WEAK QUERY CALIBRATION FOR MULTI-VIEW UNIQUE COUNTING UNDER UNRELIABLE OVERLAP PROXIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce DAWSON, a weak-query calibration framework for multi-view unique counting with a small budget of aggregate shared-person answers. It treats each answer as sparse count-level supervision: per-view counts describe observations, while a queried pair supplies a shared-person cardinality without requiring cross-view identities. People visible in several cameras create duplicate mass, and appearance proxies weaken under occlusion, viewpoint changes, crop loss, and detector noise. Thus, the pair with the largest estimated overlap need not be the query that most improves the final union estimate. DAWSON combines per-view counts and pairwise overlap proxies into a fused unique-count estimate. Its expected union-error reduction (EER) rule evaluates candidate answers through an observable response surrogate and selects the query with the largest expected decrease in a calibration residual. The residual couples count consistency with uncertainty in unrevealed overlaps, so the weak-label budget follows the counting state. Each answer updates the fused estimate and ranking of remaining candidates. Selection uses observable counts, proxy similarities, and revealed responses, while keeping cross-view identities and the ground-truth unique count outside acquisition. Experiments on synthetic scenes and the static-camera WILDTRACK image-proxy bridge use annotation-simulated responses. With matched inputs, estimators, thresholds, and two-query budgets, EER yields lower paired MAE than Largest-overlap in all twelve degraded settings across four frozen encoders, with every paired 95% confidence interval below zero. The gap reaches -14.68 MAE for DINOv3 at Strong degradation. For YOLO26m, relative MAE reductions are 12.8%, 11.1%, and 13.2% at Mild, Medium, and Strong, and persist with noisy answers. Reliable-proxy controls retain Largest-overlap as a strong baseline, with same-input Max providing a zero-query reference for assessing weak-query calibration. These results establish residual-driven query selection as a proxy-regime-dependent mechanism for correcting overlap error under constrained budgets.

1 INTRODUCTION

Multi-view object counting is often evaluated per camera, while many applications require the number of distinct physical objects across synchronized views. Summing visible counts can over-count when the same person or vehicle appears in several cameras, disappears behind an occluder, or re-enters a scene. Dense cross-view identity supervision would resolve this ambiguity, but it is expensive and brittle for large multi-camera deployments (Chavdarova et al., 2018; Zhang & Chan, 2019; 2022; Hou et al., 2020; Daryani et al., 2025; Zheng et al., 2015; Zhou et al., 2019). The missing quantity is duplicate mass: per-view counts reveal how many observations exist, while overlap information determines how many observations refer to the same person (Figure 1).

This setting is difficult for three coupled reasons. First, duplicate mass is latent: the same per-view counts can correspond to different cross-view unions, so a count-consistent estimate is not identifiable without overlap evidence. Second, the available evidence is a fallible proxy whose errors are structured by occlusion, viewpoint, crop visibility, and detector noise; suppression can hide precisely the high-overlap pairs that most affect the union. Third, a query returns only an

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

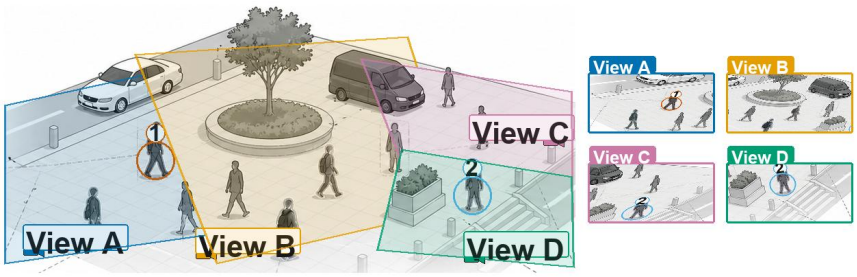

Figure 1: Overlapping camera views repeatedly observe the same people, so naively summed per-view counts over-estimate the unique total.

aggregate pair count and changes the state used to evaluate later pairs. The learner therefore must spend a finite budget while estimating which unrevealed response will reduce the current union error, without using identities or the final ground truth.

DAWSON turns this gap into a sparse aggregate weak-supervision problem. Per-view counts and pair/window overlap proxies define the observable state, and a small number of view-pair answers calibrate it. Each answer reports a pair's shared-person cardinality, replaces that pair's duplicate-mass proxy, and updates the fused set-union estimate. Because answers arrive sequentially, a finite weak-label budget becomes a sequence of calibration updates. In WILDTRACK, annotation-box crops construct the appearance proxy and the response remains an aggregate count, so weak supervision acts at final-count calibration rather than identity reconstruction.

A simple acquisition rule queries the pair with the largest estimated overlap. It is a strong operational baseline when proxy magnitude tracks duplicate mass: in the clean-query control of Table 6, Largest-overlap reaches 11.43 MAE with two queries versus 12.61 for EER. Under suppressed high overlap in the same protocol, EER reaches 7.80 versus 9.63 for Largest-overlap. These results identify the decision boundary: overlap magnitude is useful with reliable proxies, while suppression shifts value toward a residual-aware query.

Largest-overlap is strongest when its score tracks the duplicate mass that most affects the current union estimate. EER addresses the suppressed-proxy regime with an observable response surrogate, ranking candidates by expected reduction of the current residual and re-scoring after each response.

The matched comparison with Largest-overlap gives twelve negative paired MAE gaps across four frozen encoders and three degraded strengths under shared matching thresholds. The gap reaches -14.68 MAE for DINOv3 at Strong; on YOLO26m, EER reduces MAE by 11.1%–13.2% across degraded conditions and retains these gains with noisy answers. Synthetic scenes isolate overlap failures, while the static-camera WILDTRACK bridge evaluates their effect on annotation-box appearance proxies with simulated aggregate responses. Together, these experiments connect the sequential update mechanism to query-selection gains across self-supervised and detection-pretrained representations.

DAWSON denotes the full proxy-based union-calibration framework, and EER its query-selection rule. Our contributions are:

- A sparse aggregate weak-supervision formulation for multi-view unique counting, in which a finite budget of pairwise shared-person counts calibrates a set-union estimate from per-view counts and overlap proxies.
- A sequential EER rule that evaluates unrevealed weak labels under an observable response surrogate, ranks their expected calibration effect, and re-scores candidates after every response.
- Matched synthetic and WILDTRACK evidence of query-selection gains under degraded overlap evidence across four frozen encoders, twelve degraded settings, and noisy-answer tests with shared matching thresholds.

2 Related Work

Prior work provides training-stage weak labels, observable cross-view duplicate cues, and principles for spending limited queries. DAWSON operates after per-view counts and overlap proxies are available: a query returns the shared-person cardinality of a selected view pair and calibrates the final cross-view set-union count. The response is an aggregate pair count that can be used directly in fusion.

Counting and duplicate-aware fusion. Single-view counting learns density or count predictors from dense, point, and weaker supervision (Lempitsky & Zisserman, 2010; Zhang et al., 2016; Li et al., 2018; Sindagi & Patel, 2018; Lei et al., 2021; Cohen et al., 2017; Ma et al., 2019; Yang et al., 2020; Zhao et al., 2020; Ranjan et al., 2021). Multi-view methods fuse observations through geometry, correspondences, learned aggregation, BEV, occupancy, and tracking (Chavdarova et al., 2018; Zhang & Chan, 2019; Hou et al., 2020; Song et al., 2021; Hwang et al., 2024; Vora et al., 2023; Daryani et al., 2025; Lee et al., 2023; Fan et al., 2023; Alturki et al., 2025; Liu et al., 2024; Teepe et al., 2024; Woo et al., 2024). DAWSON calibrates the duplicate mass left uncertain after these per-view counts and proxies are available. Its set-union view also connects to distinct-element and set-intersection estimation (Flajolet & Martin, 1985; Flajolet et al., 2007; Broder, 1997), where the repeated entities are latent cross-view objects.

The weak labels in counting literature supervise training of a predictor. DAWSON uses count-level supervision at the final fusion stage: a selected pair's aggregate shared-person response replaces its proxy and calibrates the unique count.

Active aggregate supervision. Active learning spends limited labels using uncertainty, diversity, or expected error reduction (Settles, 2009; Roy & McCallum, 2001; Gal et al., 2017; Yoo & Kweon, 2019; Kirsch et al., 2019; Beluch et al., 2018; Sener & Savarese, 2018). In counting, active selection typically labels images or regions to improve a predictor (Zhao et al., 2020). DAWSON instead queries aggregate shared-person counts under a per-window budget. EER estimates each candidate's effect on an observable residual from a response surrogate fitted on development windows, then re-scores after each held-out response.

Cross-view appearance and association. Person ReID and multi-camera tracking supply appearance, motion, and geometric evidence of repeated targets (Zheng et al., 2015; 2016; Luo et al., 2019; Zhou et al., 2019; He et al., 2021; Wojke et al., 2017; Ye et al., 2022; Milan et al., 2016; Ristani et al., 2016; Tang et al., 2019). DAWSON uses cross-view appearance as an overlap proxy; WILDTRACK provides synchronized overlapping views (Chavdarova et al., 2018), and a frozen DINOv3 representation supplies one tested proxy (Siméoni et al., 2025). Tracking methods resolve identities across cameras and time, whereas DAWSON asks only for a pairwise shared-person total to calibrate the fused count.

3 Method

DAWSON fuses per-view counts and pairwise overlap proxies into a provisional unique count. Each aggregate pair answer replaces one proxy; EER spends a finite per-window budget on the largest expected residual reduction and re-scores after each response.

3.1 Observable Inputs and Provisional Union Estimate

For views $\mathcal{V}$ and windows $\mathcal{T}$ (Zhang & Chan, 2019), $C_{v,t}$ is the observable or predicted count in view v at window t. A synchronized pair $p = ((v_i, t), (v_j, t))$ has duplicate mass O_p, the number counted in both views.

Color-histogram or learned appearance similarity with greedy matching provides overlap proxies; uncertainty and available occlusion/motion cues characterize their reliability. A weak query returns the pair's shared-person cardinality without identity correspondences. Appendix A.1 gives the OT-inspired proxy interpretation and error bound.

