Correlated Verifier Errors in Counter-UAS Perception: Clustering, Cross-Sensor Dependence, and Label-Free Auditing

Shreyo Ghosh
Department of Biomedical Engineering, Rutgers University–New Brunswick, Piscataway, NJ, USA
Correspondence: sg2597@scarletmail.rutgers.edu ORCID: 0009-0007-6467-9939

Preprint, not peer reviewed. Posted under a CC BY 4.0 licence. Version 1, November 2026.

Abstract

Counter-UAS perception systems re-query classifiers on the same track and fuse visible and infrared evidence as if each query and each sensor were independent. We test this assumption with random-effects and latent class models from diagnostic-test evaluation, on two public drone recording sources (Halmstad and Anti-UAV RGB-T) plus AVID, a paired visible–infrared subset of the Halmstad recordings that adds bird negatives, five model families and three detector architectures. Errors cluster strongly within tracks (track-level random-effect standard deviations from 0.7 to above 5 on the logit scale), and false alarms cluster as strongly as misses. With out-of-session predictions for all 315 Anti-UAV RGB-T sequences in 37 recording sessions, recording conditions account for about two thirds of infrared clustering and half of visible clustering. For trained detectors, infrared and visible misses co-occur on the same frames far more often than independence predicts (odds ratio 6.26 across 315 sequences), so a second sensor is not independent redundancy. As a consequence, a standard label-free (Dawid–Skene) audit overstated detector sensitivity by 8 to 18 points on real votes and misestimated accuracy by up to 57 points when bird negatives were available. We show that dividing summed frame evidence by Kish's design effect is the exact Bayes rule under Gaussian, calibrated, equally correlated evidence; on real data it cut track-level overconfidence from 29 to 10 points without changing accuracy, and error dependence decays with time in sensor-specific ways. Counter-UAS systems should be fused, audited and validated as if their errors were correlated, and evaluated on recording sessions never seen in training.

Keywords: counter-UAS; sensor fusion; conditional dependence; random effects; label-free evaluation

Impact Statement. Drone-detection systems often look at the same target many times, and with more than one camera, then combine the answers as if each look were an independent check. Using real visible and thermal infrared recordings, we show this is not true: a detector that misses a drone tends to keep missing it, and the two cameras tend to fail together, largely because of shared recording conditions. Treating these looks as independent makes a system overconfident, and makes standard label-free checks of sensor accuracy report accuracy that is too high. Simple statistical corrections restore honest confidence. Counter-drone perception should be tested under recording conditions not seen in training, without counting redundant observations as independent evidence.

1 Introduction

A counter-UAS system must decide, often within seconds, whether a track is a drone or something benign such as a bird. A common design pairs a cheap always-on detector with a more expensive verifier, re-queries the verifier as a track persists, and fuses evidence from visible and thermal infrared sensors. Each of these

steps is usually justified by an implicit independence assumption: that a second query on the same track, or a second sensor viewing the same target, adds roughly as much evidence as an independent observation.

Biostatistics has long studied the failure of this assumption in a different setting. When two diagnostic tests are applied to the same patient, their errors are often correlated given true disease status, and ignoring that correlation biases estimates of each test's accuracy [1, 2]. Random-effects latent class models were developed precisely to handle it [3]. This paper carries those tools into counter-UAS perception and asks three questions:

1. Do verifier errors cluster within tracks, and how strongly?
2. Is that clustering a property of hard targets, hard scenes, or both?
3. Are visible and infrared errors conditionally independent, as label-free audits of sensor accuracy require?

The answers are: yes and strongly; both, with scenes contributing at least as much as targets; and no, not for a trained detector. We then quantify what the third answer costs a label-free audit.

1.1 Related work

Latent class models for raters and tests. Dawid and Skene [4] estimated observer error rates by EM without ground truth, assuming conditional independence of raters given the true class; Hui and Walter [5] gave identifiability conditions for two tests across populations. Vacek [1] and Torrance-Rynard and Walter [2] showed that conditional dependence between tests biases such estimates, typically overstating accuracy. Qu, Tan and Kutner [3] introduced random-effects latent class models that absorb dependence through shared subject-level effects. In machine learning, the same structure appears in crowd-label aggregation [6].

Dependent classifiers and multimodal fusion. In machine learning, Jaffe et al. estimated classifier accuracies without labels [7] and then developed methods to detect and model strongly dependent classifiers [8]; our audit results are an instance of the failure they warn about, measured on real counter-UAS perception data. In record linkage, Xu et al. [9] showed that ignoring conditional dependence matters most when prevalence is near 0 or 100% and the dependence lies in the dominant class, which is the regime of our real-data audit (96% drone frames, with dependence among them). Bayesian late fusion of RGB and thermal detectors often multiplies per-modality likelihoods under conditional independence; recent work relaxes this by conditioning on sensing context [10], and uncertainty-aware evidential fusion has been applied to Anti-UAV RGB-T [11]. A recent Bayesian sensor-fusion study lists the independence assumption as a limitation that real data must test [12]. To our knowledge, no prior work measures track-level error clustering or cross-sensor conditional dependence in counter-UAS detectors, or shows a label-free audit failing on real drone-detector votes.

Ensembles, cascades and deferral. Detection cascades [13] and learning-to-defer methods [14] route hard inputs to stronger models, and ensemble-diversity measures [15] quantify correlated errors among classifiers. These works generally treat inputs as exchangeable; we focus on the repeated-measures structure of tracks and sensors.

Drone detection data. We use the Halmstad multi-sensor drone dataset [16] and the Anti-UAV RGB-T benchmark [17], which provides simultaneous visible and infrared video.

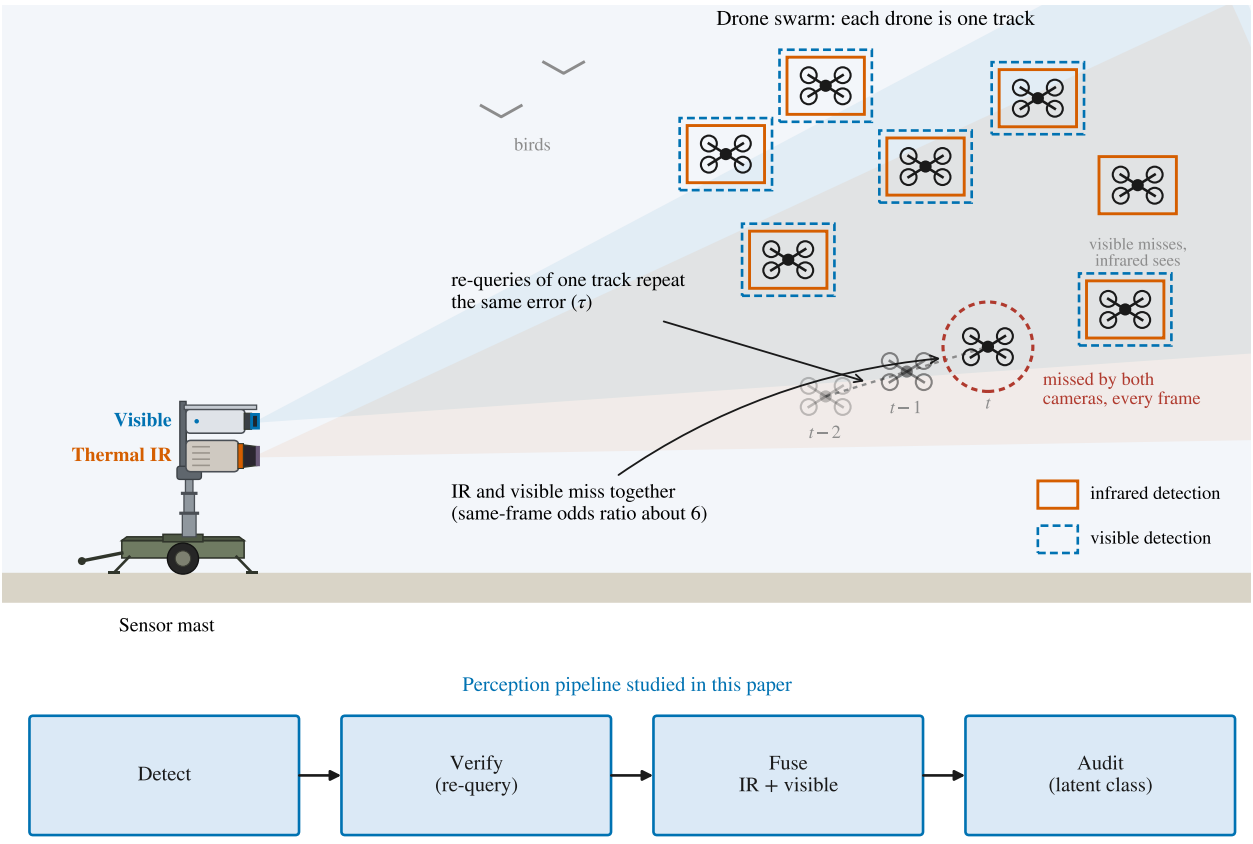


Figure 1: Schematic of the perception problem studied here. Visible and thermal infrared cameras on one sensor mast watch a drone swarm among birds; each drone is a track that is queried repeatedly. Most drones are detected by both cameras (boxes); some are missed by one camera only; and some, like the circled drone, are missed by both cameras frame after frame ($t-2, t-1, t$). These are the two dependencies the paper measures: errors repeat across re-queries of one track (τ), and infrared and visible errors co-occur on the same frame (odds ratio). The pipeline below shows the stages studied.

