Seeing Neglect: Measuring the Predictive Value of Public Aerial and Street-Level Imagery for Property Code Violations

Erik Aronesty
DirtSignal
Correspondence: aronesty@jhu.edu

June 2026

Abstract

We measure how well freely available imagery (county and USDA aerial orthophotos and Google Street View) predicts open residential code-enforcement violations, scoring hand-crafted color/texture features, zero-shot and aerial-finetuned CLIP, a DINOv2 probe, an open-vocabulary detector (OWLv2), and a small vision-language model (Qwen2-VL-2B) against matched same-neighborhood controls in Hillsborough County, Florida. We report four findings. First, a label-free Street View detector transfers across metros: with no retraining, pooled AUC in Broward County matches Hillsborough (0.58/0.59) and overgrowth replicates (0.63/0.64). Imagery is nearly universal while code-enforcement records are not, so a detector that transfers without local labels is the one that can actually be deployed. In a third metro with dated records (Jacksonville), the signal predates the record: pre-citation Street View predicts next-year first citations (overgrowth AUC 0.61, n=1,198) with significant lead up to 12 months, although it cannot predict citation timing among eventually-cited addresses. Second, model scale is not the lever: a four-number greenness statistic outperforms every deep aerial embedding, and a holistic neglect prompt outperforms targeted object detection. Third, cues are viewpoint-specific: overgrowth is detectable from the air (AUC 0.65) while debris is visible only from the street (0.50 vs 0.62), and the 0.6-0.7 AUC ceiling is consistent with a simple label-noise attenuation model. Fourth, the viewpoints are complementary: fusing them raises every signal, although the gains are not yet significant at n=352. We release an anonymized benchmark, code, and per-parcel features.

Keywords: Street View imagery; aerial imagery; code enforcement; vision-language models; urban physical disorder; transferability

1. Introduction

Distressed residential property is of interest to municipalities (code enforcement, blight remediation), to researchers (neighborhood-change dynamics), and to real-estate practitioners. Much of what makes a property distressed is visible from the outside: an overgrown lot, a green pool, junk and debris, a tarped roof, a derelict vehicle. Both overhead and street-level imagery of most U.S. parcels are now freely or cheaply available. This raises a measurable question. Does public imagery carry predictive signal about property distress, and if so, which signals, from which viewpoint, and with what model?

We treat open code-enforcement violations as an objective, externally-generated proxy for distress, and we measure how well imagery separates violating from non-violating parcels. We evaluate

against matched same-neighborhood controls: a violating parcel against a non-violating parcel about 90 m away. This removes the confound that distressed parcels cluster in older, denser neighborhoods. It is the hardest version of the question, because it asks whether imagery sees the parcel-level condition rather than the neighborhood.

Beyond any one city, imagery is nearly universal and code-enforcement records are not. Thousands of jurisdictions keep code data on paper or not at all. A distress signal that can be extracted from imagery alone is therefore useful regardless of temporality, because it can stand in where the public record does not exist. That shifts the goal from predicting a future citation (a product concern) to a transfer question: does a label-free imagery detector, calibrated where labels happen to exist, generalize to jurisdictions where they do not? We test this directly by replicating in a second, independent county (§5.7), and we test whether the signal exists before the public record does with dated panels in a third (§5.8).

Our contribution is a measurement study and a reusable dataset and pipeline, not a new model. The answer is a qualified yes. The result worth reporting first is that a label-free Street View signal transfers to a second metro. The structure behind that transfer is which signals are visible from which source, how cheaply they can be extracted, and where the current evidence remains underpowered.

2. Related work

Visual and vision-language foundation models. Contrastive language-image pretraining (CLIP) [Radford et al., 2021], scaled in the open via LAION-5B [Schuhmann et al., 2022] and OpenCLIP [Cherti et al., 2023], enables zero-shot recognition from natural-language prompts; self-supervised backbones such as DINOv2 [Oquab et al., 2024] provide strong transferable features for linear probing. For localization, open-vocabulary detectors (OWL-ViT and its scaled successor OWLv2 [Minderer et al., 2023], Grounding DINO [Liu et al., 2024b], and YOLO-World [Cheng et al., 2024]) detect arbitrary text-named objects, and the Segment Anything models (SAM, SAM 2) [Kirillov et al., 2023, Ravi et al., 2024] provide class-agnostic masks. Small instruction-tuned vision-language models such as Qwen2-VL [Wang et al., 2024a] answer free-form visual questions. We use representatives of each family off the shelf; our interest is comparative measurement, not new architectures.

Remote-sensing adaptation. Because web-trained models underperform on overhead imagery, several efforts re-train CLIP on aerial/satellite image-text data: RemoteCLIP [Liu et al., 2024a], GeoRSCLIP with the RS5M dataset [Zhang et al., 2024b], and SkyScript [Wang et al., 2024b]. We include RemoteCLIP as our aerial-finetuned baseline. On the specific, localized distress cues we target, it does not beat a four-number color statistic, which is consistent with the rest of our results.

Street-level imagery and the built environment. Google Street View has been used to audit neighborhood physical disorder in place of in-person surveys [Rundle et al., 2011], to quantify urban perception at scale (Place Pulse) [Dubey et al., 2016], and to measure physical urban change and its socioeconomic predictors [Naik et al., 2017]; recent work surveys this "urban visual intelligence" broadly [Zhang et al., 2024a] and detects urban physical disorder with interpretable transformers [Hu et al., 2023]. Closest to our work, Zou and Wang [2021] detect individual abandoned houses from Street View using façade and overgrown-vegetation patch classifiers (F1 = 0.84 across five shrinking U.S. cities).

Our position. Prior applied work is largely single-viewpoint (street or overhead) and validated

against curated or perception labels. We measure both viewpoints against an objective public-record label (open code-enforcement violations), per distress sub-type, against matched same-neighborhood controls, and we quantify the complementarity and fusion of aerial and street imagery, with an emphasis on which cheap, deployable features actually carry signal.

3. Data

Region. Hillsborough County, Florida (Tampa metro).