For homogeneous visibility multiplicity, a target seen in m views contributes m to the first-order sum and $\binom{m}{2}$ to the pairwise sum, yielding Eq. (2). Heterogeneous multiplicity makes it an approximation; Appendix A.1 bounds the effect of multiplicity and proxy error.

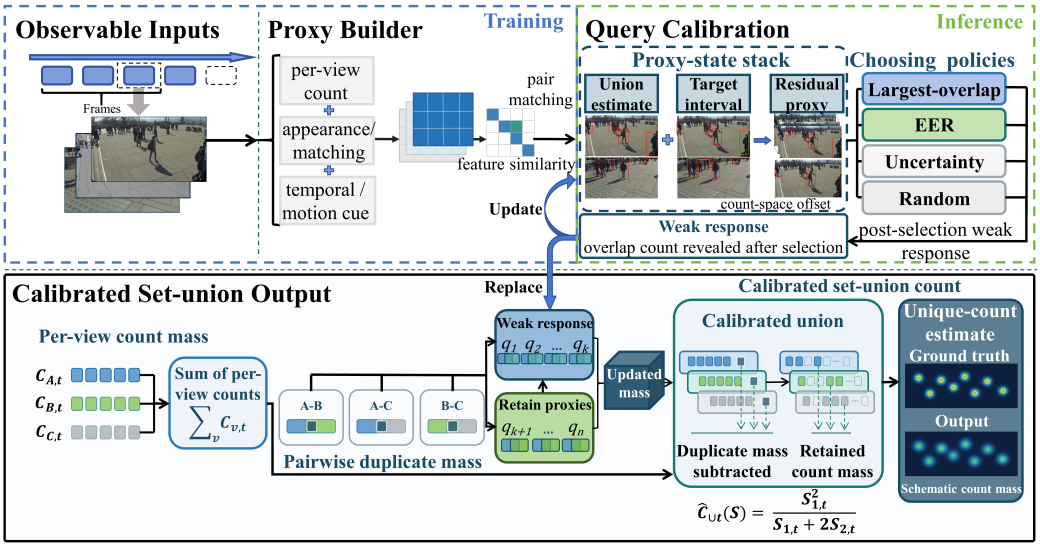

Figure 2: Overview of DAWSON. Appearance features from WILDTRACK annotation boxes build duplicate-mass proxies; aggregate weak answers replace selected proxies to update duplicate mass and the calibrated unique-count estimate, while EER ranks queries by expected fused-count residual reduction.

Acquisition uses counts, overlap estimates $\widehat{O}_p$, reliability features, and earlier responses. WILDTRACK pair answers are annotation-simulated and revealed only after selection; identities, world positions, and the true unique count serve evaluation alone.

3.2 Weak-Query Calibration Loop

For each window t, DAWSON forms the fixed observable count-consistency interval

$$\mathcal{I}_t = [\max_{v \in \mathcal{V}} C_{v,t},\ S_{1,t}],$$

with $\mathcal{I}_t = [0, 0]$ for empty windows. It depends only on per-view counts and remains fixed after queries.

Let $\bar{O}_p(S)$ be the current overlap value used for fusion:

$$\bar{O}_p(S) = \begin{cases} y_p, & p \in S, \\ \widehat{O}_p, & p \notin S, \end{cases} \tag{1}$$

where y_p is the revealed aggregate pair answer. The current union estimate applies the multiplicity-corrected closed form:

$$\widehat{C_{\cup t}}(S) = \frac{S_{1,t}^2}{S_{1,t} + 2\,S_{2,t}}, \qquad S_{1,t} = \sum_{v \in \mathcal{V}} C_{v,t}, \qquad S_{2,t} = \sum_{p \in \mathcal{P}_t} \bar{O}_p(S), \tag{2}$$

where $\mathcal{P}_t$ is the window's candidate set. The estimate is zero for $S_{1,t} = S_{2,t} = 0$ and is otherwise unclipped. A response above its proxy raises $S_{2,t}$ and lowers the estimate; a response below its proxy has the opposite effect. We score this state by distance to the count-consistency interval plus unresolved proxy uncertainty:

$$R_t(S) = d\Big(\widehat{C_{\cup t}}(S), \mathcal{I}_t\Big) + \lambda W_t(S) \sum_{p \in \mathcal{P}_t \setminus S} \sigma_p, \tag{3}$$

where $d(x, [\ell, u]) = \max(\ell - x, 0, x - u)$, $T_t(S) = \sum_{p \in S} y_p$, $W_t(S) = 2S_{1,t}^2/(S_{1,t} + 2T_t(S))^2$ (0 for empty windows), $\lambda = 1$, and σ_p is expected absolute proxy deviation. The shared W_t weights unresolved uncertainty; the true count error is unavailable to acquisition.

For candidate q, Q_q is discrete-Laplace on $y \in \{0, \ldots, c_q\}$, with $c_q = \min(C_{v_i,t}, C_{v_j,t})$, initial proxy $e_q = \widehat{O}_q$, and scale $b_q = c_q(\beta_0 + \beta_a a_q + \beta_o o_q)$ for $\beta = (0.5, 0, 0)$. We fix $Q_q(y) \propto$

$\exp(-|y - e_q|/b_q)$ before acquisition, define $\sigma_q = \sum_y Q_q(y)|y - e_q|$, and use a point mass at zero for $c_q = 0$. Development calibration is separate from held-out windows; the budget counts answers revealed within each held-out window. Appendix A.2 specifies response construction.

Expected union-error reduction (EER) chooses the query with the largest expected decrease in this observable residual:

$$\mathrm{EER}(q \mid S) = R_t(S) - \mathbb{E}_{\tilde{o}_q \sim Q_q}\left[R_t(S \cup \{q\}; \tilde{o}_q)\right]. \tag{4}$$

The expectation enumerates Q_q's support and recomputes Eqs. (2) and (3) for each hypothetical answer, ranking calibration effect rather than proxy magnitude.

After revealing y_q, the loop replaces $\widehat{O}_q$, updates the estimate and residual, and re-scores eligible pairs. Noisy answers are capacity-clamped before insertion; the final prediction is unclipped. Appendix A.2 gives eligibility, ties, and pseudocode.

The common candidate set supports no-query, random, uncertainty, largest-overlap, EER, and a diagnostic hybrid with a frozen gate between EER and Largest-overlap.

3.3 Final Estimate and Proxy Regimes

At the query limit, the loop returns $\widehat{C_{\cup t}}(S)$. Controlled degradation and reliable-proxy tests identify when EER or Largest-overlap is the stronger selector.

4 Experiments and Results

Synthetic scenes isolate set-union correction and suppressed overlap; static-camera WILDTRACK tests image-derived proxies with annotation-simulated pair answers. We compare no-query, random, uncertainty, largest-overlap, EER, and a diagnostic hybrid across query budgets, proxy regimes, answer noise, and encoders. Tables 1 and 2 report the Clean external comparison and Clean/Mild policy ablation.

Our primary readout is paired EER-minus-Largest MAE at matched inputs, estimator, and budget. Table 10 tests four frozen encoders at three degradation strengths; Table 11 adds noisy-answer YOLO26m results. Clean controls, external methods, and zero-query Max give reliable-proxy, operating-regime, and absolute-count references.

Clean uses original annotation-box crops and proxies. Mild adds crop dropout, box jitter, local occlusion, proxy noise, and high-overlap suppression (Figure 3).

4.1 Experimental Setup

The synthetic generator supplies per-view counts, unique counts, pair overlaps, noisy estimates, uncertainty, and motion/occlusion severity under six stress modes. WILDTRACK's seven-view Color and frozen F0 DINOv3 controls use undegraded static-camera observables on held-out windows 200–399. Their source-specific builders define union targets and selected-pair answers. In Color, EER re-scores sequentially, Largest takes a fixed top-k overlap list, and Hybrid audits its native gate.

The F0 control in Table 7 uses frozen DINOv3 ViT-B/16 annotation-crop features (seed 7), thresholds raw $(0.475, 0.450)$ or shifted $(0.7375, 0.725)$ fixed on development windows 160–199, held-out windows 200–399, and budgets 2 and 4 with one deterministic repeat. The separate M3b bridge (Tables 5 and 8) uses thresholds $(0.55, 0.35)$, four strengths, deterministic/noisy answers, ten degradation seeds, 200 windows per condition, and budget 2; its own visibility rules define targets and responses. Formal policy replays are in Tables 10, 13, and 14; zero-query Max and Sum diagnostics appear in Tables 15 and 16.

Formal replay profile. Tables 10, 11, 13, and 14 share the multiplicity estimator, Q with $\beta = (0.5, 0, 0)$, $\lambda = 1$, sequential EER, and `expected_run_metrics_v1`. Metrics first cover each repeat's full window set, then aggregate repeats and degradation seeds. Deterministic runs use 1 repeat per policy except Random's 100; noisy runs use 30 for querying policies, 100 for Random, and 1 for No-query. M3b retains its source-specific contract; F0 uses one deterministic repeat.

Table 1: WILDTRACK Clean unique-count comparison. DAWSON uses crop/proxies; external methods use native inputs. Bias is signed and ranked by magnitude.