2 Materials and Methods

2.1 Data

Halmstad. The public Halmstad dataset [16] contains visible and thermal infrared clips, each with one target class (drone or bird), per-frame boxes, and a distance label. We used 158 visible clips (111 drone, 47 bird; 902 crops) and 224 infrared clips (152 drone, 72 bird; 1,315 crops). Crops were sampled at least 2 s apart, cropped at three times the box size (minimum 48 px), and resized to 224 px; clips with fewer than three crops were dropped. A subset of the visible and infrared clips was recorded simultaneously and has since been paired frame by frame in the AVID benchmark [18]; in the Halmstad analyses we treat each sensor's clips separately, and we use the paired AVID subset for the cross-sensor analyses in Section 3.10.

Anti-UAV RGB-T. Anti-UAV RGB-T [17] (public copy on Hugging Face) contains 318 sequences with simultaneous visible and infrared video and per-frame annotations. The recording session (date and time) is encoded in each sequence name, which gives real scene labels. We used it for three analyses: zero-shot CLIP on 60 test sequences (1,046 paired frames); a trained YOLO11n detector (Section 2.10); and a far-range

subset of 18 sequences from 6 sessions with at least two sampled frames whose infrared box diagonal is at most 30 px. Within those sequences, 245 frames per sensor meet the 30 px criterion; we call them *far frames*.

AVID. AVID [18] is derived from the Halmstad recordings, so our three benchmark resources come from two independent recording sources. It pairs visible and infrared frames from 115 single-class Halmstad videos (bird or drone), mostly with targets of 12 to 16 px, and defines a train/validation split by video; it is released under CC BY-NC 4.0. We used its 13 validation videos (9 drone, 4 bird; 3,250 frame pairs) as a source of real negatives in both sensors, with truth taken from each video's class.

2.2 Within-track clustering

For each model, sensor and class we fit a random-intercept logistic model

$$\operatorname{logit} P(\text{error}_{ij}) = \beta_0 + \beta_1 \log(\text{size}_{ij}) + \gamma_{\text{dist}(i)} + \tau u_i, \qquad u_i \sim N(0,1), \tag{1}$$

where i indexes clips or sequences and j crops or frames. The standard deviation τ measures how much error propensity varies between tracks beyond what target size and distance explain; the latent intraclass correlation is $\tau^2/(\tau^2+\pi^2/3)$. We estimate by exact maximum likelihood with 30-point Gauss–Hermite quadrature, report 95% profile-likelihood intervals, and test $\tau = 0$ with the boundary-corrected likelihood ratio test [19], applying Holm's adjustment [20] across tests. On synthetic data, the estimator returned 0.00 (CI 0 to 0.53) when $\tau = 0$ and 1.75 (CI 1.40 to 2.18) when $\tau = 1.5$. A label-permutation null (shuffling clip labels within size-by-distance cells) gave the same conclusions.

2.3 Target versus scene

To separate hard targets from hard scenes we fit nested random intercepts,

$$\operatorname{logit} P(\text{error}_{ijk}) = X_{ijk}\beta + \sigma v_k + \tau u_{ik}, \qquad v_k, u_{ik} \sim N(0,1), \tag{2}$$

with scenes k, tracks i and frames j, integrated by 15×15 Gauss–Hermite quadrature. We report the sequence share of clustering variance, $\tau^2/(\tau^2+\sigma^2)$. On Halmstad, which lacks session metadata, scenes were reconstructed from each clip's first frame with the target masked (grey thumbnail plus brightness histogram), grouped by Ward clustering into $K = 10$, 20 or 40 groups. On Anti-UAV RGB-T we used the real recording sessions.

2.4 Cross-sensor dependence

For paired visible and infrared frames we report (i) the Mantel–Haenszel odds ratio [21] pooled over per-sequence 2×2 tables of infrared and visible errors on the same frame, which measures same-frame dependence within sequences; and (ii) the correlation of per-sequence error rates, which measures whether the same sequences are hard for both sensors, with sequence-level and session-level bootstrap intervals. We do not apply a continuity correction: adding 0.5 to every cell of small per-sequence tables biased the pooled odds ratio toward 1 by about 25% in simulation.

2.5 Label-free audit

The audit fuses several views of each track, each reduced to one vote, with a two-class latent class model. The independence model is Dawid–Skene [4]. The dependence-aware variants are a random-effects latent

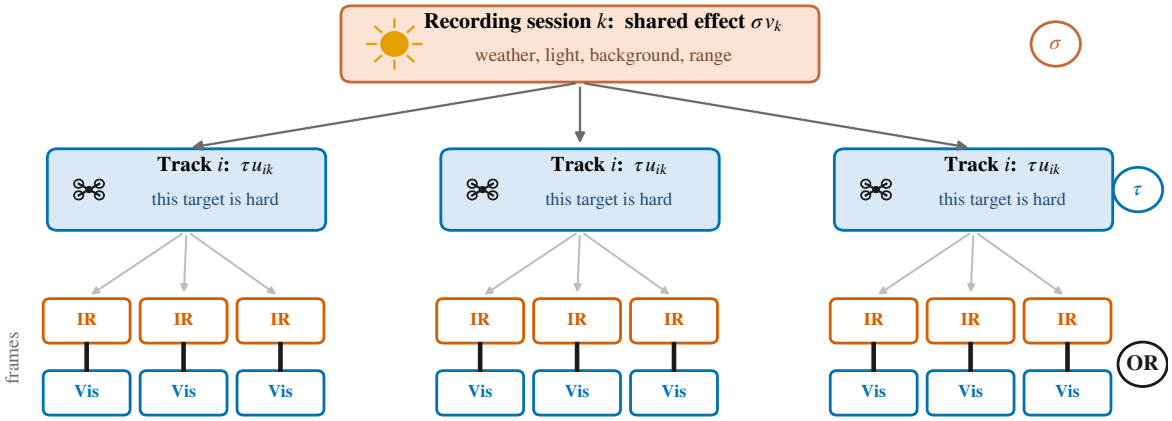


Figure 2: The three levels of error dependence measured in this paper. Session effects σv_k are shared by every track in a recording session; track effects τu_{ik} are shared by every re-query of one target; and infrared and visible errors on the same frame can co-occur beyond what the shared effects explain.

class model with a shared hard-track loading [3] (used on Halmstad) and, for the infrared–visible pair, a model that treats the pair as a single four-level categorical view within each class, leaving the remaining views conditionally independent (used in the calibrated simulation of Section A.2).

2.6 Combining correlated evidence within a track

Let $\ell_1, \ldots, \ell_n$ be the frame-level log-likelihood ratios of a track, and suppose that given the class $c \in \{0,1\}$ they are Gaussian with mean μ_c, common variance s^2 and within-track correlation matrix R. The exact track-level log-likelihood ratio is then

$$\Lambda = \frac{\mu_1 - \mu_0}{s^2}\, \mathbf{1}^\top R^{-1}\big(\ell - \bar{\mu}\,\mathbf{1}\big), \qquad \bar{\mu} = \tfrac{1}{2}(\mu_0 + \mu_1). \tag{3}$$

Under exchangeable correlation $R_{jk} = \rho$ $(j \neq k)$, $\mathbf{1}^\top R^{-1} = \mathbf{1}^\top/(1+(n-1)\rho)$, so $\Lambda = \frac{\mu_1-\mu_0}{s^2}\,(\sum_j \ell_j - n\bar{\mu})/(1+(n-1)\rho)$. If the frame log-likelihood ratios are calibrated, a Gaussian log-likelihood ratio satisfies $\mu_1 = -\mu_0 = s^2/2$, and this reduces exactly to

$$\Lambda = \frac{\sum_j \ell_j}{1+(n-1)\rho}, \tag{4}$$

the sum of frame evidence divided by Kish's design effect [22]. The design-effect correction is therefore the Bayes-optimal combination rule under Gaussian, calibrated, exchangeably correlated frame evidence; treating frames as independent ($\rho = 0$) overstates the evidence by the factor $1+(n-1)\rho$. Because exchangeability can fail, we also use Eq. (3) with the structure standard in longitudinal analysis [23], a random intercept plus serial correlation, $R_{jk} = a + b\,\phi^{|j-k|}$ for $j \neq k$, with a, b, ϕ, μ_c and s^2 estimated on training videos.

