CORRESPONDENCE, NOT SIMILARITY: GATING MULTI-VIEW EVIDENCE IN DRIVING VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving vision-language models are given several cameras, and a growing number of evaluations conclude that a model used the extra view. The usual evidence is a *matched negative*: show the model a comparable image containing no target object, and check that its answer returns to baseline. We show that this control certifies less than it is taken to certify. Holding scene, timestamp, ego pose and ground truth fixed across 610 Paired Exposure Units built from nuScenes, we replace the true side camera with one from another vehicle's world, matched on city, hour, weather and road-user status. Across five vision-language models the true and the substituted view prove equally persuasive, and two models find the substituted view *more* persuasive. The distinction is easy to draw from pixels, where an off-the-shelf matcher separates the two at 0.973 AUROC. We benchmark ten verification methods and find that those measuring correspondence keep the evaluation valid at every difficulty, while those measuring similarity either leave it invalid or hold it only by admitting a third of the foreign views, regardless of supervision. Gating on correspondence needs no extra generation. We release the benchmark, the pools, and all $122{,}000$ responses.

1 INTRODUCTION

A driving vision-language model (VLM), a system that reads camera images alongside a text question and answers in text, is usually given more than one camera. Evaluations of such models routinely conclude that the model *used* the additional view. They license that conclusion with a control known as a **matched negative**: the model is shown a comparable image in which the object of interest is absent, and if its answer falls back, the extra camera is credited with the difference. We find controls of this shape wherever a multimodal system is asked whether it really looked (Goyal et al., 2017; Xie et al., 2025; Zafar et al., 2026).

This control certifies something weaker than what it is used to certify. A matched negative varies whether the target *content* is present in the added image, not whether that image is a picture of *the scene the question is about*. We call the second property **provenance**, and we show that the two come apart. An image can contain a pedestrian, from the right camera, in the right city, at the right hour, and still be a photograph of somewhere else.

The models do not notice. We run five open-weight VLMs over 610 units whose scene, timestamp, ego pose (where the vehicle is and which way it faces) and hidden ground truth we hold frozen, and we compare the effect of the true synchronized side view against the effect of the substituted one. The two are indistinguishable: our correspondence-specific causal effect is not positive in 13 of 20 model$\times$difficulty cells, and in both Qwen3-VL models it turns negative, so the foreign view moves the answer *more* than the true one does. The evidence needed to catch the swap is not hidden in the pixels, where a feature matcher published in 2021 tells our true pairs from our substituted ones at 0.973 AUROC on the hardest pool we build. The models never look for it, and the control meant to certify that they did cannot see the difference either.

The obvious repair fails. If a model will not verify provenance, we can verify it on the model's behalf, scoring the two views against each other and refusing the pair when they do not belong

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

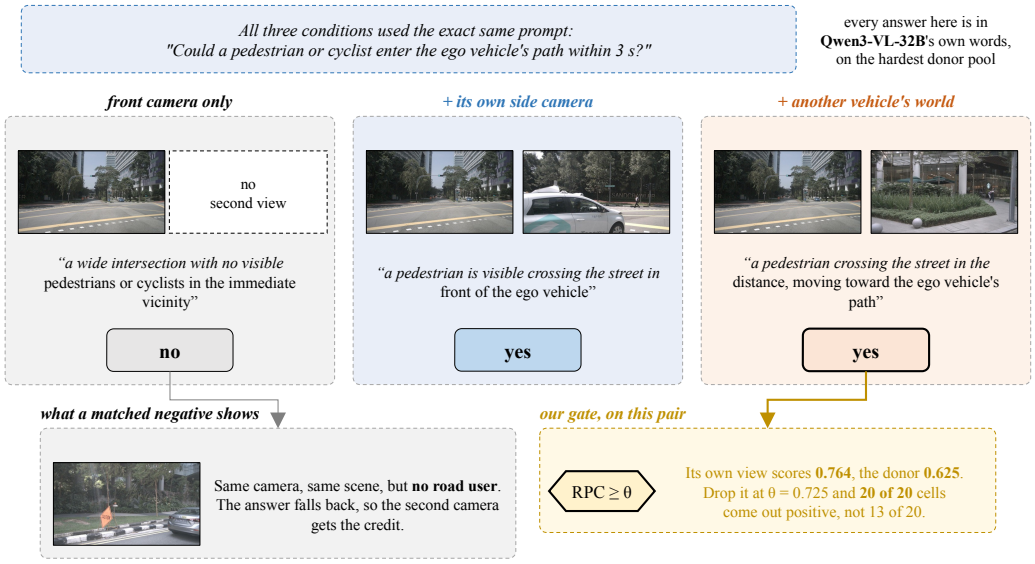

Figure 1: **The control that licenses "the model used the second camera" tests content, not provenance.** One unit of our benchmark, in Qwen3-VL-32B's own words. Shown the front camera alone the model commits to *no*; shown this vehicle's synchronized front-left camera, *yes*; shown a front-left image from a different vehicle kilometres away in the same city and daylight, *yes* again, describing a pedestrian who is not in this scene. A matched negative removes the pedestrian instead, so the answer falls back and the control is passed. Refusing views that fail a correspondence check turns 13 failing model×difficulty cells into none. Photographs and quotations are real.

together. The instrument a practitioner reaches for first is embedding similarity, from CLIP, SigLIP2 or DINOv2, and we find that it works exactly as long as the substitution is easy. As we match the substituted view on city, then on lighting and weather, then retrieve it for maximal resemblance, every global embedding we test falls to the level of a colour histogram, and gating on one still leaves the measurement negative in a quarter to a third of the cells. Supervision does not rescue it. A linear classifier we train on this very task, on the same features, with labels, is still negative in 5 cells, and on the hardest pool all three visual place recognition models we test score below a keypoint pipeline from 2004.

What works, and why. The axis that separates these methods is not how much supervision they received but what they measure. We distinguish **similarity**, one number summarising how alike two images look overall, from **correspondence**, an explicit set of point-to-point matches between the two images together with a test of whether those matches are geometrically consistent. A well-chosen foreign scene can look alike; it cannot supply a consistent set of matching surfaces, because there are none to find. Every method we test that measures correspondence keeps the evaluation valid at every difficulty. No method that measures resemblance matches that: a global embedding leaves the measurement negative however it was trained, and a place recognition model holds the sign only while admitting 31 to 43% of the foreign views it was asked to reject.