	Count error				Count agreement		
Method	MAE ↓	RMSE ↓	MAPE (%) ↓	Bias ↓	CP ↑	CR ↑	CF1 ↑
DAWSON (ours), Clean							
No-query	1.60	2.00	6.95	**-0.20**	0.966	**0.968**	0.966
Random	1.54	1.93	6.58	-0.37	0.972	0.966	0.967
Uncertainty	**1.51**	**1.90**	**6.34**	-0.46	**0.975**	0.964	**0.968**
Largest-overlap	1.67	2.07	7.10	-0.45	0.971	0.962	0.964
EER	1.75	2.13	7.34	-0.62	0.973	0.957	0.963
Strong-information setting (clean)							
CaMuViD (Daryani et al., 2025)	7.01	7.78	32.11	+6.95	0.772	**0.999**	0.867
TrackTacular (Teepe et al., 2024)	2.38	3.26	9.06	-2.26	0.997	0.912	0.951
Booster-SHOT (Hwang et al., 2024)	**1.92**	**2.71**	**7.30**	**-1.81**	**0.998**	0.929	**0.961**

Frozen encoders. The bridge freezes DINOv3 ViT-B/16, DINOv2 ViT-B/14, YOLO11m, and YOLO26m (Siméoni et al., 2025; Oquab et al., 2023; Jocher et al., 2024; 2026), transfers the same M1b thresholds (raw $(0.475, 0.450)$; shifted $(0.7375, 0.725)$), and evaluates windows 200–399. YOLO26m uses official `yolo26m.pt` layers 0–9 and 512-dimensional pooled features from 640×640 letterboxed annotation crops, with development-crop mean centering before normalization as for YOLO11m. Appendix B.2 details feature extraction and paired seeds.

Table 6 mechanism protocol. This independent noisy synthetic-answer test uses the sequential core, four views, 10,000 scenes per condition/seed, ten scenario seeds, and budgets 1/2/3. Random uses 100 repeats; other policies use one. Metrics aggregate complete-scene repeats before scenario seeds.

Cache comparisons are point estimates; cross-encoder paired confidence intervals use degradation seeds as their uncertainty unit.

Independent Fig. 4 diagnostic. The independent five-window WILDTRACK diagnostic (windows 360–364) uses frozen F0 DINOv3 (seed 7), strict annotation-box visibility, two queries, and a response surrogate calibrated on development windows 160–199. Booster-SHOT uses native raw-frame/calibration inputs; DAWSON uses annotation crops and simulated aggregate answers. Appendix A.3 defines this diagnostic.

4.2 WILDTRACK Clean External Comparison

Table 1 compares Clean DAWSON with CaMuViD (Daryani et al., 2025), TrackTacular (Teepe et al., 2024), and Booster-SHOT (Hwang et al., 2024) on matched timestamps under native inputs. The external training splits are unverified for CaMuViD/TrackTacular, and the bound Booster-SHOT checkpoint overlaps 160/200 evaluation timestamps, so this is an operating-regime reference. In the matched five-policy ablation (Table 2), EER has the lowest Mild MAE, 12.76 versus Largest-overlap's 13.78 (7.3% lower before rounding). Table 6 isolates selection when high overlap is suppressed: at two queries, EER reaches 7.80 MAE versus 9.63 for Largest-overlap. Clean and no-suppression controls favor Largest-overlap, as does the three-query high-overlap condition, establishing a budget- and proxy-dependent selection effect.

4.3 MultiviewX Held-Out Comparison

Table 3 compares Sequential EER with TrackTacular and CaMuViD + count adapter on clean held-out frames 0360–0399. TrackTacular uses stronger bird's-eye-view spatial information and has the lowest MAE (1.33); EER uses aggregate weak queries and reaches 10.03 MAE. The table provides an input-regime comparison rather than a matched policy ranking.



Figure 3: Clean and Mild crop examples with box jitter, local occlusion, and crop dropout. The schematic shows subsequent overlap-proxy perturbations.

Table 2: DAWSON WILDTRACK policy ablation under Clean and Mild crop-and-proxy degradation; Bias is ranked by magnitude.

	Count error				Count agreement		
Policy	MAE ↓	RMSE ↓	MAPE (%) ↓	Bias ↓	CP ↑	CR ↑	CF1 ↑
Clean							
No-query	1.60	2.00	6.95	**-0.20**	0.966	**0.968**	0.966
Random	1.54	1.93	6.58	-0.37	0.972	0.966	0.967
Uncertainty	**1.51**	**1.90**	**6.34**	-0.46	**0.975**	0.964	**0.968**
Largest-overlap	1.67	2.07	7.10	-0.45	0.971	0.962	0.964
EER	1.75	2.13	7.34	-0.62	0.973	0.957	0.963
Mild							
No-query	17.18	17.75	74.34	+17.18	0.584	**1.000**	0.735
Random	13.97	14.50	60.43	+13.97	0.633	**1.000**	0.772
Uncertainty	14.11	14.63	60.65	+14.11	0.631	**1.000**	0.771
Largest-overlap	13.78	14.26	60.02	+13.78	0.635	**1.000**	0.774
EER	**12.76**	**13.18**	**55.79**	**+12.76**	**0.651**	**1.000**	**0.786**

4.4 WILDTRACK Proxy Validation and Controlled Degradation

WILDTRACK provides a static-camera image-proxy bridge: annotation-box crops yield overlap estimates, and selected pairs receive annotation-simulated shared-person counts. Tables 1 and 2 report the Clean external comparison and Clean/Mild policy ablation. External methods retain their native raw-frame and world-plane inputs; DAWSON uses per-view counts and crop-based proxies. EER has the lowest policy MAE under Mild degradation (12.76 versus 13.78 for Largest-overlap), while Clean tests reliable proxy evidence.

In the separate Color protocol (Table 4), EER-selected answers reduce MAE from 5.06 without queries to 4.36 and 3.33 at budgets two and four. Largest-overlap reaches 4.55 and 3.76 at those

Table 3: Clean MultiviewX held-out unique-count comparison (frames 0360–0399). EER is Sequential; CaMuViD uses a calibrated count adapter. Bias is signed and ranked by magnitude.

	Count error				Count agreement		
Method	MAE ↓	RMSE ↓	MAPE (%) ↓	Bias ↓	CP ↑	CR ↑	CF1 ↑
EER	10.03	10.12	26.82	-10.03	**1.000**	0.732	0.845
TrackTacular (Teepe et al., 2024)	**1.33**	**1.86**	**3.55**	**-1.23**	0.999	0.966	**0.982**
CaMuViD + count adapter (Daryani et al., 2025)	15.88	16.21	42.71	15.88	0.704	**1.000**	0.825

Table 4: Deterministic WILDTRACK color-proxy robustness: policy MAE and EER-minus-Largest Δ (negative favors EER).

	Policy MAE ↓			EER (ours) ↓			EER − Largest
Weak Queries	Largest	EER	Hybrid	RMSE	MAPE (%)	Bias	Δ
0	**5.06**	**5.06**	**5.06**	5.44	23.48	5.06	0.00
1	4.75	**4.72**	4.75	5.12	21.90	4.72	−0.03
2	4.55	**4.36**	4.44	4.78	20.38	4.36	−0.19
3	4.19	**3.93**	4.04	4.37	18.59	3.92	−0.26
4	3.76	**3.33**	3.55	3.84	16.02	3.30	−0.43

budgets. The fixed M1b F0/F1/F2 diagnostic uses another source family; its complete results are in Appendix C.

The M3b bridge tests crop dropout, partial visibility, correlated proxy noise, and suppressed high-overlap estimates while retaining WILDTRACK's static cameras and annotation-simulated responses. Its source-specific visibility rules determine the surviving union target and pair intersections; the full builder and acquisition protocol is in Appendix B.1.

Independent sequential EER diagnostic. Figure 4 shows three preselected windows from the separate five-frame F0 diagnostic. In window 362, the second response updates overlap from 14 to 19 and the estimate from 21.34 to 20.86, illustrating sequential state calibration. Projection and matching rules and all five traces are in Appendix A.3.

Under Strong noisy M3b degradation, EER reaches 31.38 MAE versus 43.16 for Largest-overlap at two queries (paired $\Delta = -11.77$, 95% CI $[-13.18, -10.36]$). Table 5 reports all six policies over 200 held-out windows and ten degradation seeds; Appendix B.1 gives the paired-seed protocol and intervals.

4.5 Ablation and Proxy-Regime Analysis

Table 6 isolates the budget-dependent mechanism: EER leads Largest-overlap under high-overlap suppression at two queries (7.80 versus 9.63 MAE), while Largest leads at three (6.72 versus 8.48) and in the clean-query control.

Table 7 adds undegraded Color and separately sourced frozen F0 DINOv3 controls. In F0, Largest-overlap leads at budgets two and four; Tables 5 and 8 give the M3b degradation comparison.

The policy replay compares No-query, Random, Uncertainty, Largest, EER, and the Hybrid diagnostic. Max and Sum are separate zero-query count diagnostics (Tables 15 and 16); the development switching analysis is in Table 17.

Cross-encoder generalization. Four frozen encoders (DINOv3, DINOv2, YOLO11m, YOLO26m) share M1b thresholds selected with DINOv3. Appendix Figure 5 shows the paired EER-minus-Largest gaps under fixed-threshold transfer.

All twelve degraded-condition paired gaps and 95% confidence intervals favor EER. The Strong gap reaches -14.68 MAE for DINOv3; YOLO26m shows 12.8%–13.2% MAE reductions across degraded strengths, retained with noisy answers (Table 11). The Clean DINOv3 control favors Largest-overlap by 0.08 MAE. Appendix B.2 reports complete metrics and intervals.

Table 5: DINOv3 WILDTRACK degradation bridge with two queries; entries show MAE and policy-minus-Largest differences.

Strength	Mode	No-query	Largest	EER	Random	Uncertainty	Hybrid
Clean	Det.	3.07	3.06	3.06 (+0.00)	**2.96** (-0.11)	2.99 (-0.07)	3.06 (+0.00)
Clean	Noisy	3.07	3.02	3.02 (+0.00)	**2.92** (-0.10)	2.96 (-0.06)	3.02 (+0.00)
Mild	Det.	3.22	3.20	3.21 (+0.01)	**3.10** (-0.10)	3.12 (-0.08)	3.21 (+0.01)
Mild	Noisy	3.22	3.17	3.18 (+0.01)	**3.07** (-0.10)	3.09 (-0.08)	3.18 (+0.01)
Medium	Det.	3.84	3.55	3.31 (-0.24)	3.27 (-0.28)	**3.21** (-0.34)	3.46 (-0.09)
Medium	Noisy	3.84	3.58	3.33 (-0.25)	3.29 (-0.29)	**3.23** (-0.35)	3.49 (-0.09)
Strong	Det.	53.41	42.90	**31.26** (-11.64)	38.15 (-4.75)	38.78 (-4.12)	36.53 (-6.37)
Strong	Noisy	53.41	43.16	**31.38** (-11.77)	38.33 (-4.83)	38.98 (-4.18)	36.72 (-6.43)

Table 6: Sequential EER occlusion-mechanism study; L/E are Largest/EER MAE and Δ is paired EER$-$Largest (negative favors EER).