Parcels and imagery. Parcel geometry, addresses, and folio identifiers come from the Hillsborough County Property Appraiser ArcGIS service. Aerial imagery is the county's 0.15 m/px orthophoto service (multi-year 2016–2025); we render an adaptive per-parcel window (sized to the parcel footprint, clamped 50–180 m) at the cache's finest scale. Street View is fetched via the Street View Static API. For the matched Hillsborough/Broward panels, we query metadata at the parcel centroid, use the returned outdoor road pano location, and set the image heading to the bearing from that pano toward the parcel centroid. For the Jacksonville parcel-geometry temporal panel, we also query multiple parcel-boundary samples, deduplicate candidate panos, select the pano closest to the parcel boundary, reject far panos, and aim at the nearest parcel-boundary point. This aims the camera at the parcel rather than down the road; it does not guarantee a facade-facing image of the target structure. Coverage is 95% of labeled parcels.

Labels. Open code-enforcement violations come from a production ingestion pipeline (Tampa and Hillsborough County Accela portals), joined to parcels by the 22-character STRAP identifier. A generic "any violation" label is dominated by issues with no aerial signature (permits, fences, interior), so we restrict positives to violations whose free-text description is aerial or ground relevant: overgrowth, accumulation, debris, junk, trash, or inoperable vehicles. This yields 185 distress parcels. We tag each with its sub-type (overgrowth / debris / vehicle) from the description.

Matched controls. For each distress parcel we draw the nearest non-violating parcel (about 90 m median, same block) as a clean control, so the two classes share neighborhood, housing stock, and imagery conditions. This controls out location and makes the task deliberately hard.

Vacant-lot exclusion. An overgrown vacant lot is a different phenomenon from a distressed structure, so we exclude vacant parcels from both sets in two steps: the county assessor land-use code (Hillsborough `LU_GRP`/`DOR_CODE`; Broward `USE_TYPE`), then a zero-building-value filter that catches mis-coded empty lots. This trims Hillsborough from 185/185 to 162/173 and finally to 154 distress / 165 clean (n=319); Broward is 185/184 (no vacant parcels flagged). Vacant lots were over-represented on the distress side (Hillsborough: 23 vs 12 of the clean controls), but excluding them leaves results essentially unchanged: overgrowth-greenness 0.645 to 0.640, pooled SV-VLM 0.588 to 0.580, overgrowth SV-VLM 0.646 to 0.629, and the cross-jurisdiction conclusions are identical. The signal therefore reflects the condition of improved property rather than empty overgrown land. Sections 5.1–5.6 report the original Street View-covered sample (n=352), whose conclusions the exclusions leave unchanged; the cross-jurisdiction comparison (§5.7) uses the use-code-filtered sets; and the reader ladder (§5.9, Appendix B) and the released benchmark tables use the final n=319 set.

A note on dates. A leading-indicator analysis (does neglect appear before the citation?) requires violation open-dates. These are not published by the Tampa/Hillsborough Accela portals (verified by live inspection of the detail pages). Jacksonville exposes a multi-year dated history (about 239k MyJax records, 2019–2026), so we run the temporal study there (§5.8). Pre-citation Street View

predicts next-year first citations against no-record controls, with significant lead out to 12 months, but a hard later-cited risk-set panel is near chance: the imagery sees the condition before the record exists, while citation timing among eventually-cited addresses appears to be set by the enforcement process rather than by visible condition.

4. Methods

The pipeline schematic summarizes the workflow: each parcel yields an aerial crop and a Street-View image; each viewpoint is scored by several feature extractors; scores are fused per signal; and the result is evaluated by ROC-AUC against the violation label with matched controls.

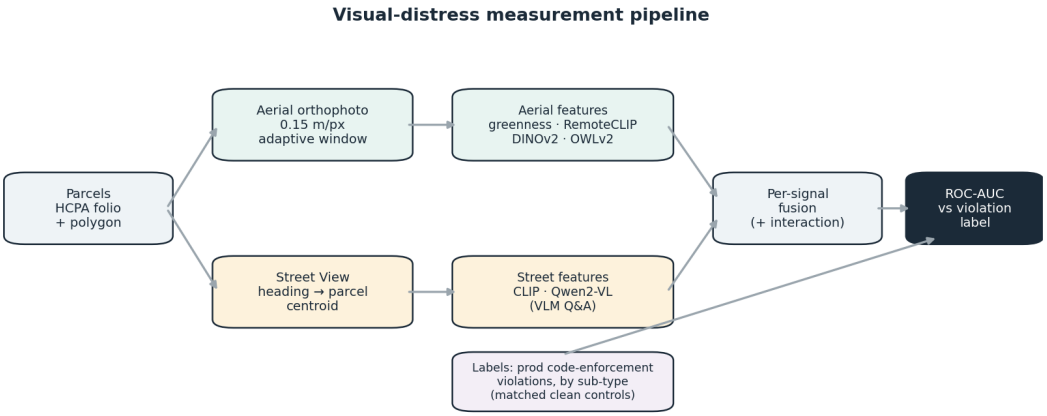

Figure 1: Pipeline overview: parcel records are joined to aerial and Street View imagery, scored with cheap features and zero-shot vision models, fused per signal, and evaluated against matched public-record labels.

We evaluate a ladder of feature extractors, scoring each with 5-fold stratified-CV logistic regression (ROC-AUC), and we additionally report single-feature AUC where a feature is one-dimensional.

- **Hand-crafted aerial features.** Excess-green (2G−R−B) mean/fraction/texture, vivid-blue fraction (pool/tarp), brown/bare fraction, brightness, edge density, dark-canopy fraction, computed over the parcel-centered crop.
- **Zero-shot CLIP** (ViT-L/14, LAION), prompt pairs per signal.
- **RemoteCLIP** (ViT-L/14, aerial-finetuned), same protocol.
- **DINOv2 linear probe**, frozen embeddings plus logistic regression.
- **OWLv2** open-vocabulary detection, per-class max score / count.
- **Street View CLIP** ("run-down/neglected" vs "well-kept").
- **Street View VLM** (Qwen2-VL-2B), P(Yes) on targeted questions ("poorly maintained?", "junk/debris?", "derelict vehicle?", "overgrown?").