Contributions.

1. **A graded wrong-world benchmark** of 610 units on nuScenes (Caesar et al., 2020), with four donor pools of increasing difficulty and a mining criterion kept deliberately disjoint from the quantity any gate scores (§3).
2. **A benchmark of ten verification methods** spanning image statistics, global embeddings, a supervised head, three place recognition models and two feature matchers, each choosing its own *readout*, the particular score it reduces a pair to, on the same fitting data ours uses (§5.1).

3. **Correspondence gating**, and the result that decides between the families: we recover a positive CSCE in all 20 cells, reaching 0.196 on the hardest pool while admitting 15.6% of foreign views (§5.2).

4. **RPC**, a training-free correspondence score over frozen patch tokens whose per-pair cost, after caching frame features, is a single matrix product (§4).

2 Related work

Visual place recognition. NetVLAD (Arandjelović et al., 2016) introduced learned aggregation of local descriptors for deciding whether two images show the same place, and CosPlace (Berton et al., 2022), MixVPR (Ali-bey et al., 2023), EigenPlaces (Berton et al., 2023) and SALAD (Izquierdo & Civera, 2024) train on millions of geotagged street-level images. These models are the natural off-the-shelf answer to our question and we report three as baselines. Their objective, however, is retrieval invariance: two views of one place should match *despite* differences in viewpoint, season and hour. Our substitution inverts that setting: it holds the conditions the objective learns to see past and removes the geographic identity it learns to keep, leaving the aliasing case that invariance was never asked to separate. This is why a model trained for the right task still admits a foreign view. AnyLoc (Keetha et al., 2024) builds such descriptors on frozen features; we test the supervised line.

Local feature matching. The alternative is to establish correspondences and verify them geometrically. SIFT (Lowe, 2004) with a ratio test and a RANSAC fundamental matrix (Fischler & Bolles, 1981), a robust fit that discards matches inconsistent with a single camera geometry, is the classical pipeline; SuperGlue (Sarlin et al., 2020), LightGlue (Lindenberger et al., 2023) and the detector-free LoFTR (Sun et al., 2021) are the learned ones, and dense matchers have gone further since, RoMa (Edstedt et al., 2024) and DUSt3R (Wang et al., 2024), the natural candidates for closing the gap we report to LoFTR. Neighbourhood consensus (Rocco et al., 2018) observes that correct matches agree with their spatial neighbours while spurious ones do not, and we reuse it directly. We show that this family, rather than the retrieval family, is what a provenance check needs.

Controls for evidence use. Balanced image pairs that share a question but not an answer were introduced to stop VQA models answering from language priors (Goyal et al., 2017), and the idea recurs wherever a benchmark may be answerable without looking: blank or shuffled inputs expose driving models that answer confidently without pixels (Xie et al., 2025), and hard-negative grounding in medicine shows a model scoring well while attending elsewhere (Zafar et al., 2026). Closest to us, Hsieh et al. (2023) find hard negatives distributionally separable across compositionality benchmarks, so a blind model beats a sighted one. A blank frame removes the evidence; a foreign frame keeps it and removes only its provenance.

Why a VLM might fail this check. Affirmation bias (Li et al., 2023; Guan et al., 2024; Chan et al., 2026), poor handling of negation (Alhamoud et al., 2025), unstable resolution of text-image conflict (Li et al., 2026) and sensitivity to prompt formatting (Sclar et al., 2024; Zhuo et al., 2024) could all be mistaken for the effect we measure, and §3 says what we do about each. Two we cannot design around: a model need not attend to the region that determines its answer (Liu et al., 2025), and object hallucination traces to identifiable circuits rather than to missing visual input (Li et al., 2025; Liu et al., 2026b).

Multi-view and causal evaluation for driving. Recent work targets cross-view grounding (Wang et al., 2026a; Yan et al., 2026), asks which view an answer came from (Wang et al., 2026b), probes spatial reasoning over nuScenes (Tian et al., 2025; Fruhwirth-Reisinger et al., 2025), measures how much visual evidence a model absorbs (Liu et al., 2026a; Zhao et al., 2026), or tests causal and structured visual reasoning (Komanduri et al., 2025; Chen et al., 2025; Wüst et al., 2025). All of them ask whether a model reaches the right answer. We ask whether it was *entitled* to that answer, and supply the check that makes the entitlement testable.

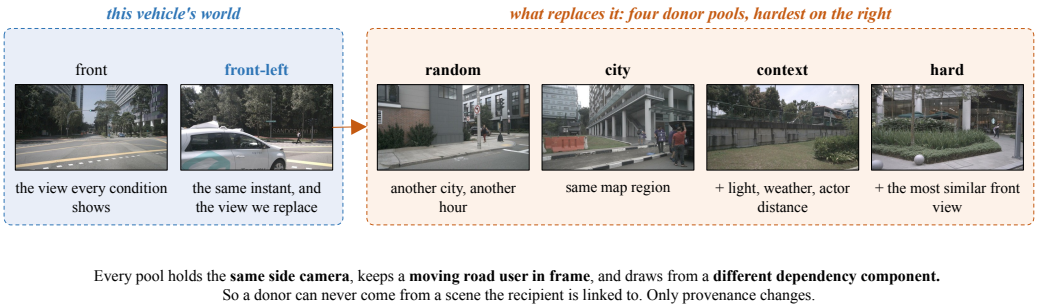

Figure 2: **Each rung of the ladder keeps one more of the cues a matched negative can inspect.** Left, one unit: the front camera and the synchronized front-left camera the substitution replaces. Right, the four donors the released pools assign to it. By the rightmost pool the donor is the same city at the same hour on the same kind of street, and provenance is all that separates it from the view it replaces.