2.7 Query-policy replay

To test correlation-aware querying on real data, we replayed sequential policies on the Halmstad per-crop votes of zero-shot CLIP, Qwen yes/no and the CLIP probe, querying each clip's crops in frame order to

classify the clip as drone or bird. An independence policy updates a posterior assuming i.i.d. votes given the class; a correlation-aware policy instead models each class's vote rate as beta-binomial across clips, the simplest form of track-level clustering. Both stop when the posterior passes a threshold. Parameters were fit on four folds and evaluated on the fifth (5-fold, grouped by clip), and we report accuracy, mean queries, and overconfidence (mean stated confidence minus accuracy).

2.8 Session cross-fitting

Session-level variance is poorly estimated from few sessions, and a detector evaluated on sessions it was trained on overstates performance. To obtain an out-of-session prediction for every Anti-UAV RGB-T sequence, we split its 37 recording sessions at random (seed 7) into two folds of 19 and 18 sessions. For each fold we trained a YOLO11n (15 epochs, image size 640, every second sampled frame of the fold's sequences, with 10% of its sequences held out for validation) and used it to score every frame of the other fold, sampled every 50 frames. Sequences with fewer than three usable frames were excluded (3 of 318), leaving 315 sequences and 11,246 frame–sensor pairs, with 7 to 10 sequences per session. Because each session is scored by exactly one of the two detectors, a difference between the detectors could be confounded with session effects; we therefore also fit the nested model with the fold as a fixed effect and test the fold term by a likelihood-ratio test. Uncertainty in the cross-sensor odds ratio is assessed by resampling whole sessions (2,000 session-cluster bootstrap resamples), so that dependent frames within a session are never treated as independent.

2.9 Use of AI tools

Claude (Anthropic) was used throughout this project to write and debug analysis code, run experiments in a cloud notebook, prepare figures and draft text. Every reported number was produced by code that is publicly available (Data Availability Statement) and can be rerun. The derivation in Section 2.6 was checked numerically against exact multivariate Gaussian log-likelihood ratios (maximum absolute discrepancy 4×10^{-13} over 900 random cases), and the estimators were checked on synthetic data with known parameters before use on real data. The author reviewed all analyses and takes full responsibility for them.

2.10 Trained detector

We fine-tuned YOLO11n [24] for 25 epochs on frames from 35 Anti-UAV RGB-T training sequences (both sensors pooled), with 5 further training-split sequences for validation. We evaluated on two held-out test sets: set A, 40 test-split sequences from 5 sessions; and set B, 30 sequences drawn 3 per session from 10 sessions spread across the test split. Training sequences were chosen so that no training session appears in either test set or in the far-range set. A frame counts as a miss when no detection with confidence at least 0.25 lies within one box size of the annotated center. Frames were sampled every 50 frames (25 for far range). To check robustness we retrained with two further seeds and re-evaluated set B, and re-evaluated set B's visible frames at 1280 px input to test whether downscaling the 1920 px visible frames drove the visible results. For a real-data audit, we also ran the detector on set-B frames with no drone in either sensor and, as a temporal view, on the same sensor 12 frames (about half a second) later.

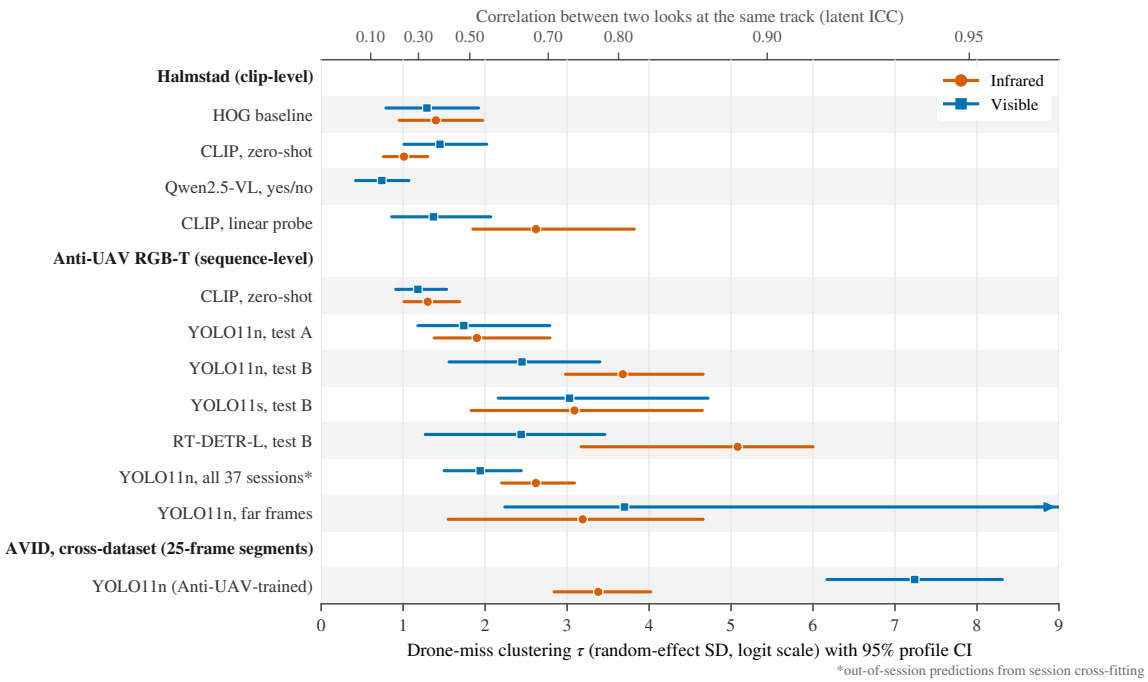


Figure 3: Drone-miss clustering across five model families, three detector architectures and three benchmarks. Points are maximum-likelihood estimates of τ (infrared circles, visible squares) with 95% profile-likelihood intervals; the top axis translates τ into the latent correlation between two looks at the same track. The visible far-frame interval extends to 20.9 (arrow). Test sets A and B are held-out Anti-UAV RGB-T evaluations; the 37-session row uses out-of-session predictions from session cross-fitting; AVID uses 25-frame segments as tracks.

3 Results

3.1 Errors cluster within tracks

Figure 3 and Table 8 summarize drone τ across all models. Every drone test with enough errors to measure is significant after Holm adjustment. Bird tests are mostly underpowered (47 to 72 clips, often under 5% error). At $\tau \approx 1.3$ (latent ICC ≈ 0.34), three repeated queries on one clip carry the information of about 1.8 independent ones. Errors also fall with target size: zero-shot CLIP missed 24% of 16 to 32 px visible crops and 7% of 32 to 64 px crops.

3.2 Zero-shot VLMs are weak and fail together

Zero-shot CLIP caught 76% of visible and 46% of infrared drones; Qwen2.5-VL-3B [25] caught 13% and 0.4% with a three-way prompt and 40% and 3% with a yes/no prompt, while rarely calling birds drones (92% to 100% correct). Both models are conservative in the costly direction for counter-UAS, and changing only the prompt tripled Qwen's visible recall. CLIP and Qwen per-clip error rates correlate at 0.61 (visible) and 0.63 (infrared), so a second zero-shot model adds little independent evidence.

3.3 Correlation-aware querying on real votes

Table 1 and Figure 8 (Appendix) show the replay. The main claim of our simulation (Appendix A.1) did not transfer: on real votes the correlation-aware policy used more queries, not fewer, because the independence

Table 1: Sequential query policies replayed on real Halmstad votes (stopping threshold 0.95, held-out folds). Over = mean stated confidence minus accuracy.

	Majority	Independence policy			Correlation-aware policy		
Verifier	of 5	Acc.	Queries	Over	Acc.	Queries	Over
CLIP, infrared	0.603	0.732	4.5	+0.14	0.830	4.9	−0.01
CLIP, visible	0.861	0.886	1.9	+0.08	0.886	2.8	+0.02
Probe, infrared	0.906	0.888	1.9	+0.08	0.911	3.2	+0.00
Probe, visible	0.949	0.930	1.4	+0.05	0.943	2.1	+0.02
Qwen yes/no, visible	0.563	0.829	3.8	+0.07	0.829	3.9	+0.01

policy stops early on overconfident posteriors. What did transfer is calibration. The independence policy's stated confidence exceeded its accuracy by 5 to 14 points for every usable verifier, while the correlation-aware policy stayed within about 2 points. Accuracy gains appeared only where errors were most heterogeneous: for zero-shot CLIP on infrared the correlation-aware policy reached 83.0% accuracy, while the independence policy never exceeded 78.6% at any stopping threshold and grew *less* accurate as it was allowed more queries, because it kept over-weighting correlated votes. For the probe and visible CLIP, the two policies were within about 2 points of each other at matched query budgets.