		1 query			2 queries			3 queries		
Condition	Strength	L	E	Δ	L	E	Δ	L	E	Δ
Overall / main stress results										
Clean query	0.8	**14.78**	14.95	+0.17	**11.43**	12.61	+1.18	**8.96**	12.93	+3.97
Query occluded	1.0	15.35	**15.23**	-0.12	**12.30**	12.32	+0.02	**9.57**	11.27	+1.69
Ablation / diagnostic slices										
No estimate suppression	0.8	**14.69**	15.24	+0.55	**11.62**	13.40	+1.78	**9.74**	13.74	+4.00
Low-overlap mix	0.8	**17.08**	18.21	+1.13	**13.19**	16.25	+3.06	**10.71**	16.13	+5.42
High-overlap mix	1.0	12.59	**11.16**	-1.43	9.63	**7.80**	-1.83	**6.72**	8.48	+1.77

Table 7: Color and frozen DINOv3 F0 proxy calibration on 200 held-out windows at budgets 2 and 4 ; Δ is EER minus Largest MAE.

Proxy	Weak Queries	Largest	EER	Δ	Hybrid
Color	2	4.55	**4.36**	-0.19	4.44
Color	4	3.76	**3.33**	-0.43	3.55
DINOv3	2	**1.67**	1.75	+0.08	1.75
DINOv3	4	**1.68**	1.70	+0.03	1.70

5 Conclusion

DAWSON calibrates multi-view unique counts with a small budget of aggregate shared-person answers. EER selects each query by its expected effect on an observable calibration residual, then updates the estimate and re-scores the remaining pairs. Across four frozen encoders with shared matching thresholds, this sequential rule reduces MAE relative to Largest-overlap in all twelve degraded settings. YOLO26m retains the gain with noisy answers, supporting transfer across both feature models and response conditions. Synthetic and reliable-proxy controls link the choice of selector to overlap quality. Detector-produced inputs, moving-camera benchmarks, and higher-order union correction are natural next tests of the calibration framework.

AI USE STATEMENT

Generative AI tools assisted with research-methodology and experiment feedback, implementation and verification code, experimental-result interpretation, and manuscript editing and formatting. The authors reviewed and verified all AI-assisted work, including experimental configurations, source data, numerical results, text, code, figures, and other artifacts, and take responsibility for the final content of this submission.

REFERENCES

Reef Alturki, Adrian Hilton, and Jean-Yves Guillemaut. Enhanced multi-view pedestrian detection using probabilistic occupancy volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9368–9377, June 2018.

Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997 (SEQUEN)*, pp. 21–29, 1997.

Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5030–5039, June 2018.

Joseph Paul Cohen, Genevieve Boucher, Craig A. Glastonbury, Henry Z. Lo, and Yoshua Bengio. Count-ception: Counting by fully convolutional redundant counting. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

Amir Etefaghi Daryani, M. Usman Maqbool Bhutta, Byron Hernandez, and Henry Medeiros. Camuvid: Calibration-free multi-view detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Siqi Fan, Zhe Wang, Xiaoliang Huo, Yan Wang, and Jingjing Liu. Calibration-free bev representation for infrastructure perception. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences (JCSS)*, 31(2):182–209, 1985.

Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Proceedings of the 2007 Conference on Analysis of Algorithms (AofA)*, pp. 127–146, 2007.

Francois Fleuret, Joan Berclaz, Regine Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2), 2008.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15013–15022, 2021. URL `https://openaccess.thecvf.com/content/ICCV2021/html/He_TransReID_Transformer-Based_Object_Re-Identification_ICCV_2021_paper.html`.

Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–18, 2020.

Jinwoo Hwang, Philipp Benz, and Pete Kim. Booster-shot: Boosting stacked homography transformations for multiview pedestrian detection with attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, 2024. URL `https://github.com/ultralytics/ultralytics`.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019.

Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. Multi-view target transformation for pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023.

Yinjie Lei, Yan Liu, Pingping Zhang, and Lingqiao Liu. Towards using count-level weak supervision for crowd counting. *Pattern Recognition (PR)*, 109:107616, 2021.

Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pp. 1324–1332, 2010.

Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. Multi-view attentive contextualization for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1487–1495, June 2019.

Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL `https://arxiv.org/abs/2304.07193`.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016 Workshops (ECCVW)*, volume 9914 of *Lecture Notes in Computer Science*, pp. 17–35, Cham, 2016. Springer International Publishing.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pp. 441–448, 2001.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Vishwanath A. Sindagi and Vishal M. Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters (PRL)*, 107:3–16, 2018.

Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8797–8806, June 2019.

Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Lifting multi-view detection and tracking to the bird's eye view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Bringing generalization to deep multi-view pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017.

Sanghyun Woo, Kwanyong Park, Inkyu Shin, Myungchul Kim, and In So Kweon. Mtmmc: A large-scale real-world multi-modal camera tracking benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–17, 2020.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):2872–2893, 2022.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Qi Zhang and Antoni B. Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8297–8306, June 2019.

Qi Zhang and Antoni B. Chan. Calibration-free multi-view crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 227–244, 2022.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 589–597, June 2016.

Zhen Zhao, Miaojing Shi, Xiaoxiao Zhao, and Li Li. Active crowd counting with limited supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 565–581, 2020.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision – ECCV 2016 (ECCV)*, Lecture Notes in Computer Science, 2016.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

A Appendix: Detailed Acquisition Protocol and Supporting Diagnostics

For each temporal window, eligible weak queries are unqueried pairs with positive capacity $\mathrm{cap}_q = \min(C_{v_i,t}, C_{v_j,t})$. The core accepts only an integer response in $[0, \mathrm{cap}_q]$; the noisy evaluator may capacity-clamp its generated response before insertion, whereas the final prediction is not clipped. The surrogate Q_q is initialized from the observable pair proxy and then frozen. Ties are resolved deterministically by exact score followed by the lexicographically largest canonical pair key. These details do not introduce evaluation-only targets into scoring: the annotation-simulated response is revealed only after selection, and true pair overlap, global identity, world position, and true unique count remain unavailable at decision time.

A.1 Observable Proxy Construction and Error Characterization

For a view-window (v, t), let

$$\mu_{v,t} = \sum_{a=1}^{n_{v,t}} m_a \delta_{x_a} \tag{5}$$

be an empirical measure over observable detections, crops, or tracklet fragments. Here, $n_{v,t}$ is the number of observable items in view-window (v, t), x_a is the feature of item a, m_a is its count mass, and δ_{x_a} is a unit point mass at feature x_a. For two view-windows, a partial cross-view coupling

$$\Pi_{ij,t} = \{\pi \geq 0 : \pi \mathbf{1} \leq m^{(i)},\ \pi^\top \mathbf{1} \leq m^{(j)}\} \tag{6}$$

represents feasible duplicate assignments without forcing every detection in one view to match another. In this expression, π is a nonnegative pair-assignment matrix, $\mathbf{1}$ is an all-ones vector of the appropriate dimension, and $m^{(i)}$ and $m^{(j)}$ are the vectors of item masses for the two view-windows. A transport-style proxy is:

$$\begin{aligned} \widehat{\pi}_{ij,t} &\in \operatorname*{arg\,min}_{\pi \in \Pi_{ij,t}} \langle C_{ij,t}, \pi \rangle + \varepsilon \Omega(\pi) + \tau U(\pi), \\ \widehat{O}_{ij,t} &= \langle \widehat{\pi}_{ij,t}, \mathbf{1}\mathbf{1}^\top \rangle, \end{aligned} \tag{7}$$

where $C_{ij,t}$ is the pairwise cost matrix built from decision-time appearance, temporal/geometric, confidence, occlusion, and motion features; ε weights the regularizer $\Omega(\pi)$, Ω is an optional coupling regularizer, τ weights the unbalanced-mass penalty $U(\pi)$, and U measures unmatched mass or mass imbalance permitted by the partial coupling. The brackets $\langle \cdot, \cdot \rangle$ denote the matrix inner product, and $\widehat{\pi}_{ij,t}$ is an optimal feasible coupling. The resulting $\widehat{O}_{ij,t}$ is the estimated duplicate mass, obtained by summing all entries of that coupling. This is an OT-inspired modeling view consistent with occupancy-map and association-metric tracking formulations (Fleuret et al., 2008; Wojke et al., 2017).

Let $C_t^\star$ be the true unique count, $O_p^\star$ the true duplicate mass for pair p, and $H_t = S_{3,t}^\star - S_{4,t}^\star + \cdots$ the omitted higher-order inclusion-exclusion terms, where $S_{k,t}^\star$ is the true k-th order intersection sum (the total number of targets visible in all of a k-view subset, summed over all k-view subsets). Under the condition that $S_{1,t}$ is the accurate first-order count sum, let $S_{1,t}$ denote that sum (Eq. (2)), so $C_t^\star = S_{1,t} - \sum_p O_p^\star + H_t$. The linear pairwise truncation $\widehat{C_{\cup t}}^{\mathrm{lin}}(S) = S_{1,t} - \sum_p \bar{O}_p(S)$ obeys

$$\left| \widehat{C_{\cup t}}^{\mathrm{lin}}(S) - C_t^\star \right| \leq \sum_{p \in \mathcal{P}_t} \left| \bar{O}_p(S) - O_p^\star \right| + |H_t|, \tag{8}$$

where the first term is the duplicate-mass proxy error and $|H_t|$ is the higher-order residual. The $|H_t|$ term is what makes the linear truncation fail when many targets span three or more views. The multiplicity-corrected closed form in Eq. (2) absorbs this term under homogeneous multiplicity. Let m denote the number of views in which a target is visible, with mean $\bar{m}$ and variance $\mathrm{Var}[m]$ over targets in window t. With heterogeneous multiplicity, Eq. (2) estimates $C_t^\star\, \bar{m}^2 / \mathbb{E}[m^2] = C_t^\star / (1 + \mathrm{Var}[m]/\bar{m}^2)$, carrying a relative bias of $-r/(1+r)$ for $r = \mathrm{Var}[m]/\bar{m}^2$; $-r$ is its first-order approximation when r is small. Both the proxy error and the multiplicity heterogeneity are therefore the calibration targets.