3 THE WRONG-WORLD BENCHMARK

Paired Exposure Units. The unit of our benchmark is a Paired Exposure Unit (PEU), a pair of nuScenes *keyframes*, the annotated moments the dataset releases at 2 Hz, that share a scene and a side camera. In the *positive* member, a moving pedestrian or cyclist appears in the side view and not in the front view; in the *matched negative* member, the same side camera of the same scene is revealed and contains no road user. An *exposure ladder* then varies what the model sees: the front camera alone (LOW), the front camera plus this unit's own synchronized side camera (HIGH), or the front camera plus a side view from another vehicle's world (WRONG-WORLD). Scene, timestamp, ego pose and hidden ground truth stay frozen, and only the revealed camera set changes. Two known sensitivities shape the rest. Models affirm more readily than they deny (Li et al., 2023; Chan et al., 2026) and read negation poorly (Alhamoud et al., 2025), so we ask every unit in both polarities and average over them; they move with prompt formatting alone (Sclar et al., 2024; Zhuo et al., 2024), so schema, verdict set and surrounding text are one fixed template in which only the asserted sentence changes. Every probe offers `true`, `false` and `insufficient` as equally acceptable verdicts, and our outcome is the *commitment rate*: the share of answers taking `true` or `false` over the abstention the prompt licenses, read from the verdict and never from a confidence field.

What we estimate. For each model we take a difference-in-differences, the change in commitment rate on positives minus the change on matched negatives, which removes any drift that affects both members of a unit equally. Writing τ_{true} for this quantity on LOW→HIGH and τ_{wrong} for LOW→WRONG-WORLD, the quantity we report is their contrast,

$$\text{CSCE} \;=\; \tau_{\text{true}} - \tau_{\text{wrong}},$$

the *correspondence-specific causal effect*. Both arms share one bootstrap, so the correlation between them survives into the interval. A value near zero says the measured effect is carried entirely by the presence of target content, wherever it came from, and a valid multi-view evidence claim requires $\text{CSCE} > 0$. We also report the *wrong-world leverage ratio* $\text{WLR} = |\tau_{\text{wrong}}|/|\tau_{\text{true}}|$, which is 0 when a measurement is entirely about this vehicle's world and 1 when a foreign view is exactly as persuasive as the true one.

Dependency-aware splitting. We call two units *dependent* if anything links them that a model could exploit to answer one from the other. We count six such links: a shared scene, temporal adjacency, a shared actor instance, a shared released frame, a perceptual near-duplicate (dHash Hamming distance ≤ 4), and a shared recording log. We build the graph of these links and allocate whole connected components, which we call *dependency components*, in a $65/15/20$ split. Splitting by scene is not sufficient: near-duplicates collapse 259 scenes into 241 components, and 40 recording logs straddle a scene-level boundary. The released allocation leaves zero dependency edges across splits.

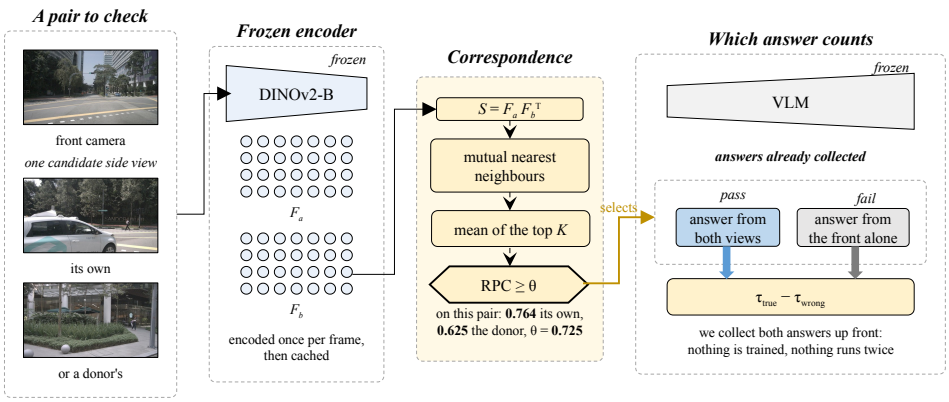

Figure 3: **The gate reads correspondence off frozen patch tokens and then decides which of two *already collected* answers counts.** Nothing in the diagram is fine-tuned: the encoder is frozen, the VLM is frozen, and the only free parameters are the top-K fraction and the threshold, both fixed on the RANDOM pool. The crops and the two scores are the running example of Figure 1.

The graded ladder. How hard a substitution is to catch depends entirely on how similar the donor is, and a negative drawn without that care is separable on its own statistics (Hsieh et al., 2023), so we build four pools of increasing difficulty. Every pool holds the same side camera and the same road-user status as the unit it replaces, and draws from a different dependency component. Each then removes one further cue that is not correspondence.

RANDOM draws any foreign side view, and only 30.7% of its donors happen to share the recipient's map region; CITY matches that region, so all of the remaining three pools sit at 100%. CONTEXT adds day or night, wet or dry, and the actor-distance bin, then keeps a random survivor by a seeded derangement. HARD draws from that same candidate set and keeps its most confusable member instead, by the rule below. Figure 2 shows all four for one unit.

All 610 units receive a donor in all four pools, no unit draws from its own dependency component, and where a strict context group runs dry we relax one cue at a time and record which rung the match landed on, with $577/610$ units landing on the strictest. We run all four pools as a complete ladder from one shared random stream, so LOW and HIGH are the *same responses* in every pool. This makes τ_{true} identical across the difficulty ladder by construction, and places every difference between pools in the wrong-world arm alone.

One design decision decides whether HARD is honest. To build HARD we retrieve, for each unit, the donor whose *front* view most resembles the recipient's *front* view, and never the resemblance between the recipient's front view and the donor's *side* view, which is the pair a gate actually scores. Mining on the scored pair would hand every gate the donor its own method is most wrong about, and the number would report the miner rather than the method. Because "most confusable" still depends on who is asked, we build the pool once per mining method and report the full mining×scoring matrix in Appendix A.

4 METHOD: CORRESPONDENCE GATING

The policy. Our gate is a function of the two views and a threshold. If the added view passes, the model answers from both views; if it fails, the model answers from the front view alone. Both branches already exist in the response set, so applying a gate is a re-selection over responses the models have already produced, with no fine-tuning, no auxiliary head and no second inference pass. When we refuse a pair the model sees only the front view, so its answer *is* the LOW answer and its contrast against LOW is zero by construction rather than by assumption. The only choice left is what the gate scores, and §5.2 shows that this choice, not the supervision behind it, decides how much the gate recovers.

