper; surrogate-validity evaluation of representation-level interventions.

Anonymous. Necessity without sufficiency: The read-write boundary for computation directions in llms, 2026b. Companion paper; the read–write writability boundary across the necessity spectrum and across linear and invertible-nonlinear actuators.

Paredes La Torre et al. Adversarial attacks against modern vision-language models, 2026. Reports 6.5–15.5% attack success rate for Qwen2.5-VL versus 52–67% for other VLMs.

A. T. Wasi, S. Z. Ridoy, K. A. Tonmoy, K. Tshering, S. M. M. Hasan, W. Faisal, T. Mohiuddin, and M. R. Parvez. Timespot: Benchmarking geo-temporal understanding in vision-language models in real-world settings, 2026.

Jiancong Wu, Yifan Xie, Shibo Lin, Shijie Zhao, and Xiangyan Chen. Knowing without acting: The disentangled geometry of safety mechanisms in large language models, 2026.

Yongzhong Xu. Global low-rank, local full-rank: The holographic encoding of learned algorithms, 2026.

C. Yang, Y. Jiang, Y. Deng, and C. Wu. Skill-conditioned visual geolocation for vision-language models, 2026.

Haeun Yu, Seunghyun Jeong, Sachin Pawar, Jinwoo Shin, Jinyoung Jin, Jongsu Myung, Alice Oh, et al. Entangled in representations: Mechanistic investigation of cultural biases in large language models, 2025. Method named CultureScope.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of SIGIR*, 2024. Method named ETU (Effective and Transferable Universal).