Algorithm 1 DAWSON weak-query acquisition loop

Input: per-view counts $\{C_{v,t}\}$, pair proxy statistics $\{\widehat{O}_p, \sigma_p, z_p\}_{p\in\mathcal{P}_t}$, number of queries B

$S \leftarrow \emptyset$; initialize fixed $\mathcal{I}_t$ and $\widehat{C_{\cup t}}(S)$
for $b \leftarrow 1, \ldots, B$ **do**
  $E \leftarrow \{q \in \mathcal{P}_t \setminus S : \mathrm{cap}_q > 0\}$
  if $E = \emptyset$ **then**
    break
  end if
  for all $q \in E$ **do**
    Retrieve frozen Q_q and enumerate $y = 0, \ldots, \mathrm{cap}_q$
    $\mathrm{score}(q) \leftarrow \mathrm{EER}(q \mid S)$
  end for
  $q^\star \leftarrow \arg\max_{q\in E}(\mathrm{score}(q), \mathrm{key}(q))$
  $y_{q^\star} \leftarrow$ RevealAndValidateResponse$(q^\star)$
  $S \leftarrow S \cup \{q^\star\}$; $\bar{O}_{q^\star}(S) \leftarrow y_{q^\star}$
  Update $\widehat{C_{\cup t}}(S)$, $T_t(S)$, $W_t(S)$, and $R_t(S)$; keep $\mathcal{I}_t$ fixed
end for
Return: $\widehat{C_{\cup t}}(S)$

A.2 Surrogate Query Responses

For a candidate pair q, the response surrogate is the finite discrete-Laplace family used by the formal sequential core:

$$Q_q(y) = \begin{cases} 1, & \mathrm{cap}_q = 0,\ y = 0, \\ \dfrac{\exp(-|y - e_q|/b_q)}{\sum_{z=0}^{\mathrm{cap}_q} \exp(-|z - e_q|/b_q)}, & \mathrm{cap}_q > 0,\ y \in \{0, \ldots, \mathrm{cap}_q\}, \end{cases} \tag{9}$$

where $e_q = \widehat{O}_q$, $b_q = \mathrm{cap}_q(\beta_0 + \beta_a a_q + \beta_o o_q)$, and $(\beta_0, \beta_a, \beta_o) = (0.5, 0, 0)$. The distribution is initialized once from the initial proxy and held fixed; its uncertainty is

$$\sigma_q = \sum_{y=0}^{\mathrm{cap}_q} Q_q(y)|y - e_q|, \tag{10}$$

an expected absolute deviation from that initial proxy, not a standard deviation. The implementation derives Q_q only from decision-time proxy statistics and never from a test unique-count oracle.

A.3 Fig. 4 Sequential EER Candidate

This independent diagnostic instance instantiates the common sequential definitions in Eqs. (2)–(4) and Algorithm 1. The scope below fixes only this figure's inputs, feature construction, calibration, and trace collection; it neither changes the common U, fixed interval, finite-Q, σ, residual, or re-scoring definitions nor retrospectively validates the sequential implementations behind other tables.

Inputs and scope. Each timestamp has seven views and all 21 unordered pairs. A view entry is visible exactly when $x_{\min} \geq 0$, $y_{\min} \geq 0$, $x_{\max} > x_{\min}$, and $y_{\max} > y_{\min}$. Counts, anonymous crops, frozen F0 DINOv3 features, hard matched-count proxies, and simulated responses share this scope. Write C_v for a view count, $c_p = \min(C_i, C_j)$ for pair capacity, and $e_p = \widehat{O}_p \in \{0, \ldots, c_p\}$ for its hard proxy. Similarity-score sums are not duplicate counts. The frozen shifted-cosine matching and weak thresholds are 0.7375 and 0.725; these are transferred from the earlier clipped-bbox development calibration, not reselected on the five frames.

Figure-specific response inputs. Suppressing the timestamp index, let $S_1 = \sum_v C_v$ and $S_2(S) = \sum_{p\in S} y_p + \sum_{p\notin S} e_p$. The count update and fixed interval are the common definitions above; in this instance $g(x) = S_1^2/(S_1 + 2x)$ and the common finite response family uses the figure-specific proxy e_p and features below. Responses are validated by the core, and the final estimate is not clipped. Here a_p is the existing ambiguity feature: for each left-hand similarity row, average the two indicators for a top score above the weak threshold with top-two gap below 0.08, and for a top score below the weak threshold, then average over rows. Thus $a_p \leq 0.5$ for positive-capacity pairs; it is

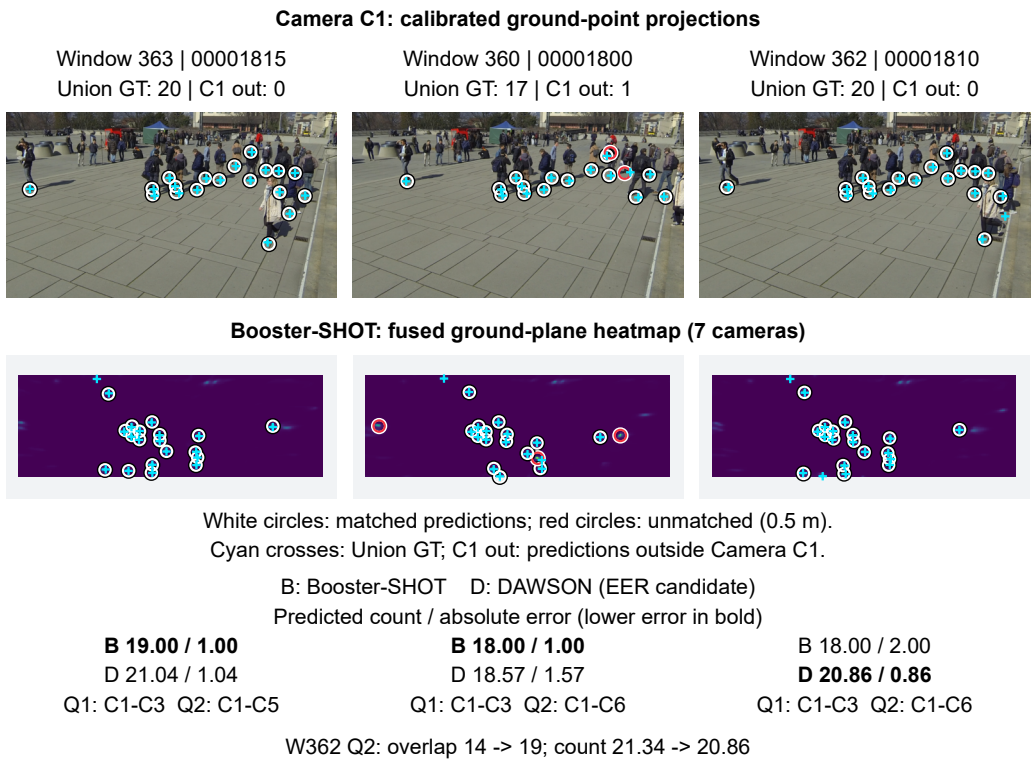

Figure 4: Fixed WILDTRACK cases: calibrated Booster-SHOT projections in Camera C1 (top) and seven-camera BEV predictions (bottom), with Union GT in cyan. DAWSON's count estimates and sequential query traces appear below.

not rescaled. The pair feature h_p averages the two views' occlusion proxies, each equal to half the mean maximum bbox IoU plus half the mean truncation fraction. The latter is the fraction of the four bbox boundaries within two pixels of their corresponding image boundary. Both features are dimensionless. The location e_p is Q's mode, not generally its mean on the bounded support. Sigma has pair-count units and is an absolute deviation, not a standard deviation. Zero-capacity pairs are ineligible and retain the common point-mass convention.

Development-only calibration. The 252 tuples were frozen before accessing development responses: $\beta_0 \in \{1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1\}$ and $\beta_a, \beta_o \in \{0, 1/16, 1/8, 1/4, 1/2, 1\}$. On windows 160–199 (frames 00000800–00000995), observable artifacts were frozen before an independent identity-based process generated strict pair overlaps. For positive-capacity pairs P_t^+ and nonempty windows D^+, calibration minimized the window-then-pair negative log likelihood

$$J(\beta) = \frac{1}{|D^+|} \sum_{t \in D^+} \frac{1}{|P_t^+|} \sum_{p \in P_t^+} -\log Q_{t,p,\beta}(y_{t,p}^{\mathrm{dev}}).$$

Equal represented objectives use ascending lexicographic beta order. The selected tuple is $(\beta_0, \beta_a, \beta_o) = (0.5, 0, 0)$, hence $b_p = 0.5 c_p$. Its zero slopes do not support an ambiguity- or occlusion-adaptive claim for this fitted configuration. Calibration uses no query outcomes or global-count errors from windows 360–364.