Reciprocal patch correspondence. We instantiate the gate with a score that needs no training and no pairwise network. Let $F_a \in \mathbb{R}^{N\times d}$ and $F_b \in \mathbb{R}^{M\times d}$ be the ℓ_2-normalised *patch tokens* that a frozen vision transformer emits, one d-dimensional vector per image patch, and let $S = F_a F_b^\top$ hold the cosine similarity of every patch of a against every patch of b. Write $\mathrm{fwd}(i) = \arg\max_j S_{ij}$ and $\mathrm{bwd}(j) = \arg\max_i S_{ij}$, and let $\mathcal{M} = \{\, i : \mathrm{bwd}(\mathrm{fwd}(i)) = i \,\}$ be the set of *mutual nearest neighbours*, the patches that pick each other as their best match in both directions. With $K = \lceil \kappa N \rceil$,

$$\textsc{rpc}(a, b) \;=\; \frac{1}{K} \sum_{i \in \text{top-}K(\mathcal{M})} S_{i,\mathrm{fwd}(i)},$$

zero-padded when fewer than K mutual matches exist. The score is therefore a *rate* of confident mutual matches and does not grow with resolution.

Both ingredients carry weight. Mean best-match similarity over all patches reaches only 0.778 on HARD and the fraction of mutual matches only 0.740, against 0.836 for the two combined (Table 4). The *number* of mutual matches barely separates the pools at all, at 45.0 per true pair against 44.1 per wrong one, so the signal lies in where the matches land rather than in how many of them there are.

Everything free is chosen once. We rank backbone, scoring variant, K and θ on the **train and validation components of the RANDOM pool only**. No harder pool, no test component and no downstream response enters any of it. DINOv2-large leads at 0.969; we run the paper on DINOv2-base at 224×392, a 16×28 patch grid, $\kappa = 0.10$ and $\theta = 0.725$, 2.2 points behind at under a third of the parameters and priced on every pool in Table 4 at 0.7 to 1.9 points.

Three variants that did not work. Table 4 has them all. Lowe's ratio test is worse at every K (0.701 against 0.836), because on a 16×28 grid a patch's second-best match is almost always the spatial neighbour of its best; local consensus (Rocco et al., 2018) gives 0.813, and a RANSAC fundamental matrix over the mutual matches, the step that makes SIFT work, gives 0.835. Frozen patch tokens saturate as a correspondence signal near 0.84 here, so the gap to a trained matcher is not an aggregation gap.

Cost. We encode each frame once and cache it, leaving one matrix product per pair. LoFTR caches only its convolutional backbone: the coarse and fine feature maps are per-frame, but the attention that decides the matches runs on the pair, so the bulk of its 99 ms is paid again for every candidate. Where one front view is checked against many candidates that difference is the whole cost of the gate; on a single cold pair the two are within a factor of two.

5 EXPERIMENTS

Setup. We run five open-weight VLMs, Qwen3-VL 4B and 32B, Gemma3 4B, MiniCPM-V 8B and LLaVA 7B, over the full exposure ladder on all four pools, for $122{,}000$ responses in total, decoded at temperature 0 with a fixed output schema. Parse rates are 1.000 for four of the five models and 0.995 for LLaVA. We score every verification method on the identical $2{,}927$ distinct pairs, and every method, ours included, selects its readout and its threshold on the train and validation components of the RANDOM pool alone (494 pairs), so the harder pools, the test components and every downstream response are transfer for all methods equally. Intervals resample the 49 dependency components.

The models do not check. Before any gate, CSCE is indistinguishable from zero on RANDOM for every model, and on HARD both Qwen3-VL models are *negative* (-0.230 and -0.220), meaning a side view from another world moves them further than the true synchronized one. Asked head-on whether two views show the same scene, with no hazard mentioned and nothing else to do, no model performs above chance.

5.1 CROSS-VIEW VERIFICATION

Each unit contributes two pairs to a binary problem the field already has metrics for: a positive pair, the front view with this unit's own side view, and a negative pair, the front view with the donor side

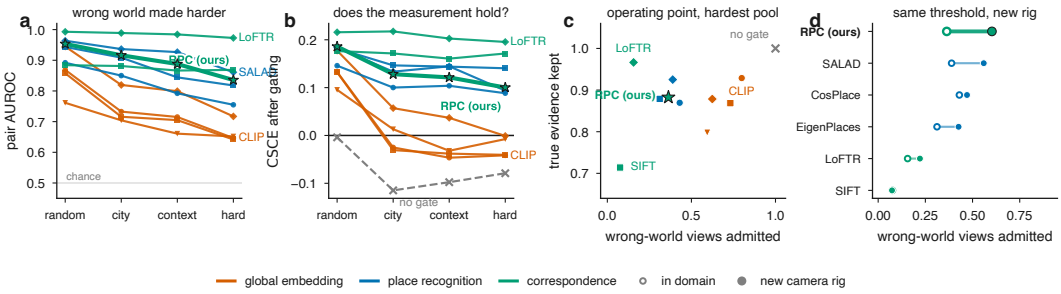

Figure 4: **One distinction runs through every panel, and it is not supervision.** Hue carries the *family* of measurement, marker and label the method, and ours is drawn heaviest. **(a)** As the wrong world is matched on city, then lighting and weather, then retrieved for resemblance, every global embedding falls towards chance while correspondence holds. **(b)** The same split appears downstream: a global-embedding gate lifts CSCE and still leaves it negative on every matched pool. **(c)** On HARD, up and to the left is better, and the correspondence gates are more selective *and* more retentive. **(d)** Carried to a different vehicle, a count of verified correspondences barely moves and a continuous score does; the six methods are those we could run on Argoverse 2.

Table 1: Cross-view verification AUROC, 610 true against 610 wrong pairs per pool. Best in **bold**, second underlined.