Per-signal evaluation. We score each detector against its own violation sub-type (for example, greenness against overgrowth-cited parcels), because a pooled label mixes independent conditions and dilutes each.

Fusion. We combine aerial and Street-View detectors (i) by concatenating scores into the logistic model, and (ii) via an explicit overgrowth×debris interaction (rank product), using identical CV folds for fair comparison.

5. Results

5.1 Signal by sub-type

A pooled "any-distress" classifier barely separates the classes (about 0.55). Matched to its own sub-type, overgrowth is detectable from aerial imagery (excess-green AUC 0.645), while debris is invisible from overhead (about 0.50). The per-signal structure, not a single distress score, is the unit of analysis we use throughout.

5.2 Deeper models do not help on aerial imagery (Fig. 2)

On aerial imagery, deeper models do worse: CLIP ViT-L/14 (0.44), DINOv2 probe (0.45), RemoteCLIP (0.51), all at or below chance pooled, while a four-number greenness feature reaches 0.57. The global embeddings are distracted by scene content irrelevant to the localized distress cue. Naive feature-stacking also hurts: a 9-feature heuristic model (0.53) and a 28-feature everything model (0.45) both score below the single best feature.

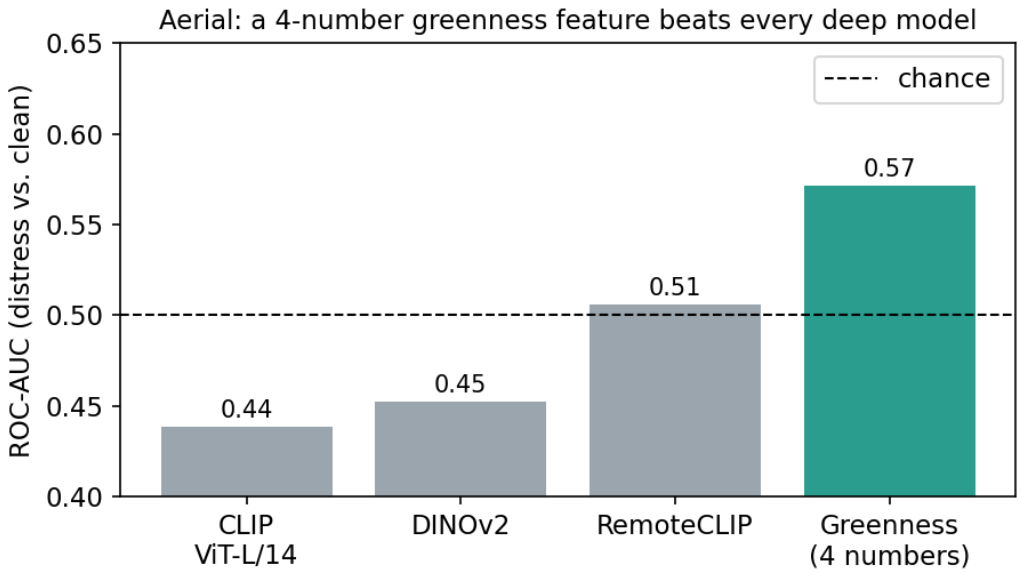

Figure 2: Aerial model comparison

5.3 Street View recovers debris; the VLM scores above CLIP (Fig. 3)

Ground-level imagery is in-domain for web-trained models and sees what overhead cannot. Street-View CLIP lifts debris from 0.50 to 0.574; a small VLM reaches 0.617 on debris (CI [0.55, 0.67]), 0.668 on overgrowth, and 0.596 on vehicles. The VLM trends above CLIP zero-shot (+0.03 to +0.04), but the difference is not significant at this sample (paired bootstrap p=0.08 to 0.15). The holistic question ("does this look neglected?") scores higher than the specific one ("is there debris?"), which echoes the aerial finding that targeted detection does not beat holistic judgment.

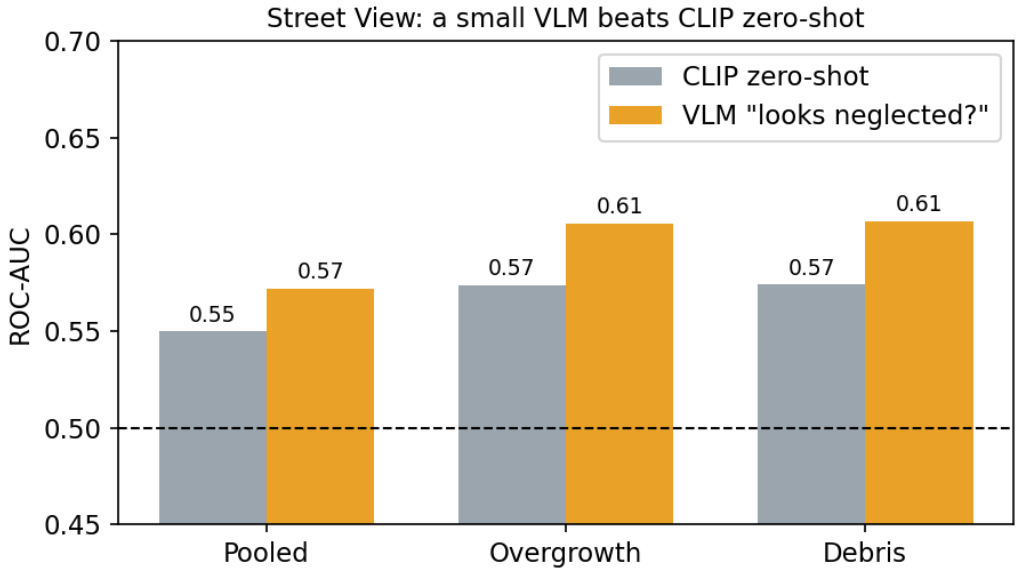

Figure 3: VLM vs CLIP on Street View

5.4 Source complementarity and fusion (Fig. 4, Fig. 5)

Aerial and Street View see different cues: overgrowth from above, debris and derelict vehicles from the curb (Fig. 4). Fusing them trends above either source alone on every signal: pooled 0.584 to 0.613, overgrowth 0.661 to 0.699, debris 0.581 to 0.636 (Fig. 5, error bars are 95% bootstrap CIs). The direction is consistent and the debris lift is borderline (+0.054, p=0.053), but none of the fusion gains reach significance at n=352, and the CIs overlap. We read this as encouraging evidence of complementarity that a larger, multi-jurisdiction sample is needed to confirm (§6). A blunt-CLIP fusion showed no such trend at all.