Residual and exact response expectation. Set $T(S) = \sum_{p \in S} y_p$, $B_\sigma(S) = \sum_{p \notin S} \sigma_p$, and instantiate the common Eq. (3) with $\lambda = 1$ and

$$W(S) = \frac{2S_1^2}{(S_1 + 2T(S))^2}, \qquad R(S) = d(g(S_2(S)), I) + W(S) B_\sigma(S).$$

For $S_1 > 0$, $0 < W \leq 2$ bounds the magnitude of g' on $x \geq T(S)$, converting pair-count deviation into count sensitivity. Lambda is a fixed, dimensionless normalization, not a fitted parameter. For a

hypothetical answer y to unqueried q, Eq. (4) is evaluated by the full finite sum using

$$S_2' = S_2 - e_q + y, \qquad T' = T + y, \qquad B_\sigma' = B_\sigma - \sigma_q,$$
$$W' = \frac{2S_1^2}{(S_1 + 2T')^2}, \qquad R_{q,y} = d(g(S_2'), I) + W' B_\sigma',$$
$$A(q \mid S) = R(S) - \sum_{y=0}^{c_q} Q_q(y) R_{q,y}.$$

Each hypothetical branch is an independent copy of the parent state and makes no call to the response service. After an actual response, S_2, T, W, B_σ and R update; counts, capacities, interval, and unresolved Q_p, σ_p remain fixed. All remaining candidate scores are then recomputed.

Sequential sealed execution. Eligible pairs are unqueried pairs of positive capacity. With seed 7 recorded and budget two, each step enumerates all eligible candidates, then breaks equal represented scores by descending zero-based string pair key i-j. It persists all candidate scores, Q supports/probabilities, hypothetical residuals and the selected choice, and verifies that choice's hash before a separate service returns the selected pair's aggregate strict overlap. The service supplies no identities, unselected labels or global GT to the scorer. Choice and response events reference the previous event's hash. The first response updates the state before the second scoring pass. Execution stops at the budget or when no eligible pair remains; there is no extra score threshold. Only after all five final predictions are frozen does independent evaluation read global GT. Arithmetic is float64 with finite log-sum-exp normalization; ties use no epsilon. Synthetic tests use 10^{-10} absolute/relative tolerance and 10^{-12} probability-normalization tolerance, not tunable method parameters.

Interpretation of the diagnostic. Writing $d = d(g(S_2), I)$, the score decomposes as

$$A(q \mid S) = d - \mathbb{E}d' + W\sigma_q + (W - \mathbb{E}W')(B_\sigma - \sigma_q).$$

The last term tightens a common sensitivity bound; it conveys no new labels to other pairs and does not reduce their stored Q or sigma. It contributes 67.4%–78.3% of the ten selected scores. Under the fixed interval, nonnegative responses and fixed unresolved sigma, the derivative bound gives $\mathbb{E}_{Q_q} R' \le R$. This concerns the specified surrogate, not true expected count error. A realized response may increase count error even as R falls, as in the first query of window 364. Heterogeneous visibility still biases Eq. (2) with exact pair answers; $R = 0$ need not imply zero count error. Development endpoint mismatch and the marginal, rather than joint, nature of Q further limit its probabilistic interpretation. The five previously viewed timestamps support this local mechanism diagnostic, not a generalization claim or retrospective validation of the main tables.

B Proxy Calibration Diagnostics

These diagnostics collect the proxy-regime evidence for DAWSON's weak-query calibration. Frozen DINOv3 F0 marks the reliable-proxy control, the M3b bridge varies degradation, and the cross-encoder replay checks the same pattern with other encoders. The protocols remain separate so each selector comparison keeps its input and evaluation identity.

B.1 Frozen DINOv3 Controls and Degradation Diagnostics

With two queries, the audited Color proxy gives EER an MAE of 4.36 versus 4.55 for Largest-overlap. In the frozen F0 control, Largest-overlap has lower MAE than EER at budgets 2 and 4 (1.67/1.68 versus 1.75/1.70). EER targets pairs with high ambiguity or expected residual effect, directing queries toward reducing the fused-count residual under a noisy proxy.

The DINOv3 rows in Table 7 use frozen F0 features (model seed 7), M1b matching thresholds raw $(0.475, 0.450)$ (shifted $(0.7375, 0.725)$), and 200 held-out windows (200–399), with one deterministic repeat per policy and budget. Native annotations and the F0 builder's crop-visibility rule define the union targets and queried intersections. We verify the saved two-query replay and execute a separate four-query replay, retaining complete predictions and query traces. Hybrid uses the original

Table 8: Six-policy DINOv3 replay: MAE and paired EER-minus-Largest Δ with 95% CIs.

		(a) Policy MAE ↓						(b) EER − Largest	
Strength	Mode	No-query	Largest	EER	Random	Uncertainty	Hybrid	Δ	95% CI
Clean	Det.	3.07	3.06	3.06	**2.96**	2.99	3.06	+0.00	—
Clean	Noisy	3.07	3.02	3.02	**2.92**	2.96	3.02	+0.00	—
Mild	Det.	3.22	3.20	3.21	**3.10**	3.12	3.21	+0.01	[+0.01,+0.01]
Mild	Noisy	3.22	3.17	3.18	**3.07**	3.09	3.18	+0.01	[+0.01,+0.01]
Medium	Det.	3.84	3.55	3.31	3.27	**3.21**	3.46	-0.24	[-0.73,+0.24]
Medium	Noisy	3.84	3.58	3.33	3.29	**3.23**	3.49	-0.25	[-0.74,+0.24]
Strong	Det.	53.41	42.90	**31.26**	38.15	38.78	36.53	-11.64	[-13.05,-10.24]
Strong	Noisy	53.41	43.16	**31.38**	38.33	38.98	36.72	-11.77	[-13.18,-10.36]

fixed observable gate and independently executes EER in all 200 windows at each budget. These are single-control MAEs; differences and rankings use unrounded values.

Table 8 reports the same complete six-policy estimates as Table 5, from the accepted isolated replay of the frozen M3b inputs. This is separate from the Table 7 F0 DINOv3 control, which uses one frozen seed-7 input identity, no bridge degradation, and one deterministic repeat at budgets 2 and 4. For policy p and seed s, let $\mathrm{MAE}_{p,s}$ denote the mean of the complete 200-window repeat MAEs. We form the paired difference $d_s = \mathrm{MAE}_{p,s} - \mathrm{MAE}_{\mathrm{Largest},s}$ before rounding and report $\bar{d} \pm t_{0.975,9} s_d/\sqrt{10}$, where s_d is the sample standard deviation across the ten degradation seeds. Clean reuses one distinct observable input across the ten condition labels and is reported as a point estimate; these labels retain their configured acquisition streams. Random uses 100 repeats per condition, the other querying policies use one deterministic or 30 noisy repeats, and No-query uses one. Hybrid applies the recovered M3b development-quantile gate and executes the selected EER or Largest branch with that branch's response RNG.

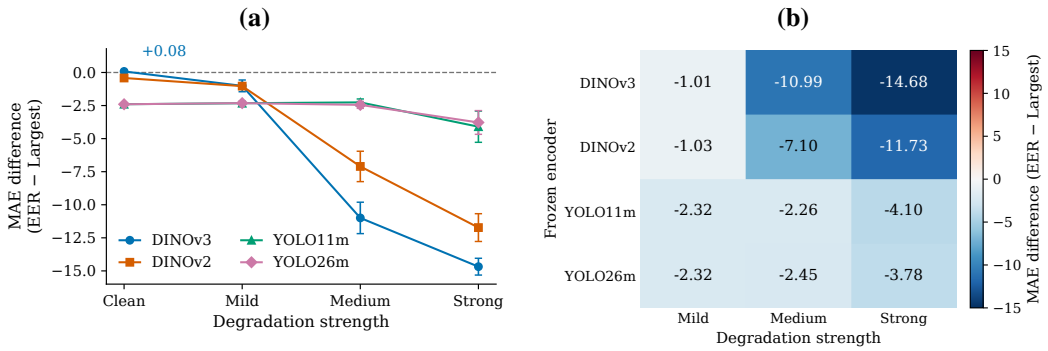

Figure 5: EER − Largest MAE across four encoders at two queries: (a) paired means and 95% CIs over ten degradation seeds (Clean: point estimate); (b) the 4×3 degraded-condition matrix. Negative values favor EER.

B.2 Cross-encoder generalization

Table 10 reports the deterministic replay for DINOv3, DINOv2, YOLO11m, and YOLO26m behind Figure 5(a). All four use the same M1b matching thresholds, shared multiplicity estimator, and sequential EER core at a two-query budget. Their twelve degraded encoder–strength pairs have negative EER-minus-Largest MAE differences and paired 95% intervals below zero. The table also reports the same-input zero-query Max baseline and all Clean rows.

Figure 5(b) and Table 9 summarize the four-encoder, three-strength matrix. Complete policy metrics and confidence intervals follow in Table 10.

Table 9: Paired MAE difference (EER − Largest); negative values favor EER.

Encoder	Mild	Medium	Strong
DINOv3	-1.01	-10.99	-14.68
DINOv2	-1.03	-7.10	-11.73
YOLO11m	-2.32	-2.26	-4.10
YOLO26m	-2.32	-2.45	-3.78

Table 10: Deterministic replay across four encoders at two queries. Policy MAE and paired EER − Largest differences use ten seeds; Clean uses point estimates.