			AUROC ↑, by how well the wrong world is matched			
Method	Sup.	ms/pair	RANDOM	CITY	CONTEXT	HARD
Image statistics						
Colour histogram	–	–	0.761	0.704	0.661	0.652
Global embedding similarity						
CLIP ViT-L/14	–	–	0.869	0.733	0.715	0.646
SigLIP2	–	–	0.848	0.719	0.677	0.634
DINOv2-B	–	–	0.858	0.716	0.705	0.642
Supervised head, DINOv2-B	✓	–	0.945	0.820	0.799	0.717
Visual place recognition						
CosPlace	✓	87.2	0.892	0.850	0.792	0.755
EigenPlaces	✓	90.3	0.944	0.909	0.844	0.818
SALAD	✓	95.7	<u>0.964</u>	<u>0.936</u>	<u>0.927</u>	0.861
Local feature matching						
SIFT+RANSAC	–	75.6	0.884	0.881	0.867	<u>0.869</u>
LoFTR	✓	99.4	**0.993**	**0.989**	**0.984**	**0.973**
Ours						
RPC (ours)	–	2.1	0.953	0.916	0.888	0.836

view. We report AUROC, the probability that a method scores a randomly drawn true pair above a randomly drawn wrong one, where 0.5 is chance.

Table 1 compares ten methods in four families on the identical pairs: a sky-and-ground colour histogram as a floor that uses no semantics; global cosine similarity of CLIP (Radford et al., 2021), SigLIP2 (Tschannen et al., 2025) and DINOv2 (Oquab et al., 2024) embeddings, plus a logistic classifier we train on DINOv2 features with labels from the fitting pool; place recognition with CosPlace (Berton et al., 2022), EigenPlaces (Berton et al., 2023) and SALAD (Izquierdo & Civera, 2024); and correspondence, SIFT (Lowe, 2004) with a ratio test and a RANSAC fundamental matrix (Fischler & Bolles, 1981), the detector-free LoFTR (Sun et al., 2021) where SuperGlue (Sarlin et al., 2020) and LightGlue (Lindenberger et al., 2023) are not, and our RPC.

Similarity collapses; correspondence does not. On RANDOM, where the wrong world may come from another city, every family works: even the colour histogram reaches 0.761 and CLIP reaches 0.869. One rung up the ladder the two halves of the table separate. By HARD, every global embed-

Table 2: Gating the second view. CSCE after gating, averaged over five models, with the share of true pairs kept and wrong pairs admitted on HARD. The last column counts the cells a gate leaves no better than no check at all.

	on HARD		CSCE after gating ↑				
Gate	kept$^+$	adm.$^-$	RANDOM	CITY	CONTEXT	HARD	CSCE ≤ 0
No gate	1.00	1.00	-0.004	-0.115	-0.098	-0.079	13/20
Gating on a global embedding							
Colour histogram	0.80	0.59	0.095	0.013	-0.032	-0.008	7/20
CLIP ViT-L/14	0.93	0.80	0.133	-0.025	-0.046	-0.042	6/20
DINOv2-B	0.87	0.73	0.132	-0.031	-0.038	-0.041	7/20
Supervised head, DINOv2-B	0.88	0.62	0.177	0.057	0.037	-0.001	5/20
Gating on place recognition							
CosPlace	0.87	0.43	0.146	0.100	0.104	0.088	0/20
EigenPlaces	0.88	0.31	0.182	0.146	0.144	0.140	0/20
SALAD	0.93	0.39	0.184	0.133	0.145	0.094	0/20
***Correspondence gating (ours)**, scored by*							
RPC (training-free)	0.88	0.36	0.186	0.128	0.121	0.100	0/20
SIFT+RANSAC	0.71	0.08	0.176	0.171	0.161	0.171	0/20
LoFTR	0.97	0.16	**0.216**	**0.217**	**0.203**	**0.196**	0/20

ding sits between 0.63 and 0.65, *below* the sky-and-ground histogram at 0.652, while the two matchers hold 0.869 and 0.973. Once the two worlds look alike, a 400M-parameter language-supervised encoder is worse than a histogram at deciding whether they are one world.

Supervision is not the axis. The classifier we trained on this exact task, on the same DINOv2 features, with labels, reaches 0.945 on the pool it was fitted on and 0.717 on HARD: it learns the fitting pool's regularities rather than correspondence. The three place recognition models, heirs to NetVLAD (Arandjelović et al., 2016) and trained on millions of geotagged street images for the retrieval version of this question, lead SIFT while the substitution is easy and all three fall behind it by HARD: SALAD gives up 0.103 across the ladder where SIFT gives up 0.015. What SIFT has and they lack is an explicit correspondence set and a geometric consistency test on it.

Where our own scorer stands. RPC is the strongest method here needing no training beyond a frozen backbone, and it sits 13.7 points below LoFTR on HARD. None of the three attempts in §4 closes that gap, which locates the limit in the frozen tokens rather than in how we aggregate them.

5.2 Does the gate fix the measurement?

Verification AUROC is a proxy. The quantity that matters is whether a gated measurement supports the claim it is used to license.

Each row of Table 2 is one method of Table 1 at its own fitted threshold, and its last column counts the model×difficulty cells in which the gated CSCE is not positive. Ungated, the measurement does not survive its own control: CSCE is not positive in 13 of the 20 cells and averages -0.079 on HARD. Three tiers separate above it, which Figure 4b shows at every difficulty at once.

A global embedding does not repair it. All four such gates, the supervised classifier included, roughly halve the failing cells and leave five to seven of the twenty negative. Against a matched donor pool they refuse true pairs at close to the rate they refuse wrong ones, stripping genuine evidence alongside the foreign evidence rather than instead of it. This is the repair a practitioner reaches for first, and on a matched pool it moves the measurement without fixing it.

Place recognition works, but leaks. CosPlace, EigenPlaces and SALAD are positive in all twenty cells while admitting 31 to 43% of matched wrong-world views, because retrieval invariance is trained to survive precisely the changes our substitution preserves.

Correspondence gating is best at every difficulty. The two matchers admit the fewest wrong views on HARD, 8% and 16% against 31% for the best place recognition model, and keep the most true evidence: LoFTR keeps 96.6% of true pairs while admitting 15.6%, against CLIP's 92.9% kept and 79.9% admitted (Figure 4c). RPC reaches most of that separation from frozen tokens at 2.1 ms a pair and pays at the operating point, admitting 36% where EigenPlaces admits 31% at forty times the cost. The gain is largest where the ungated measurement is worst.