5.5 Conjunction of overgrowth and neglect (Fig. 6)

Parcels high on both overgrowth and neglect are violation-positive about 59 to 61% of the time against a 50% base rate, and an overgrowth×debris interaction (AUC 0.612) edges either feature alone (0.577 / 0.588). "Looks neglected" without overgrowth is the least predictive cell, so the co-occurrence appears to carry the information. As with fusion, this is a consistent pattern at a sample size too small to confirm formally, and we flag it as a hypothesis the larger sample should test.

5.6 Statistical significance

All AUCs carry 95% confidence intervals from 4,000 parcel-level bootstrap resamples; differences (fusion vs source, VLM vs CLIP) use a paired bootstrap on the same resamples. The single-viewpoint detectors are significant, with CIs that exclude chance: aerial greenness pooled 0.581 [0.521, 0.641], overgrowth 0.664 [0.576, 0.748], Street-View-VLM debris 0.612 [0.548, 0.673]. The comparative claims are not: every fusion lift (+0.02 to +0.05) and the VLM-over-CLIP gap (+0.03 to +0.04) has a CI that includes zero (p = 0.05 to 0.39). We therefore report complementarity and the VLM advantage as consistent positive trends, not confirmed effects, and we treat increasing

Figure 4: Side-by-side aerial and Street View examples

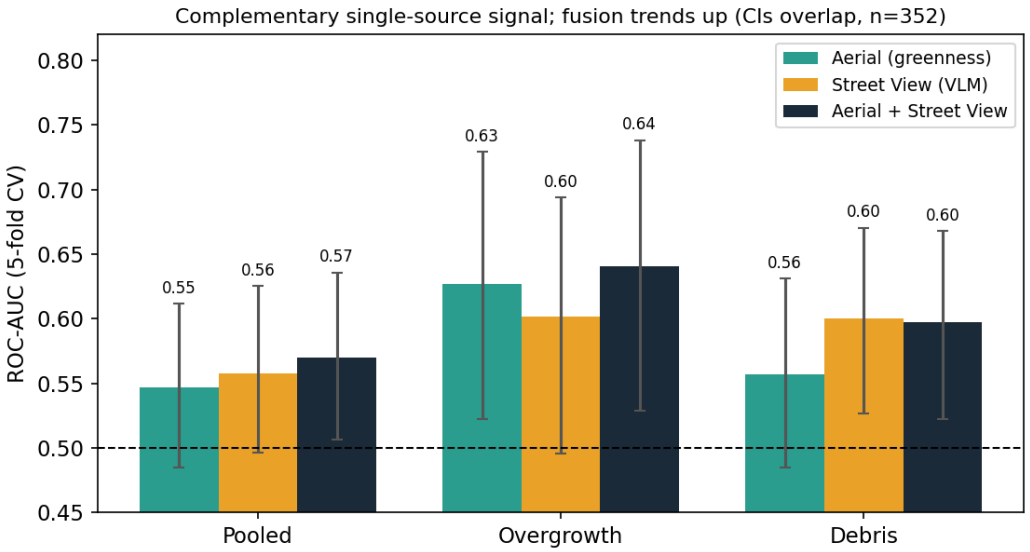

Figure 5: Source comparison and fusion

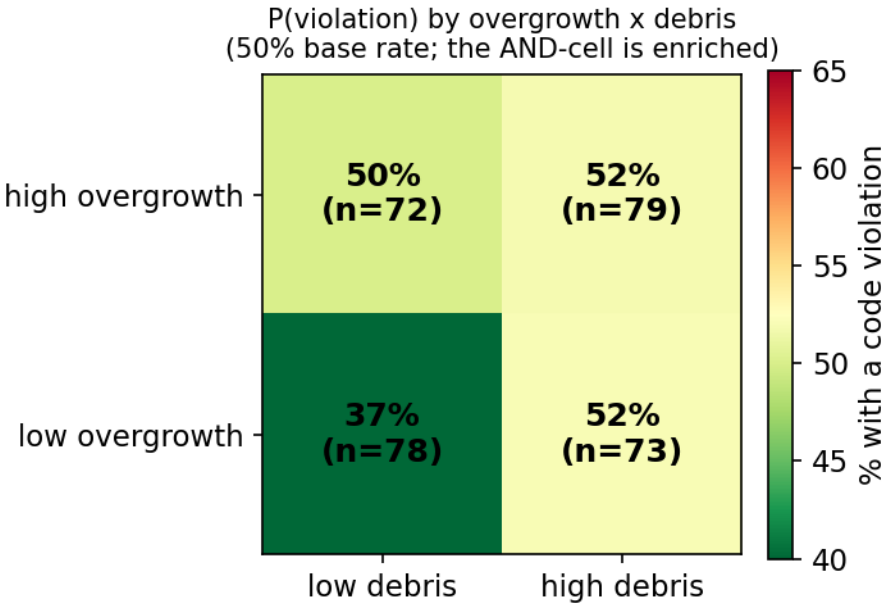

Figure 6: Overgrowth x debris conjunction

power via more jurisdictions (§6) as the way to settle them.

5.7 Cross-jurisdiction generalization (Fig. 7)

The value of an imagery proxy is greatest where public records are absent, and most of the world has no digital code-enforcement portal. The first question is therefore transfer: does a detector that saw no local labels work in a new jurisdiction? We replicate the entire protocol in Broward County (Deerfield Beach), a different metro with an independent code-enforcement system and its own property-appraiser parcel layer, with matched controls and the same zero-shot detectors (national NAIP aerial plus SV-VLM; nothing retrained).