Encoder	Strength	No-q	Rand.	Unc.	Largest	EER	RMSE	MAPE (%)	Bias	Δ (95% CI)
DINOv3	Clean	1.60	1.54	1.51	1.67	1.75	2.13	7.3	-0.62	$+0.08$ (point est.)
DINOv3	Mild	17.18	13.97	14.11	13.78	12.76	13.18	55.8	12.76	-1.01 [-1.46, -0.57]
DINOv3	Medium	63.42	44.16	45.78	45.81	34.81	35.85	149.6	34.81	-10.99 [-12.18, -9.81]
DINOv3	Strong	72.85	49.23	50.92	51.74	37.06	38.10	159.7	37.06	-14.68 [-15.31, -14.05]
DINOv2	Clean	4.48	3.71	3.72	3.57	3.16	3.89	15.5	3.08	-0.42 (point est.)
DINOv2	Mild	22.99	18.39	18.76	17.48	16.44	16.93	71.3	16.44	-1.03 [-1.21, -0.86]
DINOv2	Medium	47.35	34.73	35.87	36.17	29.07	30.19	124.3	29.07	-7.10 [-8.25, -5.95]
DINOv2	Strong	56.11	39.89	41.38	44.03	32.30	33.51	138.7	32.30	-11.73 [-12.78, -10.68]
YOLO11m	Clean	19.58	15.85	16.89	16.22	13.82	14.56	62.8	13.82	-2.40 (point est.)
YOLO11m	Mild	26.28	20.76	21.89	20.03	17.71	18.19	78.7	17.71	-2.32 [-2.47, -2.17]
YOLO11m	Medium	27.80	21.91	22.37	21.40	19.14	19.68	83.6	19.14	-2.26 [-2.52, -2.00]
YOLO11m	Strong	36.24	27.56	27.59	28.37	24.27	25.31	103.9	24.27	-4.10 [-5.28, -2.92]
YOLO26m	Clean	16.61	13.59	14.53	14.34	11.94	12.61	54.7	11.94	-2.40 (point est.)
YOLO26m	Mild	23.21	18.53	19.68	18.14	15.83	16.32	70.5	15.83	-2.32 [-2.41, -2.22]
YOLO26m	Medium	28.77	22.62	23.20	22.03	19.58	20.18	85.4	19.58	-2.45 [-2.71, -2.19]
YOLO26m	Strong	37.58	28.46	28.80	28.69	24.91	25.91	106.4	24.91	-3.78 [-4.68, -2.88]

Table 11: YOLO26m at two queries with deterministic and noisy answers. Paired intervals use ten degradation seeds; Clean uses point estimates.

Answers	Strength	EER MAE	Largest MAE	Δ (95% CI)
Deterministic	Clean	11.94	14.34	-2.40 (point est.)
Deterministic	Mild	15.83	18.14	-2.32 [-2.41, -2.22]
Deterministic	Medium	19.58	22.03	-2.45 [-2.71, -2.19]
Deterministic	Strong	24.91	28.69	-3.78 [-4.68, -2.88]
Noisy	Clean	12.04	14.42	-2.38 (point est.)
Noisy	Mild	15.94	18.24	-2.30 [-2.40, -2.20]
Noisy	Medium	19.68	22.14	-2.45 [-2.72, -2.18]
Noisy	Strong	25.01	28.79	-3.77 [-4.67, -2.87]

YOLO26m features and shared thresholds. The official `yolo26m.pt` checkpoint (Jocher et al., 2026) supplies frozen layers 0–9 through SPPF. We letterbox RGB person crops to 640×640 with padding value 114, scale pixel values by $1/255$, and globally average the resulting 512-channel feature map. A fixed feature mean from development crops at timestamps 0–159 is subtracted before ℓ_2 normalization. The backbone and feature processing follow the YOLO11m contract. All four encoders use raw matching/weak thresholds $(0.475, 0.450)$, shifted to $(0.7375, 0.725)$, and held-out windows 200–399. Frozen metadata verifies these shared thresholds across all 160 encoder–condition–seed inputs.

Paired evaluation and answer noise. For each encoder and strength, we average full-window repeat metrics within each of ten degradation seeds, then form the EER-minus-Largest difference within each seed. Degraded conditions use the paired Student-t interval defined above. Clean labels share one observable input and are reported as point estimates. Random uses 100 repeats; Largest, EER, Uncertainty, and Hybrid use one deterministic or 30 noisy repeats; No-query uses one. Noisy answers add rounded zero-mean Gaussian noise with standard deviation 1 and are clipped to the selected pair's capacity. Table 11 reports both modes for YOLO26m: each degraded mode–strength combination improves in all ten seeds, with paired intervals below zero. Max and Sum use the same observable view counts across all four encoders; their coverage and metrics are in Tables 15 and 16.

Table 12: Held-out WILDTRACK M2 replay at two queries: policy MAE and EER error metrics under fixed M1b thresholds.

		Policy MAE ↓						EER (ours) ↓		
Mode	Setting	No-query	Largest	EER	Random	Uncertainty	Hybrid	RMSE	MAPE (%)	Bias
Det.	F0	1.60	1.67	1.75	1.54	**1.51**	1.75	2.13	7.34	-0.62
Det.	F1	18.23	14.85	**13.54**	14.84	15.25	**13.54**	13.88	60.82	13.54
Det.	F2	21.18	17.08	**15.25**	17.06	17.29	**15.25**	15.62	68.53	15.25
Noisy	F0	1.60	1.67	1.74	1.54	**1.50**	1.74	2.12	7.34	-0.57
Noisy	F1	18.23	14.96	**13.65**	15.00	15.36	**13.65**	13.99	61.38	13.65
Noisy	F2	21.18	17.21	**15.39**	17.23	17.42	**15.39**	15.76	69.19	15.39

C Frozen vs. Adapted DINOv3 Diagnostics

C.1 Frozen vs. Adapted DINOv3

We evaluate three fixed representations under the held-out, annotation-simulated WILDTRACK protocol with the M1b-selected thresholds and two queries: F0 is the frozen DINOv3 control; F1 adds a development-supervised matching adapter while retaining the frozen backbone; and F2 partially unfreezes the final block and normalization layers in addition to the adapter. F0 is one deterministic control checkpoint. F1 and F2 report descriptive means across their three development-only training seeds, rather than independent significance estimates. Under the fixed M1b-threshold held-out annotation-simulated WILDTRACK protocol, the development-supervised adapter and partial-unfreeze adapter did not improve the frozen DINOv3 control.

C.2 M2 Replay and Image-Level Bridge Protocols

This new minimal M2 replay applies the frozen shared DAWSON replay core to the held-out WILDTRACK appearance-proxy inputs. It preserves each configuration's source thresholds, native gate metadata, model-seed identity, and response/RNG profile, and evaluates all six policies at budget two. F0 is the frozen control with seed 7; F1 and F2 retain their three descriptive development-only model seeds. The deterministic replay uses No-query=1, Random=100, and the four querying policies=1 repeat; the noisy replay uses No-query=1, Random=100, and Largest, EER, Uncertainty, and Hybrid=30 repeats. Within each repeat, metrics are computed on the full 200-window set, averaged over repeats, and then summarized over the model seeds for F1/F2. Max and Sum in the companion cache analysis are same-window zero-query diagnostics and are not substituted into this Table 12 M2 block.

The M2 replay and the bridge share the frozen multiplicity estimator, finite-response Q, and sequential EER update rule. They intentionally retain different experimental inputs: M2 uses the HCA appearance-proxy protocol with its representation-specific thresholds, whereas the bridge uses image-level crop degradation, its source threshold scale, and native per-configuration gate. The replay therefore compares policies within each stated configuration rather than treating M2 and the bridge as a single degradation trajectory. Table 12 reports the new seven-configuration M2 replay, while Tables 13–14 use the formal native-gate held-out protocol, and the development analysis below measures sequential policy-switching space. This M2 result is a new formal replay, not a byte-for-byte reproduction of V7 Hybrid output: the native gate is on for every M2 window in this input, so its Hybrid rows execute the EER branch throughout. Across F1/F2, EER improves over Largest-overlap in both formal modes; the Clean F0 control remains the reliable-proxy boundary where simpler queried policies are competitive. The same-window Max and Sum count baselines are reported separately in Tables 15–16; Max remains lower-error with zero queries across all conditions. This pattern identifies the role of residual-driven selection as the proxy weakens without conflating selector value with absolute fused-count accuracy.

C.3 Development switching-space analysis

This development analysis evaluates frozen F0 (model seed 7) on windows 0–199 at budget $B = 2$ with deterministic responses. We summarize each degradation seed over its full 200-window sequence; the ten Clean configurations share one observable dataset and count once, while each of

Table 13: Formal deterministic DINOv3 replay at two queries: held-out means over ten fixed degradation seeds.

		Policy MAE ↓						EER (ours) ↓		
Setting–seed	Condition	No-query	Largest	EER	Random	Uncertainty	Hybrid	RMSE	MAPE (%)	Bias
F0–7	Clean	1.60	1.67	1.75	1.54	**1.51**	1.74	2.13	7.3	-0.62
F0–7	Mild	17.18	13.78	**12.76**	13.97	14.11	13.43	13.18	55.8	12.76
F0–7	Medium	63.42	45.81	**34.81**	44.16	45.78	45.41	35.85	149.6	34.81
F0–7	Strong	72.85	51.74	**37.06**	49.23	50.92	51.12	38.10	159.7	37.06
F1–7	Clean	18.20	14.82	**13.46**	14.81	15.09	14.25	13.78	60.7	13.46
F1–7	Mild	21.89	16.61	**15.36**	17.60	18.17	16.09	15.65	68.0	15.36
F1–7	Medium	26.82	20.67	**18.66**	21.14	22.00	19.95	19.18	81.5	18.66
F1–7	Strong	28.22	23.12	**20.05**	22.08	22.65	22.36	20.82	87.2	20.05
F1–13	Clean	17.18	13.83	**12.84**	14.04	14.64	13.14	13.21	57.4	12.84
F1–13	Mild	20.79	15.77	**14.67**	16.78	17.53	15.24	14.99	64.8	14.67
F1–13	Medium	26.22	20.12	**18.16**	20.71	21.59	19.43	18.68	79.3	18.16
F1–13	Strong	27.89	22.76	**19.82**	21.85	22.32	21.99	20.59	86.3	19.82
F1–29	Clean	19.31	15.89	**14.31**	15.67	16.01	15.02	14.65	64.4	14.31
F1–29	Mild	22.75	17.41	**16.02**	18.24	18.97	16.70	16.35	71.1	16.02
F1–29	Medium	27.83	21.38	**19.19**	21.87	22.85	20.49	19.72	83.9	19.19
F1–29	Strong	29.05	23.51	**20.46**	22.66	23.50	22.55	21.19	89.1	20.46
F2–7	Clean	23.11	18.14	**16.23**	18.48	18.65	17.46	16.59	73.0	16.23
F2–7	Mild	25.52	19.05	**17.46**	20.24	20.79	18.53	17.77	77.5	17.46
F2–7	Medium	27.99	21.57	**19.24**	21.96	22.68	20.81	19.68	84.8	19.24
F2–7	Strong	27.03	22.24	**19.30**	21.22	21.67	21.46	19.93	84.8	19.30
F2–13	Clean	18.93	15.65	**13.99**	15.37	15.61	14.61	14.33	63.3	13.99
F2–13	Mild	22.43	17.06	**15.71**	18.00	18.61	16.43	15.98	69.7	15.71
F2–13	Medium	27.19	20.84	**18.83**	21.41	22.21	20.09	19.32	82.4	18.83
F2–13	Strong	28.41	23.07	**20.09**	22.21	22.81	22.22	20.87	87.4	20.09
F2–29	Clean	21.50	17.43	**15.53**	17.31	17.62	16.37	15.95	69.4	15.53
F2–29	Mild	24.41	18.39	**16.82**	19.45	20.12	17.73	17.20	74.3	16.82
F2–29	Medium	27.75	21.28	**19.04**	21.80	22.68	20.43	19.54	83.4	19.04
F2–29	Strong	27.66	22.51	**19.60**	21.67	22.27	21.60	20.25	85.8	19.60

Table 14: Formal noisy DINOv3 replay at two queries: held-out means over ten fixed degradation seeds.