5.3 Transfer to another camera rig

We fixed backbone, readout and threshold on nuScenes before this section existed. Argoverse 2 (Wilson et al., 2021) is a different vehicle: a seven-camera ring with different optics and aspect ratios, a different annotation vendor and six US cities. Pair verification needs no annotations, since a true pair is two cameras of one rig at one instant and a wrong pair is a side view from another log, so the whole evaluation is a download: 150 validation logs, five timestamps each, for 2,274 frames and 6,000 pairs. Intervals resample the AV2 log, the dependency unit here. Figure 4d plots the operating points and Table 5 gives every cell.

Every method *discriminates* at least as well on AV2 as on nuScenes, and global cosine collapses on the harder pools exactly as it does in domain (0.622 and 0.564). What separates the two halves of the table is instead the false-positive rate, the share of wrong pairs a method admits at its fitted threshold. Carrying its nuScenes threshold across, SIFT admits 7.1% and 7.6% of matched wrong-world views on AV2 and LoFTR 26.5% and 17.7%, against 35 to 62% for the three place recognition models and 43 to 78% for RPC. The split here is not the one the rest of the paper draws. A count of verified correspondences is close to rig-invariant because it counts things that are either there or not, while a score on a scale the optics set moves bodily when the lens moves. RPC emits such a score, locating its evidence by correspondence and reporting an average cosine, so it loses its threshold on a new rig as the resemblance methods do while keeping the discrimination they keep. Any of these scales recentres without annotation (Appendix A); only SIFT and LoFTR need no recentring.

5.4 Ablations

Table 4 separates the parts of the score from the parts of the encoder. Neither ingredient is enough alone, and the backbone is not the story either: applying RPC to CLIP's *own* patch tokens takes that encoder from 0.646 to 0.818, against 2 further points for swapping in DINOv2.

RPC also reads the geometry it is meant to read: its matches concentrate in the half of each frame the two cameras physically share and swap sides between the front-left and front-right pairings, while every wrong pool sits at the centre (Appendix A).

6 Conclusion

A matched negative certifies that a model received another image, not that it received an image of *this* world, and multi-view evidence claims rest on the second reading of a control that only tests the first. We built a substitution that separates them, and five vision-language models failed to make the distinction, two moving further by a view from another vehicle's world than by their own. Gating the second view on correspondence restores the measurement in every cell we tested and costs no extra generation. Every correspondence gate admits fewer foreign views than every resemblance gate however much supervision the latter had, with the margin at the operating point going to the trained matchers rather than to our frozen-token scorer.

Limitations. Our benchmark covers 610 units over 49 dependency components from one dataset's keyframes, 49 is a coarse unit for a bootstrap, and the five models are open-weight and under 32B parameters. Our gate decides between two views and one, so a model that receives a correct view and answers from the wrong part of it (Liu et al., 2025) is outside what we measure. Two results are narrower: the inversion in which a foreign view is more persuasive is specific to Qwen3-VL, and a gate scoring a pair on a continuous scale carries its discrimination to a new rig but not its threshold. RPC is on that side, admitting 43 to 78% of wrong-world views on an unseen rig at our fitted θ against 7 to 27% for the two matchers (§5.3).

REPRODUCIBILITY STATEMENT

The released package holds every input to every number in this paper except image pixels. §3 defines the units, the exposure ladder and the graded donor pools together with the rule that mining is done front to front and never on the pair a gate scores, and Appendix A gives the full mining$\times$scoring matrix. §4 fixes the gate, its two free parameters and the split they are chosen on, and §5 lists the five models, the decoding settings and the bootstrap. The package ships all 1,452 unit definitions, the 610 evaluated here among them, the dependency graph and its allocation, the three donor pools, all $122{,}000$ model responses as JSONL, and the verification scores for all ten methods over every pair, so every table and figure rebuilds without a GPU and without re-running inference. nuScenes and Argoverse 2 pixels are third-party data we do not redistribute; the manifest records each frame by its dataset token together with a checksum and a perceptual hash, and a script rebuilds the exact frames we evaluated from a reader's own copy. A second script re-checks the invariants the design rests on, among them that no dependency component straddles a split boundary and that no donor in any pool comes from the recipient's own component.

ETHICS STATEMENT

This work analyses two public driving datasets and collects no new recordings. The images contain pedestrians and cyclists in public streets, recorded and released by the dataset authors under their own terms, and we redistribute no pixels, so nobody's likeness leaves the upstream release through us. Our derived data is licensed as a derivative of nuScenes and inherits its non-commercial terms. The models we evaluate are open-weight and were run locally.

The result carries a safety reading worth stating. A driving system that answers as confidently from another vehicle's camera as from its own is one whose evidence use cannot be certified by the control the field currently applies. We release the substitution and the gate so that this can be measured rather than assumed. The benchmark scores an evaluation protocol, not a deployed vehicle, and a gate that passes our check certifies correspondence between two views and nothing further about the answer that follows.

AI USE STATEMENT

We used generative AI tools for implementing methods, for feedback on experimental design, for retrieval and discovery of related work, for creating and revising the figures, and for drafting and polishing prose that the author then revised. Every analysis script and every figure-generating script was read, run and checked by the author, and the numbers in the paper are emitted by those scripts and verified against the manuscript by an automated checker over 105 values. Related work found by retrieval was read in the original before being cited, and every reference was checked against its published record. We did not use generative AI tools to propose the hypothesis, to generate synthetic datasets, or to formulate or prove mathematical claims. The $122{,}000$ model responses analysed here are the object of study rather than an authoring aid: they are the raw output of the five open-weight vision-language models under evaluation, collected under the protocol of §3 and released unedited. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. *arXiv preprint arXiv:2501.09425*, 2025.

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. MixVPR: Feature mixing for visual place recognition. In *WACV*, 2023. arXiv:2303.02190.

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. arXiv:1511.07247.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *CVPR*, 2022. arXiv:2204.02287.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. EigenPlaces: Training viewpoint robust models for visual place recognition. In *ICCV*, 2023. arXiv:2308.10832.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. arXiv:1903.11027.

Tsan Tsai Chan, Varsha Suresh, Anisha Saha, Michael Hahn, and Vera Demberg. System-mediated attention imbalances make vision-language models say yes. *arXiv preprint arXiv:2601.12430*, 2026.