All AUCs are directional (pre-specified sign: higher score means distress) with unfolded 95% bootstrap CIs. A folded or oriented AUC overstates near-chance features, because it cannot reveal a reversed feature, which turns out to matter here.

signal / source	Hillsborough	Broward
pooled / SV-VLM neglect	0.580 [0.52, 0.64]	0.588 [0.53, 0.65]
overgrowth / SV-VLM	0.629 [0.54, 0.72]	0.638 [0.54, 0.73]
vehicle / SV-VLM	n/a (n=10)	**0.694 [0.61, 0.78]**
debris / SV-VLM	0.617 [0.55, 0.68]	0.533 [0.50, 0.60]
overgrowth / aerial greenness (0.15–0.16 m)	0.640 [0.54, 0.73]	**0.428 [0.34, 0.52]** (reverses)
overgrowth / NAIP greenness (0.6 m)	0.547 [0.50, 0.61]	0.481 [0.39, 0.58]

The Street-View neglect detector transfers: pooled AUC is near-identical across the two metros (0.58 / 0.59) and overgrowth replicates (0.63 / 0.64), both with CIs above chance, so a detector calibrated on neither county generalizes to an unseen one. Vehicles are the strongest signal where sample size permits (Broward 0.694, n=37; Hillsborough's n=10 was uninformative). Debris did not transfer (0.62 to 0.53), plausibly because Broward's "bulk-trash" citations are more transient than persistent accumulation.

The aerial side is the clear negative. Hand-crafted aerial features do not transfer. The greenness feature that scored 0.640 in Hillsborough reverses in Broward (0.428): in lush Deerfield Beach the maintained lawns are the greenest, while cited "overgrown grass" lots are browner and patchier, so green-fraction tracks clean parcels. This is not a resolution artifact, because the 0.16 m imagery is crisp and the NAIP-coarse versions behave the same. Testing six aerial features (greenness, green-texture, edge density, gray-texture, brown-fraction, dark/canopy), every one flips sign or goes to chance across the two metros. A VLM applied to the aerial crops (overhead-phrased) only weakly and non-significantly judges overgrowth (about 0.53 to 0.59), though unlike the pixel features it stays directionally consistent and detects vehicles from above (Broward 0.66). Overhead imagery yields no transferable per-parcel condition signal. The transferable distress signal lives at ground level, in the Street-View VLM, which is the universally-available source useful where code records do not exist.

5.8 Temporal lead: the signal predates the record (Jacksonville)

Jacksonville MyJax exposes dated, geocoded visible-property requests, so a third metro lets us ask the temporal question directly: does the imagery signal exist before the public record does? We

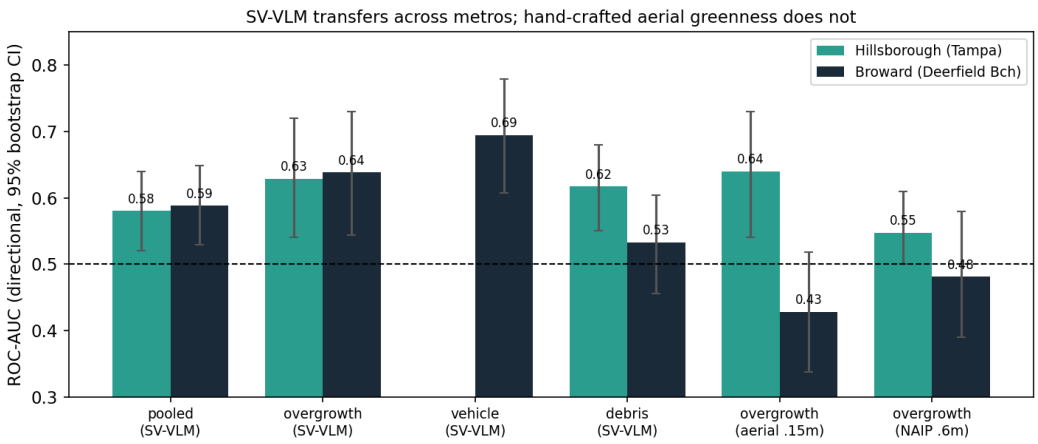

Figure 7: Cross-jurisdiction replication

built two 2025 panels whose control definitions test different claims.

The first is a hard risk-set panel: each positive address is observed before its first visible citation, and each control is an address observed on the same date whose first visible citation is more than one year away. We filtered to rows whose Street View pano date precedes the future event and scored 293 images (113 future-violation / 180 control) with Qwen2-VL-2B using the same Street View questions as §5.3. Every score is at chance on this contrast (neglect 0.524 [0.454, 0.592]; debris 0.504; vehicle 0.549; overgrown 0.525; three of four raw directions inverted). The image cannot predict *when*, within the eventually-cited population, the record arrives. This is a structurally hard test rather than evidence of no visual lead: controls here are not clean parcels but addresses cited later than one year out, so a persistent condition is typically present on both sides and the discriminating variable is largely complaint and inspection timing.

The second panel asks whether the signal precedes the record at all. Positives are the same kind of future visible-property citations; controls are random geocoded MyJax addresses with no visible-property record in the available history. Because address points alone aim Street View poorly, we resolved rows to Duval parcel polygons from the public parcel ArcGIS layer, queried Street View at the parcel centroid and boundary samples, deduplicated candidate panos, selected the pano closest to the parcel boundary, and aimed each camera at the nearest parcel-boundary point. After parcel resolution and date gating, 1,198 parcel-targeted pre-event images were scored.

score	n	future / no-visible-record	oriented AUC
VLM neglect	1,198	832 / 366	0.591 [0.556, 0.624]
VLM debris	1,198	832 / 366	0.581 [0.546, 0.616]
VLM vehicle	1,198	832 / 366	0.586 [0.552, 0.621]
VLM overgrown	1,198	832 / 366	**0.615 [0.581, 0.647]**

Unlike the risk-set panel, every score is significantly above chance. The imagery sees the condition before any record of it exists.