3.4 Adaptation raises accuracy but concentrates errors

A linear probe on CLIP [26] image embeddings reached 89% drone recall in both sensors. In infrared, drone recall rose from 46% to 89% while drone τ rose from 1.01 (0.76 to 1.30) to 2.62 (1.85 to 3.82); the intervals do not overlap. The probe's remaining errors fall on a small set of hard clips, and in infrared they fall on different clips than zero-shot CLIP's (per-clip correlation 0.30). Better models therefore do not remove the need for correlation-aware fusion; they can make it more important.

3.5 A trained detector clusters as strongly or more

Table 3 reports YOLO11n. Validation mAP50 reached 0.994, but held-out recall was 0.79 to 0.88; validation sequences share sessions with training, so the validation figure reflects familiar scenes. Misses cluster strongly within sequences in every set, and more strongly for far frames. In set B the single-level infrared τ of 3.68 absorbs a large session effect (Section 3.6). Clustering is robust to the training seed: across three seeds, set-B τ ranged from 2.12 to 3.68 in infrared and 1.84 to 2.45 in visible, every interval excluding zero. Recall itself was less stable (visible 0.80 for the original seed, 0.90 for the other two), so single-run accuracy figures should be read cautiously. Re-evaluating visible frames at 1280 px changed visible recall only from 0.801 to 0.805, so downscaling does not explain the visible results.

Other architectures. To test whether these findings are specific to one detector, we trained YOLO11s and the transformer-based RT-DETR-L [27] on the same split and evaluated them on set B (Table 2). Clustering replicated in every architecture and sensor. The most accurate model, RT-DETR-L (recall about 0.95 in both sensors), had the most concentrated infrared errors ($\tau = 5.08$), repeating the pattern seen with the CLIP probe: higher accuracy, more concentrated residual errors. Cross-sensor dependence replicated for YOLO11s (odds ratio 5.80, 95% CI 2.33 to 14.43). For RT-DETR-L the point estimate was similar (4.57) but the interval included 1 (0.49 to 42.5), because a 95%-recall detector leaves too few misses to estimate dependence; we report it as inconclusive. YOLO11s was less accurate than YOLO11n on held-out sessions (visible recall

Table 2: Three detector architectures on test set B (30 sequences, 10 sessions).

Detector	Recall IR	Recall vis	τ IR [95% CI]	τ vis [95% CI]	MH odds ratio [95% CI]	Val. mAP50
YOLO11n	0.876	0.801	3.68 [2.98, 4.66]	2.45 [1.56, 3.40]	2.45 [1.08, 5.58]	0.994
YOLO11s	0.836	0.726	3.09 [1.83, 4.65]	3.03 [2.16, 4.72]	5.80 [2.33, 14.43]	0.995
RT-DETR-L	0.951	0.949	5.08 [3.17, 6.00]	2.44 [1.27, 3.46]	4.57 [0.49, 42.52]	0.985

Table 3: YOLO11n misses on held-out Anti-UAV RGB-T data. All τ have boundary-corrected LR $p < 10^{-15}$.

Set	Sensor	Recall	τ [95% CI]	Sequences / sessions	Frames
Test A	Infrared	0.790	1.90 [1.38, 2.79]	40 / 5	699
Test A	Visible	0.845	1.74 [1.18, 2.79]	40 / 5	699
Test B	Infrared	0.876	3.68 [2.98, 4.66]	30 / 10	507
Test B	Visible	0.801	2.45 [1.56, 3.40]	30 / 10	507
Far sequences	Infrared	0.726	2.24 [1.50, 2.91]	18 / 6	705
Far sequences	Visible	0.670	2.11 [1.44, 4.39]	18 / 6	705
Far frames	Infrared	0.661	3.19 [1.55, 4.66]	18 / 6	245
Far frames	Visible	0.543	3.70 [2.24, 20.86]	18 / 6	245

0.726 against 0.801) despite slightly higher validation mAP50 (0.995), consistent with greater capacity fitting the training sessions more closely, though single-seed differences of this size are within the seed variation we observed.

3.6 Hard targets and hard scenes

With reconstructed scenes on Halmstad (HOG baseline only), clip-level τ stayed significant at every K. Scenes explained much of the visible-drone clustering (clip share 15% to 30%) and a minority in infrared (59% to 69% clip share), while bird clustering was almost entirely clip-level (79% to 100%). Repeating the analysis for the learned verifiers at $K = 20$, drone clustering was mostly clip-level: 96% for zero-shot CLIP in visible, 84% in infrared, 68% for Qwen, and 100% for the visible probe, with clip-level τ significant in every case ($p \leq 0.01$). The exception was the infrared probe (41% clip share, $\sigma = 1.93$). Two caveats apply: drone clips fell into only 5 to 8 scene groups, so σ is weakly identified, and in infrared, drone and bird clips occupy almost disjoint scene groups because the classes were filmed in different settings.

With real sessions on Anti-UAV RGB-T (Table 4), estimates from a single test set with 5 to 10 sessions are unstable, as expected for a variance estimated from few groups. Pooling set B with the far-range set (15 sessions), about half the clustering variance is sequence-level in both sensors (51% infrared, 45% visible). Profile-likelihood intervals, which describe the uncertainty more honestly than a test against zero when there are few groups, exclude zero for both components in both sensors: in infrared, $\sigma = 1.73$ (95% CI 1.18 to 2.32) and $\tau = 1.76$ (1.33 to 2.21); in visible, $\sigma = 2.06$ (1.46 to 3.66) and $\tau = 1.87$ (1.32 to 3.00). Both sources of clustering are therefore real, although with 15 groups these intervals rest on a large-sample approximation and may be somewhat narrow.

All 37 sessions by cross-fitting. With session cross-fitting (Section 2.8), every one of the 315 usable sequences has an out-of-session prediction. Trained on many sessions, the detectors generalized far better than our original detector trained on 35 sequences: out-of-session recall was 0.959 in infrared and 0.977 in visible. Clustering remained strong (single-level $\tau = 2.62$, 95% CI 2.20 to 3.09, infrared; 1.94, 1.50 to 2.44, visible). In the nested model, every variance component was clearly nonzero: in infrared, $\sigma = 1.89$ (1.43 to 2.78) and $\tau = 1.46$ (1.08 to 1.87), a sequence share of 0.37; in visible, $\sigma = 1.36$ (0.86 to 2.07) and $\tau = 1.36$

Table 4: Sequence versus session clustering for YOLO11n with real session labels. For the pooled 15-session and the cross-fitted 37-session fits, brackets are 95% profile-likelihood intervals.

Set (sessions)	Sensor	τ sequence	σ session	Sequence share
Test A (5)	Infrared	1.10	4.60	0.05
Test A (5)	Visible	2.01	0.00	1.00
Test B (10)	Infrared	1.15	4.07	0.07
Test B (10)	Visible	1.87	1.87	0.50
Test B + far (15)	Infrared	1.76 [1.33, 2.21]	1.73 [1.18, 2.32]	0.51
Test B + far (15)	Visible	1.87 [1.32, 3.00]	2.06 [1.46, 3.66]	0.45
All, cross-fitted (37)	Infrared	1.46 [1.08, 1.87]	1.89 [1.43, 2.78]	0.37
All, cross-fitted (37)	Visible	1.36 [0.99, 1.90]	1.36 [0.86, 2.07]	0.50

(0.99 to 1.90), a share of 0.50. Because each session was scored by only one of the two cross-fitted detectors, we refit the nested model with the fold as a fixed effect: the estimates were essentially unchanged (infrared $\sigma = 1.88$, $\tau = 1.46$, share 0.38; visible $\sigma = 1.37$, $\tau = 1.35$, share 0.49), and a likelihood-ratio test found no detector effect beyond the session and track effects (infrared $p = 0.95$, visible $p = 0.82$), so the session effect is not an artifact of the cross-fitting design. The 15-session analysis therefore overstated the target-level share in infrared (0.51); with all sessions, scene conditions account for about two thirds of infrared clustering and half of visible clustering. One pattern replicated across both test sets: infrared misses on the test split are dominated by session effects ($\sigma = 4.60$ in set A, 4.07 in set B), while visible splits evenly and far-range infrared shows no session effect. We hypothesize, without testing it, that thermal contrast varies between sessions with weather and time of day.