Yuefei Chen, Jiang Liu, Xiaodong Lin, and Ruixiang Tang. Distilling counterfactual reasoning from language to vision: Causal graph guided post-training for video understanding. *arXiv preprint arXiv:2511.19923*, 2025.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *CVPR*, 2024. arXiv:2305.15404.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Christian Fruhwirth-Reisinger, Dušan Malić, Wei Lin, David Schinagl, Samuel Schulter, and Horst Possegger. Stsbench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. *arXiv preprint arXiv:2506.06218*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. arXiv:1612.00837.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024. arXiv:2310.14566.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *CVPR*, 2024. arXiv:2311.15937.

Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. AnyLoc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2024. arXiv:2308.00688.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. *arXiv preprint arXiv:2506.11034*, 2025.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. *arXiv preprint arXiv:2511.05923*, 2025.

Sirun Li, Minghao Liu, Ling Dai, Yong Li, Haoxin Lyu, Junting Zhou, and Fan Zhang. Signpost-bench: Benchmarking text-vision conflict resolution in multimodal large language models. *arXiv preprint arXiv:2608.04244*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. arXiv:2305.10355.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *ICCV*, 2023. arXiv:2306.13643.

Anjie Liu, Ziqin Gong, Yan Song, Yuxiang Chen, Xiaolong Liu, Hengtong Lu, Kaike Zhang, Chen Wei, and Jun Wang. The perceptual bandwidth bottleneck in vision-language models: Active visual reasoning via sequential experimental design. *arXiv preprint arXiv:2605.01345*, 2026a.

Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. Dual-pathway circuits of object hallucination in vision-language models. *arXiv preprint arXiv:2605.13156*, 2026b.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. arXiv:2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. arXiv:2103.00020.

Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. arXiv:1810.10510.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. arXiv:1911.11763.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021. arXiv:2104.00680.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. *arXiv preprint arXiv:2504.03164*, 2025.

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

Jie Wang, Guang Li, Zhijian Huang, Chenxu Dang, Hangjun Ye, Yahong Han, and Long Chen. Vggdrive: Empowering vision-language models with cross-view geometric grounding for autonomous driving. *arXiv preprint arXiv:2602.20794*, 2026a.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024. arXiv:2312.14132.

Yimu Wang, Yee Man Choi, Barry Zhang, Mozhgan Nasr Azadani, Sean Sedwards, and Krzysztof Czarnecki. Where does the answer come from? benchmarking view-level visual evidence identification in multi-view mllms for autonomous driving. *arXiv preprint arXiv:2606.09644*, 2026b.

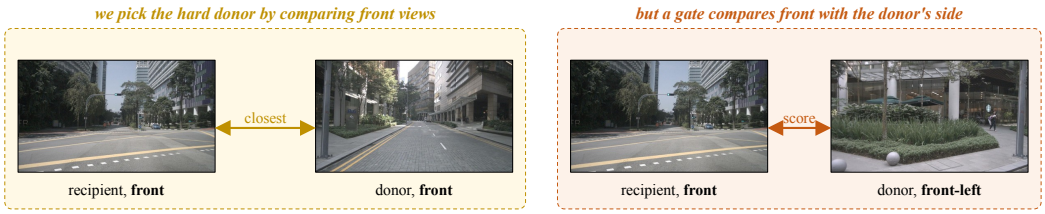

Figure 5: **The pool is retrieved on a pair no gate is ever scored on.** To build HARD we search for the donor whose front view most resembles the recipient's front view (left); every gate in this paper is then asked about the recipient's front view against that donor's *side* view (right). The two front views end up strikingly alike, which is what makes the pool hard, but the search never touched the quantity a gate computes.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2301.00493.

Antonia Wüst, Wolfgang Stammer, Hikaru Shindo, Lukas Helff, Devendra Singh Dhami, and Kristian Kersting. Synthesizing visual concepts as vision-language programs. *arXiv preprint arXiv:2511.18964*, 2025.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025.

Shannan Yan, Leqi Zheng, Keyu Lv, Jingchen Ni, Hongyang Wei, Jiajun Zhang, Guangting Wang, Jing Lyu, Chun Yuan, and Fengyun Rao. Learning cross-view object correspondence via cycle-consistent mask prediction. *arXiv preprint arXiv:2602.18996*, 2026.

Anas Zafar, Leema Krishna Murali, Siddhant Bharadwaj, Ashish Vashist, and Jia Wu. Do medical vision language models actually see? a counterfactual grounding framework and hard-negative contrastive training for visually-reliant medical vlms. *arXiv preprint arXiv:2607.03647*, 2026.

Yifei Zhao, Xiangxin Zhou, Wenhao Yang, Jiaqi Tang, Pu Jian, Huanjin Yao, Jiarui Yao, Haowei Lin, Chunchao Guo, Zhuo Chen, Wenkai Lyu, Jianzhu Ma, Xueqian Wang, and Wenxi Zhu. Sceneactbench: Can agents act on the 3d scenes they see? *arXiv preprint arXiv:2607.22393*, 2026.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. arXiv:2410.12405.

A ADDITIONAL ANALYSIS

A.1 WHO MINED THE POOL

HARD is an argmax, so "hardest" depends on who is asked. A pool mined with one method is adversarial to that method's family and merely difficult for everyone else, and reporting only the pool that happens to favour the proposed method is a second way a benchmark table can mislead. We therefore build the pool three times, mined by CLIP, by DINOv2-base and by SIFT, and score every method on all three. The mining criterion stays front-to-front throughout and is never the front-to-side pair a gate scores, as Figure 5 shows.

Table 3: Mining × scoring on HARD, AUROC over all 610 pairs. † marks an adversarial cell, mined with this method's own encoder.

	pool mined by		
scoring method	CLIP	DINOv2-B	SIFT
CLIP ViT-L/14 global	0.646†	0.639	0.678
DINOv2-B global	0.642	0.592†	0.649
CLIP ViT-L/14 + RPC	0.797†	0.781	0.821
DINOv2-B + RPC	0.836	0.805†	0.846
SIFT + RANSAC	0.869	0.870	0.869†

Two things follow. First, the adversarial cell bites the family it targets. The DINOv2-mined pool is harder for the DINOv2 family (0.592 and 0.805, against 0.642 and 0.836 on the CLIP-mined pool), and the effect is modest. The diagonal is not clean: that pool is also the hardest of the three for both CLIP rows (0.639 and 0.781 against 0.646 and 0.797), so it is harder for everyone rather than adversarial to one encoder. Attacked with a pool built against its own backbone, RPC still scores 0.805 against 0.592 for that backbone's global cosine, so the ordering this paper rests on is not an artifact of who did the mining.