		Policy MAE ↓						EER (ours) ↓		
Setting–seed	Condition	No-query	Largest	EER	Random	Uncertainty	Hybrid	RMSE	MAPE (%)	Bias
F0–7	Clean	1.60	1.67	1.74	1.54	**1.50**	1.74	2.12	7.3	-0.57
F0–7	Mild	17.18	13.87	**12.85**	14.10	14.20	13.52	13.26	56.2	12.85
F0–7	Medium	63.42	46.11	**35.02**	44.29	46.03	45.70	36.04	150.6	35.02
F0–7	Strong	72.85	52.05	**37.21**	49.45	51.20	51.43	38.25	160.5	37.21
F1–7	Clean	18.20	14.95	**13.58**	14.97	15.20	14.38	13.90	61.3	13.58
F1–7	Mild	21.89	16.71	**15.46**	17.77	18.28	16.19	15.75	68.5	15.46
F1–7	Medium	26.82	20.80	**18.76**	21.22	22.10	20.07	19.28	82.1	18.76
F1–7	Strong	28.22	23.23	**20.13**	22.17	22.76	22.47	20.90	87.6	20.13
F1–13	Clean	17.18	13.96	**12.96**	14.19	14.74	13.26	13.32	58.0	12.96
F1–13	Mild	20.79	15.87	**14.77**	16.94	17.64	15.33	15.08	65.3	14.77
F1–13	Medium	26.22	20.25	**18.26**	20.78	21.69	19.55	18.77	79.8	18.26
F1–13	Strong	27.89	22.87	**19.89**	21.94	22.43	22.10	20.66	86.6	19.89
F1–29	Clean	19.31	16.03	**14.43**	15.84	16.12	15.16	14.78	65.0	14.43
F1–29	Mild	22.75	17.51	**16.13**	18.41	19.09	16.80	16.45	71.6	16.13
F1–29	Medium	27.83	21.51	**19.30**	21.96	22.96	20.61	19.82	84.5	19.30
F1–29	Strong	29.05	23.62	**20.53**	22.77	23.60	22.65	21.26	89.5	20.53
F2–7	Clean	23.11	18.31	**16.38**	18.69	18.79	17.62	16.74	73.7	16.38
F2–7	Mild	25.52	19.17	**17.58**	20.43	20.92	18.64	17.88	78.1	17.58
F2–7	Medium	27.99	21.71	**19.35**	22.05	22.79	20.94	19.79	85.3	19.35
F2–7	Strong	27.03	22.35	**19.37**	21.32	21.77	21.56	20.00	85.2	19.37
F2–13	Clean	18.93	15.79	**14.13**	15.54	15.73	14.75	14.46	63.9	14.13
F2–13	Mild	22.43	17.17	**15.81**	18.17	18.73	16.54	16.09	70.3	15.81
F2–13	Medium	27.19	20.98	**18.94**	21.49	22.32	20.21	19.42	83.0	18.94
F2–13	Strong	28.41	23.18	**20.17**	22.31	22.92	22.32	20.94	87.8	20.17
F2–29	Clean	21.50	17.58	**15.67**	17.50	17.74	16.52	16.09	70.0	15.67
F2–29	Mild	24.41	18.50	**16.93**	19.64	20.24	17.85	17.31	74.9	16.93
F2–29	Medium	27.75	21.42	**19.14**	21.89	22.79	20.55	19.64	84.0	19.14
F2–29	Strong	27.66	22.62	**19.67**	21.78	22.37	21.70	20.32	86.2	19.67

Table 15: Same-window Max/Sum cache reanalysis for the formal replay; both use zero queries and provide count-quality diagnostics.

Absolute-count metrics.

Table	Coverage	Condition	Policy	MAE ↓	RMSE ↓	Bias	MAPE (%) ↓	q
10	4 enc. F0	Clean	Max	0.190	0.500	−0.190	0.710	0
10	4 enc. F0	Clean	Sum	90.820	92.744	90.820	390.510	0
10	4 enc. F0	Mild	Max	0.667	1.087	−0.667	2.584	0
10	4 enc. F0	Mild	Sum	87.650	89.527	87.650	376.887	0
10	4 enc. F0	Medium	Max	1.036	1.492	−1.036	4.037	0
10	4 enc. F0	Medium	Sum	85.319	87.151	85.319	366.998	0
10	4 enc. F0	Strong	Max	1.605	2.078	−1.605	6.381	0
10	4 enc. F0	Strong	Sum	81.770	83.536	81.770	352.067	0
13	7 setting seeds	Clean	Max	0.190	0.500	−0.190	0.710	0
13	7 setting seeds	Clean	Sum	90.820	92.744	90.820	390.510	0
13	7 setting seeds	Mild	Max	0.667	1.087	−0.667	2.584	0
13	7 setting seeds	Mild	Sum	87.650	89.527	87.650	376.887	0
13	7 setting seeds	Medium	Max	1.036	1.492	−1.036	4.037	0
13	7 setting seeds	Medium	Sum	85.319	87.151	85.319	366.998	0
13	7 setting seeds	Strong	Max	1.605	2.078	−1.605	6.381	0
13	7 setting seeds	Strong	Sum	81.770	83.536	81.770	352.067	0
14	7 setting seeds	Clean	Max	0.190	0.500	−0.190	0.710	0
14	7 setting seeds	Clean	Sum	90.820	92.744	90.820	390.510	0
14	7 setting seeds	Mild	Max	0.667	1.087	−0.667	2.584	0
14	7 setting seeds	Mild	Sum	87.650	89.527	87.650	376.887	0
14	7 setting seeds	Medium	Max	1.036	1.492	−1.036	4.037	0
14	7 setting seeds	Medium	Sum	85.319	87.151	85.319	366.998	0
14	7 setting seeds	Strong	Max	1.605	2.078	−1.605	6.381	0
14	7 setting seeds	Strong	Sum	81.770	83.536	81.770	352.067	0

the three degraded conditions averages ten fixed seeds. The four conditions receive equal weight. Table 17 reports Largest and EER as fixed branches, the native-metadata Hybrid diagnostic, and a window-level two-branch oracle with its headroom over the best fixed branch. This development-only analysis complements held-out results. Sequential EER is the lowest-error fixed branch in every condition (equal-weight MAE 14.574267 versus 14.555224 for the window oracle). The window oracle leaves only 0.130657% improvement headroom over sequential EER. Clean retains 5.467250% local headroom, identifying a localized switching opportunity under the reliable proxy.

Table 16: Count-quality metrics for the same Max/Sum coverage; all rows use zero queries.

Table	Coverage	Condition	Policy	CP ↑	CR ↑	CF1 ↑
10	4 enc. F0	Clean	Max	1.000	0.993	0.996
10	4 enc. F0	Clean	Sum	0.206	1.000	0.341
10	4 enc. F0	Mild	Max	1.000	0.974	0.987
10	4 enc. F0	Mild	Sum	0.212	1.000	0.349
10	4 enc. F0	Medium	Max	1.000	0.960	0.979
10	4 enc. F0	Medium	Sum	0.216	1.000	0.355
10	4 enc. F0	Strong	Max	1.000	0.936	0.966
10	4 enc. F0	Strong	Sum	0.223	1.000	0.365
13	7 setting seeds	Clean	Max	1.000	0.993	0.996
13	7 setting seeds	Clean	Sum	0.206	1.000	0.341
13	7 setting seeds	Mild	Max	1.000	0.974	0.987
13	7 setting seeds	Mild	Sum	0.212	1.000	0.349
13	7 setting seeds	Medium	Max	1.000	0.960	0.979
13	7 setting seeds	Medium	Sum	0.216	1.000	0.355
13	7 setting seeds	Strong	Max	1.000	0.936	0.966
13	7 setting seeds	Strong	Sum	0.223	1.000	0.365
14	7 setting seeds	Clean	Max	1.000	0.993	0.996
14	7 setting seeds	Clean	Sum	0.206	1.000	0.341
14	7 setting seeds	Mild	Max	1.000	0.974	0.987
14	7 setting seeds	Mild	Sum	0.212	1.000	0.349
14	7 setting seeds	Medium	Max	1.000	0.960	0.979
14	7 setting seeds	Medium	Sum	0.216	1.000	0.355
14	7 setting seeds	Strong	Max	1.000	0.936	0.966
14	7 setting seeds	Strong	Sum	0.223	1.000	0.365

Table 17: F0 development switching analysis: sequence MAE for fixed policies and an evaluation-only window oracle.

Condition	Largest	Sequential EER	Native Hybrid	Window oracle	Headroom (%)
Clean	1.449544	1.391486	1.445837	1.315410	5.467250
Mild	10.616427	9.475341	10.354546	9.475247	0.000986
Medium	31.560326	22.722308	30.814748	22.722308	0.000000
Strong	39.019603	24.707933	37.623591	24.707933	0.000000