Stratifying the positives by the gap between the pano date and the citation date shows how far in advance the signal holds. Each bin is scored against the full control pool with the fixed pre-specified

direction (higher means distress) and unfolded CIs; debris and vehicle follow the same pattern as the two columns shown.

pano-to-citation gap	n pos	VLM neglect	VLM overgrown
0–3 months	122	0.646 [0.588, 0.704]	0.656 [0.601, 0.709]
3–6 months	156	0.641 [0.590, 0.690]	0.663 [0.611, 0.713]
6–12 months	171	0.583 [0.528, 0.634]	0.614 [0.562, 0.667]
12–24 months	218	0.511 [0.463, 0.559]	0.545 [0.495, 0.594]
24+ months	165	0.615 [0.564, 0.667]	0.630 [0.578, 0.681]

The signal is strongest when the pano is within 6 months of the citation (0.64–0.66), remains significant at 6–12 months (overgrowth 0.614), and decays to chance at 12–24 months, the profile expected if neglect emerges and intensifies in the months before someone complains. The separate significant 24+ month bin is not a camera-era artifact: positives with old panos skew toward old imagery, but restricting controls to equally old panos (2022 or earlier, n=136) leaves overgrowth essentially unchanged (0.629 [0.566, 0.692]). That tail is consistent with a chronic subset: properties visibly neglected for years before their first citation.

Together, the two panels and the lead-time profile show that visible condition precedes the first enforcement record by months, and for a chronic subset by years, but does not determine when, within the eventually-cited population, enforcement arrives. Because the no-record controls are a same-city random contrast rather than matched neighbors, part of this separation may be neighborhood-level rather than parcel-specific; the matched-control results of §5.1–5.7 remain the conservative parcel-level test.

5.9 A reader-capacity probe of the aerial signal

The weak aerial result could be a property of the imagery (the signal is not there) or of the extractor (small models cannot pull it). We probe this with a reader ladder on the masked, vacant-excluded aerial crops: open VLMs (Qwen2-VL-2B, Qwen2.5-VL-7B), a frontier VLM (GPT-5.5, scored over all 319 crops via a scripted single-image pass), and one human reader. Each reader rates 0 to 100 for distress, blind to the label, and we compute AUC directionally (higher means distress) against held-out labels with 95% percentile-bootstrap CIs.

The choice of extractor changes the result, but only up to a low ceiling. Open VLMs do not extract the signal, and scale within the tested open range does not help: 2B and 7B are both flat at about 0.53, alongside the hand-crafted and embedding features (about 0.5 to 0.65). The frontier model clears chance but only modestly. GPT-5.5 reaches pooled AUC 0.59 [0.53, 0.65], and the entire lift is overgrowth (0.61 [0.55, 0.67]); debris (0.52) and vehicle (0.52) stay at chance from above. A frontier reader recovers exactly the one signal aerial can physically see, at roughly the level of the four-number greenness statistic (Appendix A). A human reader sits a little higher (about 0.70). Larger open VLMs and a full human panel remain future work rather than evidence for the present claims.

Aerial imagery provides a coarse, overgrowth-weighted view. As a standalone per-parcel condition signal it is weak and well below Street View's (§5.3–5.7), which a small VLM extracts and which transfers across jurisdictions. For deployable, standalone per-parcel distress, the useful viewpoint is ground level. Aerial still has two roles. It is a usable prefilter: it covers the entire map at near-zero

marginal cost, so even a modest overgrowth AUC lets a system narrow millions of parcels to a high-recall candidate set before spending the more expensive, less-complete Street-View pass. It is also a conjunctive data point: aerial overgrowth and ground-level debris are independent viewpoints on the same parcel, and their agreement is more informative than either alone (the per-signal fusion of §5.4 and the conjunction analysis of §5.5). Aerial is useful for map-wide coverage and for corroborating the ground-level signal, not as a standalone discriminator. The full ladder and methodological detail are in Appendix B (Fig. 8).

6. Discussion

Three themes recur. Holistic judgment outperforms targeted detection: both CLIP-vs-OWLv2 (aerial) and VLM-neglect-vs-VLM-debris (street) show that a global "is this neglected?" judgment scores above detecting specific objects. Small distress objects are hard to localize, but their gestalt is legible. Cheap features outperform expensive ones: a four-number greenness statistic out-predicts 768-dimensional foundation-model embeddings on the one signal aerial can see. The gains in this study came from matching the question to the source and combining viewpoints; none came from model scale. And the sources are complementary, so they should be fused per signal, with an interaction term.

The achievable AUC plateaus at about 0.6 to 0.7 across every model and source. We argue this ceiling is set by the label, not the imagery: code-enforcement is complaint-driven (so non-cited neighbors are often merely unreported rather than clean), type-mixed, and temporally unaligned with the imagery. Against a random background baseline, the deployment setting where location is legitimately informative, the same aerial vegetation score rises from matched-control AUC 0.56 [0.50, 0.63] to 0.63 [0.57, 0.69]. The matched-control number is therefore a conservative floor on the parcel-level visual signal. A simple contamination model gives the same conclusion: if observed positives and observed controls are mixtures of true distressed and clean parcels, then observed AUC is attenuated as 0.5 + (1-p-n)(true AUC - 0.5), where p+n is the combined contamination rate. The overgrowth AUC of 0.645 is consistent with a latent AUC near 0.8 when combined contamination is about 0.5, a plausible level for complaint-driven positives and same-neighborhood "clean" controls. Near-chance pooled or temporal tasks require much more extreme contamination, so label noise explains the ceiling where a signal exists but does not rescue every null result.

The temporal results (§5.8) should be read in that light. A future code case is not a direct physical-condition label but an enforcement-mediated outcome that depends on visibility from the public right-of-way, complaint behavior, inspector routing, local priorities, staffing, and the timing gap between the Street View pano and the later record. The lead-time stratification separates the visual part from the institutional part: the condition appears in the imagery months before the record (0.64–0.66 within 6 months of the citation, significant out to 12, with a chronic 24+ month tail), while citation timing among eventually-cited addresses is unpredictable from imagery, as expected when complaint behavior determines when enforcement happens. The remaining error is therefore not identified as a single failure mode: it mixes model sensitivity, the selectivity of complaint-driven enforcement, and conditions that appeared or were remediated between image capture and citation.