3.7 Cross-sensor dependence depends on the model

For zero-shot CLIP on 60 Anti-UAV RGB-T sequences, the two sensors' errors were close to conditionally independent: both were wrong on 22.8% of frame pairs against 22.6% expected, the per-sequence correlation was 0.02, and the Mantel–Haenszel odds ratio was 0.94 ($p = 0.7$). The trained detector behaves differently (Table 5). At normal range, its infrared and visible misses co-occur on the same frames, with odds ratios of 7.97 and 2.45 in the two test sets; pooled over the 55 unique sequences of both sets (15 appear in both), the odds ratio is 4.98 (95% CI 2.89 to 8.60). The cross-fitted predictions for all 315 sequences confirm it decisively: both sensors missed the same frame on 0.8% of frames against 0.1% expected under independence (odds ratio 6.26; session-cluster bootstrap 95% interval 3.46 to 10.57, resampling whole sessions), and the per-sequence miss rates correlate at 0.45 (session-bootstrap 95% interval 0.32 to 0.59). The direction held for every training seed (2.45, 18.6 and 7.24, each interval above 1) and with visible frames at 1280 px (2.78, $p = 0.003$), though the magnitude varies widely between seeds. A formal Breslow–Day homogeneity test could not be computed because many per-sequence tables have empty cells. On far frames, both sensors miss together more often than independence predicts (0.249 against 0.155), yet the within-sequence odds ratio is 1.08: the shared failure concerns which distant targets are hard, not which frames. An earlier analysis on only 9 far-range sequences gave a per-sequence correlation of 0.90; with 18 sequences it falls to about 0.5 with a session interval reaching zero, so we do not rely on it.

3.8 The label-free audit on real data

On Halmstad, with five views (HOG, zero-shot CLIP, the CLIP probe, Qwen yes/no, and a motion classifier) each reduced to one vote per clip, the latent class audit estimated sensitivity and specificity to within 4.1

Table 5: YOLO11n cross-sensor dependence. Per-sequence correlation intervals are session-level bootstrap 95% intervals.

Set	Miss IR	Miss vis	Both miss (obs. / indep.)	MH odds ratio	Per-seq. corr.
Test A	0.210	0.155	0.083 / 0.032	7.97 ($p = 10^{-13}$)	0.37 [0.06, 0.78]
Test B	0.124	0.199	0.047 / 0.025	2.45 ($p = 0.022$)	0.39 [0.06, 0.66]
A + B (55 seq.)				4.98 [2.89, 8.60]	
All 315 seq., cross-fitted	0.041	0.023	0.008 / 0.001	6.26 [3.46, 10.57]	0.45 [0.32, 0.59]
Far sequences	0.274	0.330	0.157 / 0.090	4.86 ($p = 0.0006$)	0.53 [0.05, 0.64]
Far frames	0.339	0.457	0.249 / 0.155	1.08 ($p = 0.95$)	0.47 [−0.27, 0.79]

(visible) and 2.5 (infrared) points on average without using labels. Fused labels reached 94.3% and 94.2% accuracy, matching the best single view in visible and beating it by 2.7 points in infrared. The random-effects model detected the dependence that existed in visible (likelihood ratio 17.5 on 4 parameters, $p \approx 0.002$, with loadings on CLIP and Qwen, whose true within-class error correlation was 0.39) but missed a stronger one between HOG and the probe (0.44), and it did not improve accuracy estimates over Dawid–Skene at these sample sizes. Both models overstated the sensitivity of the strongest views (true 0.93 to 0.97, estimated 0.98 to 1.00). HOG and the probe were trained with labels through cross-validation, so they are not fully label-free views.

3.9 An audit on real detector votes

On set B we collected 516 frames with ground truth (496 with a drone in both sensors, 20 with none) and four YOLO11n views: infrared, visible, and each sensor again half a second later. Misses repeat strongly within a sensor: among drone frames, the odds ratio between a miss now and a miss half a second later was 40.2 in infrared and 33.8 in visible, a direct real-data measure of how little a re-query adds. Fitting Dawid–Skene to these votes gave a badly wrong audit: it estimated drone prevalence at 0.72 (true 0.96) and overstated every view's sensitivity, by 8 points in infrared (0.98 against 0.90) and 15 to 18 points in visible (0.95 to 0.98 against 0.80), the direction predicted for dependent tests [1, 2]. Dropping either temporal view reduced but did not remove the bias. Because there were only 20 drone-free frames, specificity cannot be checked and the joint-pair model is not identifiable with two sensors; the calibrated simulation in Appendix A.2 supplies the comparison the real data cannot.

3.10 Real negatives: a cross-dataset test on AVID

Because Anti-UAV RGB-T contains almost no drone-free frames, we scored the 13 held-out AVID videos with the YOLO11n detector trained on Anti-UAV RGB-T. This is a cross-dataset test: the detector never saw AVID's cameras or any bird, so its accuracy is low (infrared sensitivity 0.84 and specificity 0.45; visible 0.36 and 0.76), and the results below concern error structure, not detector quality. A frame counts as a detection when any box has confidence at least 0.25.

False alarms cluster as strongly as misses. Using 25-frame segments of each video as tracks, misses on drone frames clustered strongly ($\tau = 3.38$ infrared, 7.24 visible), and so did false alarms on bird frames ($\tau = 3.29$ infrared, 12.0 visible; every interval far from zero). A bird that triggers a false alarm tends to keep triggering it. Same-frame dependence between sensors replicated on drone frames (Mantel–Haenszel odds ratio 11.98, 95% CI 1.59 to 90.2), making AVID the third dataset to show it; on bird frames, infrared and visible false alarms never co-occurred within a segment, so no dependence could be estimated there.

Dependence decays with time, and differently by sensor. The random-intercept model behind τ

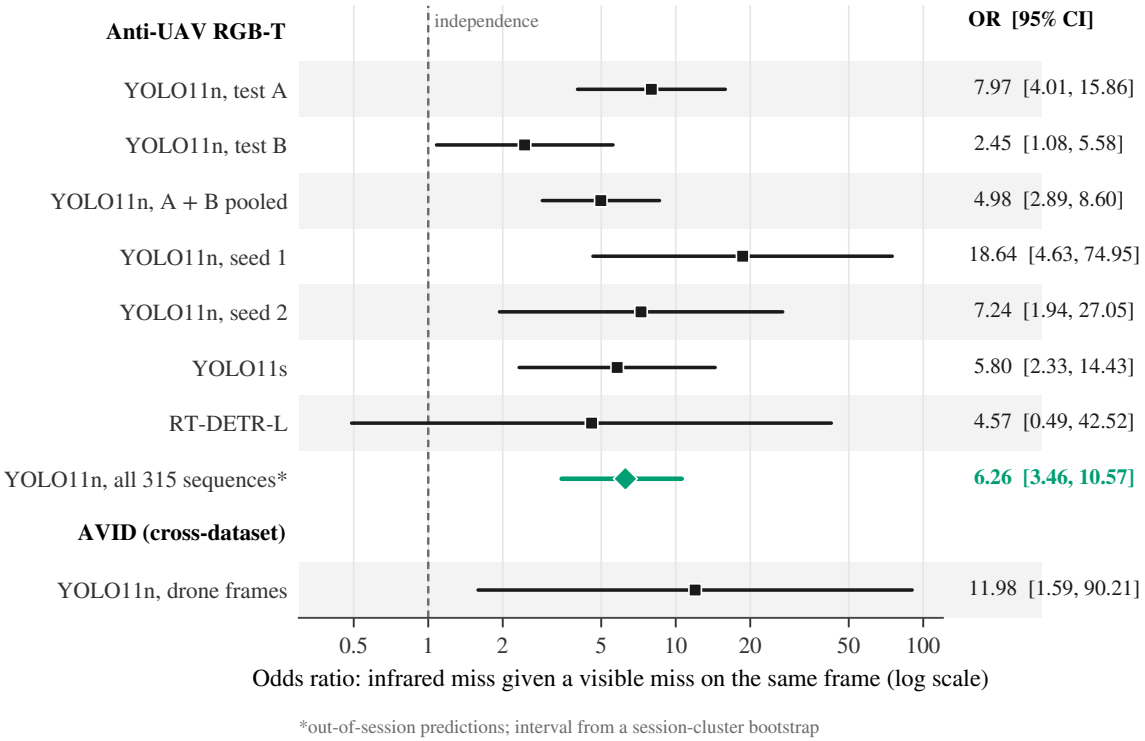


Figure 4: Same-frame dependence between infrared and visible misses across test sets, training seeds, detector architectures and datasets, with estimates and 95% intervals listed at right. Every estimate lies above independence (dashed line); only RT-DETR-L, with about 5% misses, has an interval reaching below 1. The highlighted row uses out-of-session predictions for all 315 Anti-UAV RGB-T sequences.

assumes any two frames of a track are equally correlated. AVID's full frame rate lets us check this (Figure 5). Infrared errors have two parts: a lasting track component (correlation about 0.26 to 0.30 that persists across the 7 s observed) and a short-lived component (0.34 to 0.40) that fades within about 0.7 s. Visible misses decay slowly, over about 6 s, with no lasting part in these 10 s videos, while visible false alarms barely decay at all (0.93 at 0.04 s, 0.91 at 7 s): a bird that fools the visible detector keeps fooling it. Equal correlation is therefore wrong for three of the four error types. For the Anti-UAV analyses, frames were sampled 1 to 2 s apart, beyond the infrared short-lived component, so those τ estimates mainly reflect lasting track effects; for visible errors they also absorb multi-second correlation.