Second, and we did not expect this, **a correspondence method cannot mine a hard pool at all.** SIFT scores 0.869 on the pool mined against it, 0.870 on the DINOv2-mined pool and 0.869 on the CLIP-mined pool, flat to within a thousandth, and its pool is the *easiest* of the three for every other method. Choosing the donor whose front view shares the most verified correspondences with the recipient's front view does not make the recipient's *side* view any easier to confuse, because the quantity under attack is a property of one world and a donor from another world has none of it however it was chosen. A global embedding can be attacked by finding a donor that looks alike; a correspondence count cannot.

A.2 WHERE THE MATCHES LAND

Figure 6 plots where the matches land. In a 28-column patch grid centred on 13.5, the mean horizontal position of the top-K mutual matches in the front camera is 9.6 for true front-left pairs and 16.7 for true front-right pairs, mirrored at 16.4 and 10.4 in the side view: each camera's matches concentrate in the half of the frame it physically shares with the front camera. Every wrong pool sits within 0.9 of the centre in both views, where matches fall when there is no shared surface. Figure 7 in Appendix A shows selected matches on real pairs, including the median case in which the gate gets the decision backwards.

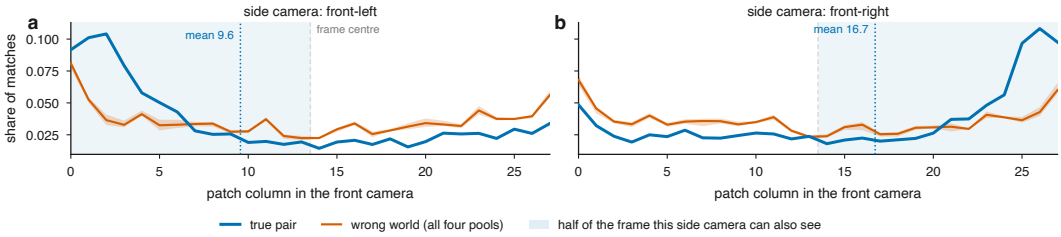

Figure 6: A prediction with a sign that could have come out wrong. CAM_FRONT overlaps the front-left camera on the left of the frame and the front-right camera on the right, so if RPC reads geometry its selected matches must concentrate there and must *swap sides* between (a) and (b). They do. Were it reading style instead, the two panels would look alike and every curve would sit at the centre, which is what all four wrong-world pools do; they are so similar to each other that we draw them as one band.

Table 4: What RPC needs, and what it does not. Reciprocity and spatial selection are both load-bearing; three additions motivated by the rest of this paper are not.

	RANDOM	CITY	CONTEXT	HARD
What the score reads				
Mean best-match similarity, all patches	0.922	0.858	0.833	0.778
Fraction of mutual nearest neighbours	0.859	0.803	0.790	0.740
Mutual + top-K mean (RPC)	0.953	0.916	0.888	0.836
+ local consensus (Rocco et al., 2018)	0.932	0.894	0.859	0.813
+ RANSAC fundamental matrix (Fischler & Bolles, 1981)	0.948	0.910	0.884	0.835
Which frozen backbone, at which resolution				
SigLIP2, 224 px	0.916	0.840	0.805	0.757
CLIP ViT-L/14, 224 px	0.948	0.882	0.877	0.818
DINOv2-B, 336 px	0.945	0.893	0.858	0.811
DINOv2-B, 448 px	0.941	0.889	0.853	0.811
DINOv2-L, 224 px	**0.972**	**0.931**	**0.902**	**0.843**
DINOv2-B, 224 px *(selected)*	0.953	0.916	0.888	0.836

A.3 QUALITATIVE RESULTS

Figure 7 draws the matches the gate actually selects, on one unit it decides correctly and one it decides backwards.



Figure 7: The gate on real pairs, at the threshold fitted on RANDOM. Dots mark the top-K mutual matches, coloured by horizontal position in the front view, so a coherent left-to-right gradient means the two views agree about where things are. The top row is the median-margin correct case and the bottom is the median *failure*, in which the gate refuses the true view and admits the wrong-world one. On HARD it inverts this way on 1.80% of units (11 of 610), against 0.16% on RANDOM, and every such unit is in the released set.

A.4 FULL TABLES

Discrimination transfers for everyone; the operating point does not. The same split as Tables 1 and 2 reappears here, with the correspondence methods holding their threshold across the rig and every similarity method losing it. Raising the input to 448 px gives the score $1{,}792$ patches instead of 448 and scores two points lower, which is why Table 4 keeps the smaller grid.

A.5 CALIBRATING A SIMILARITY WITHOUT LABELS

The repair for the similarity side needs no annotation. We subtract from each front view the median score it achieves against a small set of *unlabelled* target-domain side views. With one offset threshold, again fitted on nuScenes, RPC's AV2 false-positive rate falls from 0.361 to 0.249 on RANDOM at unchanged AUROC. The device costs 0.035 AUROC on the in-domain HARD pool, and the correspondence-counting methods need no such device at all.

Table 5: Argoverse 2 with *nothing* refitted: every method keeps the readout and threshold it selected on nuScenes. FPR@θ is the share of wrong-world views that threshold admits.

	AUROC, RANDOM		AUROC, HARD		FPR@θ, HARD	
Method	left	right	left	right	left	right
Gating on a global embedding						
DINOv2-B global cosine	0.863	0.767	0.622	0.564	–	–
Gating on place recognition						
CosPlace	0.934	0.934	0.773	0.801	0.523	0.419
EigenPlaces	0.974	0.958	0.844	0.859	0.503	0.349
SALAD	0.982	0.982	0.861	0.882	0.616	0.503
***Correspondence gating (ours)**, scored by*						
RPC (training-free)	0.982	0.959	0.874	0.851	0.775	0.431
SIFT+RANSAC	0.992	0.986	0.988	0.980	0.071	0.076
LoFTR	0.999	0.998	0.991	0.988	0.265	0.177