Two additional random-control expansion checks point in the same direction. In Lakeland, Florida, we resolved 885 visible code-enforcement positives to parcel polygons, sampled 1,200 random residential parcel controls from the same public parcel service, and used the parcel-boundary Street View targeting procedure. Street View coverage was 1,971/2,085 rows (94.5%). On this same-city random-control panel, Qwen2-VL-2B scores separate positives from random residential controls

much more strongly: VLM-neglect AUC 0.724 [0.701, 0.745], with debris, vehicle, and overgrowth scores also significant (0.651–0.678). A smaller Palm Beach County active-case panel, built from public parcel polygons and the same targeting method, gives weaker but still above-chance separation on 609 Street View-covered parcels: neglect AUC 0.608 [0.560, 0.654], vehicle AUC 0.608 [0.561, 0.652], debris AUC 0.580 [0.533, 0.626], and overgrowth AUC 0.592 [0.545, 0.636]. These are not matched-neighbor results, so they are best read as deployment-style evidence: the broad Street View distress signal is real, while same-block matched controls remain the conservative test of parcel-specific signal beyond neighborhood context.

7. Limitations

- **Conservative baseline.** Matched same-neighborhood controls make the task harder than the deployment setting and likely contain unreported distress. A random/background check with cached Hillsborough parcels plus Lakeland and Palm Beach random residential-control panels confirm higher deployment-style AUCs, but those panels answer a broader market-ranking question rather than the parcel-vs-neighbor question. Palm Beach is also active-only and has fewer than 300 Street View-covered positives, so it is an external-validity check rather than a main matched-transfer jurisdiction.
- **Limited temporal evidence.** Tampa/Hillsborough violations are undated, so the main matched-control study is cross-sectional. Jacksonville supplies dated MyJax records, and a random no-visible-record control panel is above chance with significant lead out to 12 months (§5.8), but the harder later-cited risk-set panel is near chance, and the no-record contrast is unmatched, so part of the temporal separation may be neighborhood-level. The weakness of the temporal AUCs should not be interpreted only as weak visual signal: future violations are complaint- and enforcement-mediated outcomes, so reporting behavior, local priorities, remediation timing, and Street View date mismatch can all attenuate measured predictability. The current temporal panels prevent obvious leakage by retaining only panos dated before the future event, but that does not create a complete historical backtest because the Street View API generally returns the currently available pano rather than an arbitrary prior pano. A cleaner prospective design is to freeze and cache a parcel sample's imagery today, register the scores, and evaluate against newly observed code records after a fixed horizon such as one year.
- **Still modest n.** About 185 distress parcels per matched-control metro, and some sub-type counts remain small (Hillsborough vehicles n=10). The added random-control panels broaden geography but do not replace matched-control evidence and are not a national sample.
- **Imagery availability shapes coverage.** The strong aerial overgrowth result needs about 0.15 m county imagery, which is not universal; on national NAIP (0.6 m) it washes out. Conclusions about which source sees what are resolution-dependent.
- **Street View targeting is still approximate.** The Jacksonville, Lakeland, and Palm Beach random-control panels use parcel geometry to sample multiple candidate panos and aim at the parcel boundary, but the matched Hillsborough/Broward Street View set used the older centroid heading. The boundary procedure makes the image more likely to face the property, but it does not prove frontage visibility: corner lots, deep or irregular parcels, multi-unit parcels, occlusions, and stale panos can still be misframed. A manual or model-based frontage-quality audit is the next improvement.
- **Label proxy.** A code violation is an imperfect, complaint-driven proxy for visible distress.

8. Conclusion

The central result is transfer: a label-free Street-View detector generalizes to a second metro with no retraining, so it can be used in the many jurisdictions that keep no digital code-enforcement records. In a third metro with dated labels, the same detector sees the condition before the record exists, with significant lead out to 12 months before the first citation, although when the citation actually arrives depends on the complaint process. The mechanism is modest but consistent: public aerial and street-level imagery carry complementary parcel-level distress signals, with overgrowth visible from above and debris or derelict vehicles visible from the curb. Cheap interpretable features and a small VLM capture most of what is currently measurable; larger models do not raise these numbers. Further progress requires broader geographic coverage and better labels (temporal, type-specific), and we release the datasets and pipeline to support that work.

Data and Code Availability

An anonymized benchmark bundle is archived on Zenodo (DOI: 10.5281/zenodo.20672748). The bundle contains six tables: Hillsborough matched controls, Broward matched controls, the hard Jacksonville 2025 later-cited temporal panel, the Jacksonville 2025 random no-visible-record temporal panel, and Lakeland and Palm Beach cross-sectional random-control panels. Each row has salted example IDs, labels, subtype or service metadata, imagery-source provenance, Street View targeting metadata, and model scores; it excludes addresses, coordinates, parcel identifiers, STRAP/folio IDs, geometries, and all raw image bytes. All figures regenerate via `scripts/make_figures.py`; `RESULTS.md` is the full experiment log and `RUNBOOK.md` documents the pipeline. We do not redistribute Google Street View imagery (per its terms), but the released score tables are sufficient to reproduce the analysis and the code includes the fetch scripts for users with appropriate imagery access.

Funding

This research was conducted at and funded by DirtSignal.

Competing Interests

The production code-enforcement ingestion pipeline used to derive the ground-truth violation labels in this study is operated by DirtSignal. The author has a financial interest in DirtSignal.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used Claude (Anthropic) for drafting and editing the manuscript text and analysis code. The author reviewed and edited the output as needed and takes full responsibility for the content of the published article. Separately, vision-language models (including GPT-5.5) were used as study instruments under evaluation; that use is part of the methodology reported in Sections 4–5 and Appendix B, not manuscript preparation.

References

Tianheng Cheng, Lin Song, Yixiao Ge, et al. YOLO-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, et al. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and Cesar A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, 2016.