The audit fails worse once specificity can be checked. Dawid–Skene on the four views (each sensor now and half a second later) estimated drone prevalence at 0.34 against a true 0.65 (error -32 points, video-level bootstrap 95% interval -51 to -11) and overstated visible sensitivity by 57 points (0.93 against 0.35; interval 38 to 77), while understating infrared specificity (Figure 6, left). The strongly dependent temporal pairs pulled the latent class toward "the visible detector is firing" rather than "a drone is present." With three views the prevalence estimate was close (0.69 against 0.65), but every accuracy was still overstated, by 9 to 27 points.

Fusion: calibration improves, accuracy does not. Table 7 compares frame-level fusion rules under leave-one-video-out cross-validation. At the frame level no fusion rule beat the infrared detector alone on Brier score, because the visible detector was so weak; the independent (naive Bayes) rule was the most overconfident (9.6 points). At the segment level we split each video into consecutive, complete 25-frame blocks (123 segments; the final partial block of each video is discarded) and give 95% intervals from a video-level bootstrap that resamples whole videos within class, so that the 4 bird videos, not the frames, set

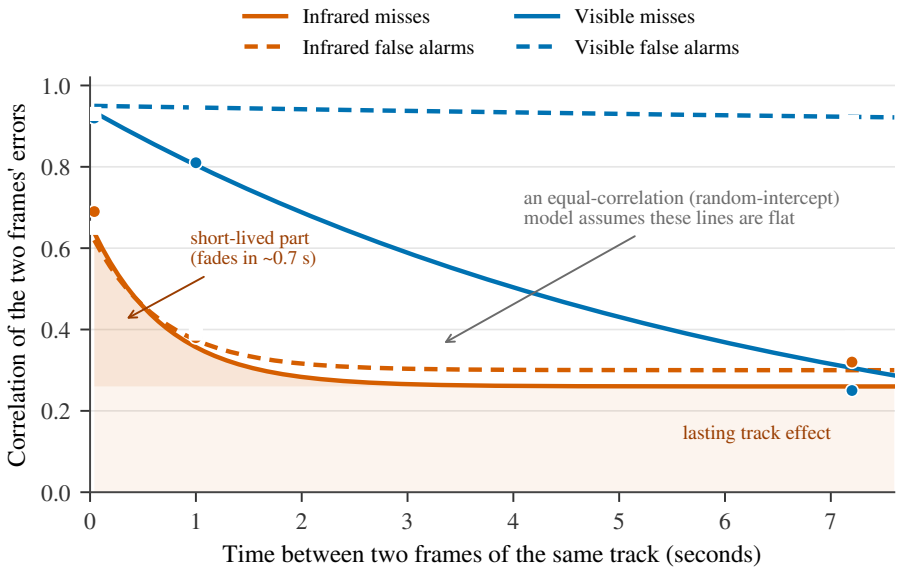


Figure 5: How the correlation between two frames' errors decays with the time between them, on AVID. Points are observed correlations and curves are fits of a lasting component plus an exponentially decaying one; for infrared misses the two components are shaded. A random-intercept model assumes each curve is flat.

Table 6: Combining evidence over complete 25-frame segments on AVID (123 segments; leave-one-video-out; cross-dataset detector). Brackets are 95% video-level bootstrap intervals (videos resampled within class). Over = mean stated confidence minus accuracy.

Rule	Accuracy	NLL	Over
Independent over time	0.691 [0.491, 0.848]	3.321 [1.401, 5.903]	+0.287 [+0.133, +0.486]
Design-effect corrected (Eq. 4)	0.699 [0.496, 0.860]	0.751 [0.449, 1.123]	+0.103 [−0.055, +0.296]
Segment mean	0.675 [0.487, 0.831]	0.704 [0.519, 0.929]	+0.052 [−0.121, +0.246]

the uncertainty (Table 6). Treating the 25 frames as independent evidence, as a standard sequential Bayes filter does, was badly overconfident (28.7 points, NLL 3.32). The design-effect correction (Eq. 4), which scales the summed evidence down to an effective sample size $n/(1+(n-1)\hat{\rho})$ with the within-segment correlation $\hat{\rho}$ estimated on training videos, cut overconfidence to 10.3 points and NLL to 0.75; the reduction, 18.4 points (95% interval 14.9 to 22.6), is robust to which videos are sampled. Simply averaging frame evidence did about as well (overconfidence 5.2 points, NLL 0.70), because with correlation this high the effective sample size is under two frames. Accuracy was about 0.69 for all three rules, with wide, overlapping intervals, so we claim no accuracy improvement. Section 2.6 shows why the correction works: it is the exact Bayes rule under Gaussian, calibrated, exchangeable evidence, and the estimated frame evidence was close to calibrated (class means -0.56 and 0.40 against ± 0.43 for perfect calibration). In an earlier pass that used fixed frame-number blocks, including short leftover blocks, the derived exchangeable rule performed almost identically to the design-effect rule, and a random-intercept plus AR(1) rule (fitted $a = 0.52$, $b = 0.33$, $\phi = 0.71$) had the highest accuracy, again within sampling error.

Table 7: Frame-level fusion on AVID held-out videos (leave-one-video-out; cross-dataset detector). Over = mean stated confidence minus accuracy. Segment-level results are in Table 6.

Level	Rule	Acc.	Brier	NLL	ECE	Over
Frame	Infrared only	0.671	0.231	0.673	0.076	+0.063
	Visible only	0.658	0.292	0.810	0.430	+0.005
	Independent (naive Bayes)	0.645	0.257	0.756	0.136	+0.096
	Stacked	0.660	0.258	0.761	0.203	+0.080
	Stacked + interaction	0.658	0.250	0.754	0.161	+0.100

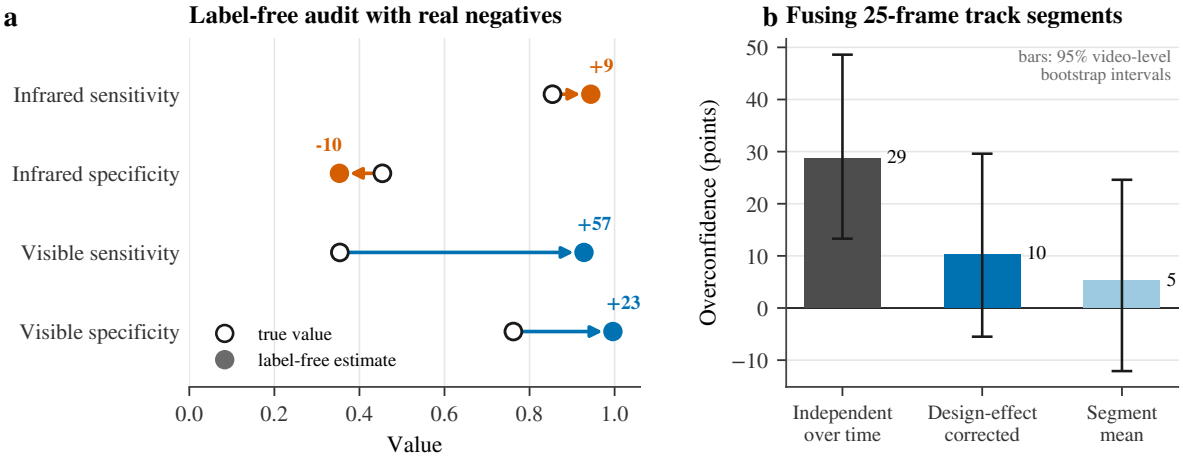


Figure 6: AVID held-out videos with real bird negatives. Left: a four-view Dawid–Skene audit misestimates every accuracy measure, by up to 57 points. Right: treating the 25 frames of a complete track segment as independent evidence is badly overconfident; the design-effect correction removes most of that overconfidence (error bars: 95% video-level bootstrap intervals, 123 segments).