Chuanbo Hu, Shan Jia, Fan Zhang, Changjiang Xiao, Mindi Ruan, Jacob Thrasher, and Xin Li. UPDExplainer: An interpretable transformer-based framework for urban physical disorder detection using street view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:209–222, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Fan Liu, Delong Chen, Ziyuan Guan, Xiaocong Zhou, Jiale Zhu, Qian Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024b.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023.

Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and Cesar A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. SAM 2: Segment anything in images and videos, 2024.

Andrew G. Rundle, Michael D. M. Bader, Catherine A. Richards, Kathryn M. Neckerman, and Julien O. Teitler. Using google street view to audit neighborhood environments. *American Journal of Preventive Medicine*, 40(1):94–100, 2011.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024a.

Zhecheng Wang, Rohit Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. SkyScript: A large and semantically diverse vision-language dataset for remote sensing. In *AAAI Conference on Artificial Intelligence*, 2024b.

Fan Zhang, Arianna Salazar-Miranda, Fabio Duarte, Lawrence Vale, Gary Hack, Min Chen, Yu Liu, Michael Batty, and Carlo Ratti. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. *Annals of the American Association of Geographers*, 114(5): 876–897, 2024a.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Sheng Zou and Le Wang. Detecting individual abandoned houses from google street view: A hierarchical deep learning approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:298–310, 2021.

Appendix A: methods table

viewpoint	extractor	best signal	AUC
Aerial	excess-green (4 numbers)	overgrowth	0.645
Aerial	RemoteCLIP / DINOv2 / CLIP	pooled	0.44–0.51
Aerial	OWLv2 detection	(all)	~0.50
Street View	CLIP zero-shot	debris	0.574
Street View	Qwen2-VL-2B “neglect”	debris	0.617
Street View	Qwen2-VL-2B “overgrown”	overgrowth	0.668
Fusion	aerial greenness + SV VLM	overgrowth	0.699
Fusion	aerial greenness + SV VLM	debris	0.636

Appendix B: aerial reader-capacity ladder (§5.9)

This appendix gives the detail behind the §5.9 probe of whether the weak aerial signal is a property of the imagery or of the extractor.

Setup. All readers scored the same masked, vacant-excluded crops: the subject parcel is marked (polygon outline plus dimmed surroundings) and zero-building-value parcels are removed, so a reader is judging one identified property rather than a patch of neighborhood. Each rating is blind to the label. “Pooled” AUC is the holistic neglect score against the binary distress/clean label, directional (higher means distress), with a 95% percentile-bootstrap CI. The GPT-5.5 row is a scripted single-image pass over all 319 crops (one independent call per crop, forced-JSON output), directly comparable to the open-VLM rows.

reader	pooled AUC	n
Qwen2-VL-2B	0.53	319
Qwen2.5-VL-7B	0.53 [0.47, 0.60]	319
GPT-5.5	0.59 [0.53, 0.65]	319

reader	pooled AUC	n
Human reader	~0.70 (acc)	23

GPT-5.5 per-signal (directional AUC, 95% CI, n=319): overgrown 0.61 [0.55, 0.67], neglect 0.59 [0.53, 0.65], debris 0.52 [0.45, 0.58], vehicle 0.52 [0.45, 0.57].

- Open VLMs do not extract the signal, and scale within the open range does not help: 2B and 7B are both flat at about 0.53, alongside greenness/CLIP/DINOv2/RemoteCLIP (about 0.5 to 0.65). Any capacity threshold is above 7B.
- A frontier VLM clears the threshold, but only modestly and only on overgrowth. GPT-5.5's pooled 0.59 is significantly above chance (CI excludes 0.5), and the lift is entirely overgrowth (0.61); debris and vehicle stay at chance. It lands at roughly the four-number greenness statistic (overgrowth 0.645, Appendix A), so a frontier model recovers the one thing aerial can see and no more. Subject- marking is what makes even this possible (greenness over the parcel polygon is 0.655 vs 0.640 for an unmarked center crop).
- Open models above 7B and a larger human panel remain future work. The present ladder is enough to show that small open VLMs are flat, while the frontier reader clears chance only modestly and only on overgrowth.

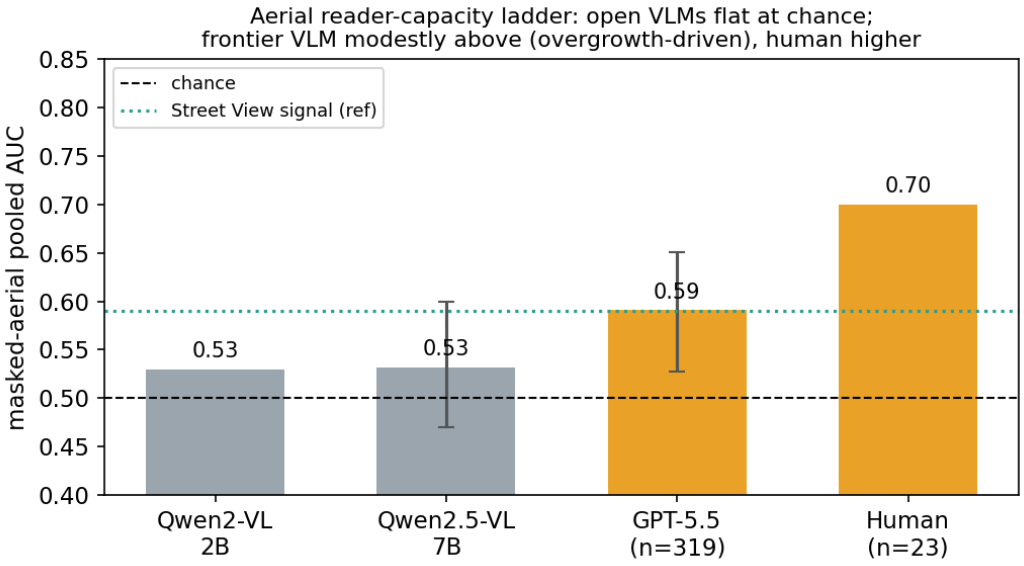

Figure 8: Aerial reader-capacity ladder