4 Discussion

Three practical consequences follow. First, re-querying a verifier on the same track buys much less evidence than independence implies, and the shortfall grows as models improve, because adaptation removes easy errors and leaves concentrated hard ones. Fusion rules and confidence estimates should model track-level correlation explicitly; on real votes, doing so mainly buys honest confidence rather than fewer queries or higher accuracy. On AVID, discounting temporally redundant evidence cut segment-level overconfidence from 29 to 10 points. Second, the remedy depends on the source of hardness. Half to two thirds of the clustering is scene-level, which favors sampling new viewpoints or conditions; the rest travels with the target, which only a physically different sensor can address. Third, that different sensor is less independent than assumed. For zero-shot CLIP, visible and infrared errors were nearly independent, but for a trained detector they were not, and at far range both sensors fail on the same targets. Label-free audits that rely on cross-sensor independence will then overstate accuracy, with a bias that more data does not fix; modeling the dependent pair jointly removes the bias at a cost in variance that is worth paying when dependence is strong or audits are large.

The gap between validation mAP50 (0.994) and held-out recall (0.79 to 0.88) is a further warning: error rates measured on familiar sessions do not transfer, and evaluation should be split by recording session, not by sequence or frame.

What the evidence establishes. The paper makes three kinds of claims, supported by different kinds of evidence. *Measurement of error dependence* rests on real data throughout: within-track clustering across five model families, three detector architectures and two independent recording sources; session effects across all 37 Anti-UAV RGB-T sessions with out-of-session predictions; same-frame cross-sensor dependence on Anti-UAV RGB-T and AVID; and the decay of dependence with time. *Correcting overconfidence* is supported on real data: discounting correlated evidence made track-level confidence honest on AVID and on Halmstad votes, without improving accuracy. *Correcting label-free accuracy estimates* is only partly supported on real data: we show on real votes that the standard (independence) audit fails, but evidence that a dependence-aware audit fixes it comes from calibrated simulation (Appendix A.2), because real data with enough negatives and enough independent views to test the fix were not available.

Beyond counter-UAS. Nothing in the statistical structure studied here is specific to drones. Any system that observes the same object repeatedly or through several instruments, such as automated industrial inspection, condition monitoring or medical imaging, has the same three levels of potential dependence (observation, object, session), and the same tools apply: random-effects models to measure clustering, stratified odds ratios to test whether sources fail together, cross-fitting by acquisition session to avoid optimistic evaluation, and design-effect corrections when combining evidence. We have validated these tools only on drone video.

4.1 Limitations

Far-range data are scarce: only 18 of 318 Anti-UAV RGB-T sequences qualified, and none contain tiny targets (10 px or less). Session-level variances in the cross-fitted analysis rest on 37 sessions, enough for usable profile intervals but still modest; other datasets would need their own session labels to test generality. Cross-sensor dependence was measured on Anti-UAV RGB-T (zero-shot CLIP and three detector architectures) and on AVID; because AVID comes from the Halmstad recordings, these are two independent recording sources, not three. On Halmstad, scenes were reconstructed from video content rather than recorded, and drone and bird clips occupy nearly disjoint scene groups in infrared. The random-intercept model behind τ assumes equal correlation within a track, which the AVID lag analysis shows is wrong for most error types; τ should be read as a summary at the sampling interval used (1 to 2 s on Anti-UAV). The size of the cross-sensor odds ratio varies widely across training seeds (2.45 to 18.6), although its direction does not. Anti-UAV RGB-T contains almost no drone-free frames (20 of 516 eligible); our specificity results come instead from AVID, scored with a detector trained on a different dataset that had never seen birds, over only 13 videos (4 of them bird videos). A detector trained on AVID itself is needed to confirm them. On Halmstad, scene groups are nearly confounded with class in infrared. Only two small VLMs were tested. Simulation parameters are ours and show mechanisms rather than expected field performance. Bird classes with under 5% error cannot separate τ from zero at these sample sizes.

5 Conclusions

Verifier errors in counter-UAS perception are not independent across repeated queries, across models, or, for trained detectors, across sensors. Random-effects and latent class tools from diagnostic-test evaluation measure this structure directly and correct the audits that depend on it. Systems should be fused, audited and validated as if their errors were correlated, because they are.

Statements

Author Contributions: Conceptualization, methodology, software, formal analysis, investigation, visualization and writing, S.G. The author has read and agreed to the submitted version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used are public: the Halmstad multi-sensor drone dataset [16], Anti-UAV RGB-T [17] (public copy on Hugging Face), and AVID [18] (CC BY-NC 4.0). Analysis code, including exact-likelihood random-effects estimation, the derived fusion rules, session cross-fitting and the figure scripts, is available at **[GitHub repository URL]** and archived on Zenodo (**[Zenodo DOI]**).

Acknowledgments: During the preparation of this work the author used Claude (Anthropic) to assist with writing analysis code, running experiments, preparing figures and drafting text. The author reviewed and edited all outputs and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

A Supplementary results

A.1 Simulation: correlation-aware querying

In a simulated track-verification pipeline with strongly clustered verifier errors ($\tau = 2$), a filter that treated repeated VLM queries as independent reached 16% error and overstated its confidence by 12 points. A correlation-aware filter with a value-of-information query rule held 5.3% error using 3.8 VLM calls per track, against 7.2 for a fixed uncertainty band. In a separate audit simulation, Dawid–Skene misestimated sensor accuracies by 8 to 11 points when views were dependent, and a nested likelihood-ratio test flagged the dependence in 15 of 16 runs at moderate and 16 of 16 at strong dependence (3 of 16 false flags), cutting audit error from 12.4 to 3.8 points in the worst case when the model was refit. All simulation parameters were chosen by us; these results illustrate mechanisms, not field performance. Section 3.3 shows that the query savings did not hold on real votes.

A.2 What measured dependence costs the audit

The YOLO11n evaluation cannot itself be audited without labels: every evaluated frame contains a drone, and there are only two views. We therefore ran a calibrated simulation. Tracks are drones with probability 0.6. Four views vote once per track: infrared (sensitivity 0.85, specificity 0.90) and visible (0.82, 0.88), loosely calibrated to YOLO11n recall; a motion cue (0.75, 0.70); and a conservative zero-shot VLM (0.60, 0.95), loosely calibrated to Halmstad. Infrared and visible errors are dependent within each class with odds ratio 1, 2.5 or 8, spanning the measured range; the other views are conditionally independent. We compared Dawid–Skene, the joint-pair model, and a direct estimate from 30 labeled tracks, over 300 replicates of 300 tracks and 40 replicates of 5,000 tracks.

Figure 7 shows the result. Dawid–Skene's error grows with dependence and does not shrink with sample size: at odds ratio 8 it overstated infrared sensitivity by 6.3 points with 300 tracks and 6.8 points with 5,000, as the dependent pair pulls the latent class toward its own votes. The joint-pair model is consistent, with error near 1.2 points at 5,000 tracks for every odds ratio, but it spends parameters, so with 300 tracks it is noisier (about 4 points) and carries a small-sample specificity bias of 3 to 4 points that vanishes with more data. The crossover lies between the two measured odds ratios: at 2.5 the simpler model is slightly better with 300 tracks (3.4 against 4.1 points), at 8 the joint model is better (3.9 against 5.1). Thirty labeled tracks alone were

Table 8: Within-clip clustering on Halmstad (exact ML). Holm adjustment is across the 12 non-HOG tests; HOG tests were adjusted separately.

Model	Sensor	Class	Error	τ (95% CI)	Holm p
HOG baseline	Visible	Drone	10.9%	1.29 (0.79–1.92)	< 0.001
	Visible	Bird	26.3%	1.39 (0.80–2.23)	< 0.001
	Infrared	Drone	9.7%	1.40 (0.95–1.97)	< 0.001
	Infrared	Bird	17.4%	1.82 (1.21–2.72)	< 0.001
CLIP zero-shot	Visible	Drone	23.6%	1.45 (1.01–2.02)	< 0.001
	Visible	Bird	4.7%	1.13 (0–2.66)	0.31
	Infrared	Drone	54.4%	1.01 (0.76–1.30)	< 0.001
	Infrared	Bird	7.6%	1.90 (1.09–3.26)	< 0.001
Qwen yes/no	Visible	Drone	59.6%	0.74 (0.42–1.07)	0.003
	Visible	Bird	2.8%	0.00 (0–2.04)	1.00
	Infrared	Bird	1.7%	2.11 (0.39–6.07)	0.08
CLIP probe	Visible	Drone	10.6%	1.37 (0.86–2.07)	< 0.001
	Visible	Bird	4.3%	0.92 (0–2.38)	0.51
	Infrared	Drone	11.2%	2.62 (1.85–3.82)	< 0.001
	Infrared	Bird	10.3%	1.06 (0.29–1.91)	0.08

Qwen on infrared drones is omitted: with 96.7% error, τ is not identifiable.

worst throughout (about 6.6 points), so a small labeled set is better used to anchor a latent class model than as a replacement for it.

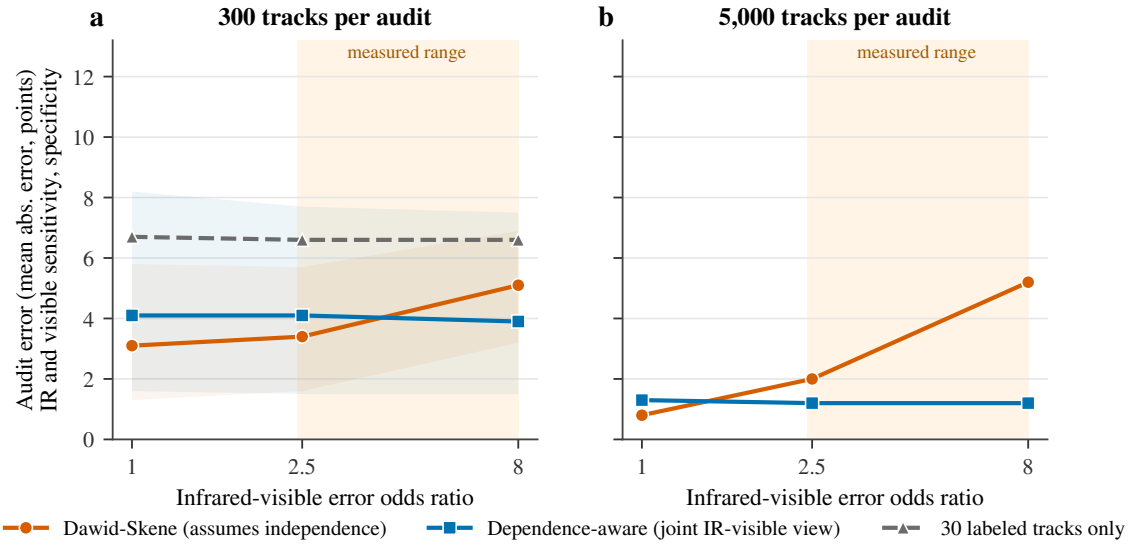


Figure 7: Audit error for infrared and visible sensitivity and specificity as infrared–visible dependence grows; the shaded column marks the odds ratios measured on real detectors. (a) 300 tracks per audit, 300 replicates, with 90% ranges shaded. (b) 5,000 tracks, 40 replicates: the independence model's error persists, while the dependence-aware model's vanishes.

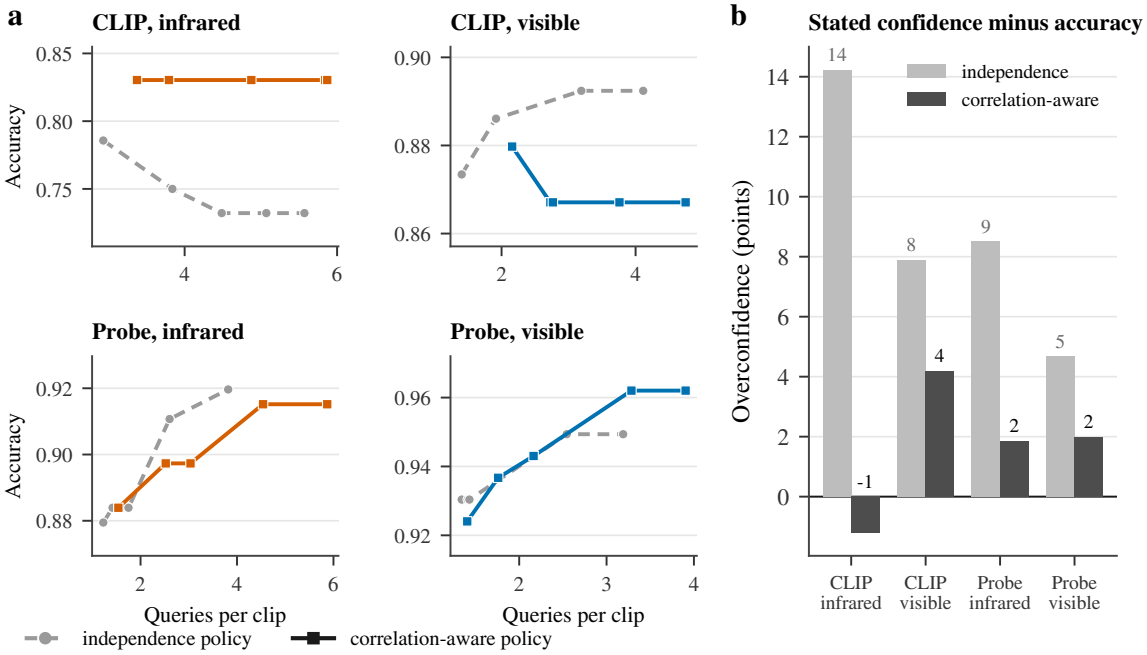


Figure 8: Query policies replayed on real Halmstad votes. (a) Accuracy against mean queries per clip as the stopping threshold varies from 0.8 to 0.999, one panel per verifier; for zero-shot CLIP on infrared the independence policy becomes *less* accurate as it queries more. (b) Overconfidence (stated confidence minus accuracy) at threshold 0.95.

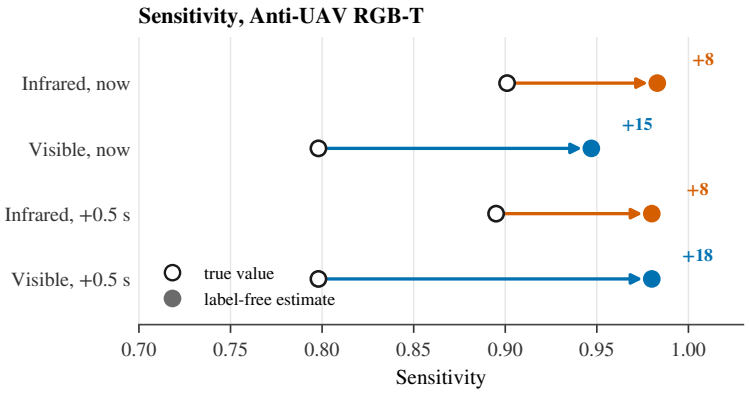


Figure 9: A Dawid–Skene audit of real YOLO11n votes (four views: each sensor now and half a second later) overstates every view's sensitivity, by 8 to 18 points.

References

[1] P. M. Vacek. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41(4):959–968, 1985.

[2] V. L. Torrance-Rynard and S. D. Walter. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 16(19):2157–2175, 1997.

[3] Y. Qu, M. Tan, and M. H. Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810, 1996.

[4] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society, Series C*, 28(1):20–28, 1979.

[5] S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980.

[6] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[7] A. Jaffe, B. Nadler, and Y. Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *Proc. AISTATS*, 2015.

[8] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger. Unsupervised ensemble learning with dependent classifiers. In *Proc. AISTATS*, PMLR 51:351–360, 2016.

[9] H. Xu et al. Incorporating conditional dependence in latent class models for probabilistic record linkage. *Annals of Applied Statistics*, 13(3), 2019.

[10] Context-aware Bayesian fusion for multimodal object detection. arXiv:2610.18741, 2026.

[11] Uncertainty-aware multimodal anti-UAV detection via evidential fusion and conflict-discounted belief aggregation. arXiv:2608.29235, 2026.

[12] An evidence hierarchy for Bayesian object classification via OSINT-aided heterogeneous sensor fusion. arXiv:2605.22259, 2026.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, 2001.

[14] D. Madras, T. Pitassi, and R. Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018.

[15] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[16] F. Svanström, F. Alonso-Fernandez, and C. Englund. A dataset for multi-sensor drone detection. *Data in Brief*, 39:107521, 2021. doi:10.1016/j.dib.2021.107521.

[17] N. Jiang et al. Anti-UAV: a large-scale benchmark for vision-based UAV tracking. *IEEE Transactions on Multimedia*, doi:10.1109/TMM.2021.3128047.

[18] AVID: a weakly aligned RGB-infrared dataset and perturbation benchmark for tiny aerial targets. arXiv:2610.40949, 2026.

[19] S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[20] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[21] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

[22] L. Kish. *Survey Sampling*. Wiley, New York, 1965.

[23] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press, 2002.

[24] G. Jocher and J. Qiu. Ultralytics YOLO11. Software, `https://github.com/ultralytics/ultralytics`, 2024.

[25] S. Bai et al. Qwen2.5-VL technical report. arXiv:2502.13923, 2025.

[26] A. Radford et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

[27] Y. Zhao et al. DETRs beat YOLOs on real-time object detection. In *Proc. CVPR*, 2024.