

Hybrid Geometry-VLM Evaluation of Surveillance Norms on a Patrol Robot

Gijs Lutje Beerenbroek

MSc Robotics, Department of Cognitive Robotics

Thesis committee: Dr. H. Caesar • Dr.ir. Y.B. Eisma • W. van der Geest • A. Steenbeek

To be defended publicly on Friday 2 October 2026 at 10:00



Hybrid Geometry-VLM Evaluation of Surveillance Norms on a Patrol Robot

by

Gijs Lutje Beerenbroek

A thesis submitted to the Delft University of Technology in partial fulfilment of the requirements for the degree of Master of Science in Robotics, to be defended publicly on Friday, October 2, 2026, at 10:00.

Student number: 5065321
Faculty: Mechanical Engineering, Department of Cognitive Robotics
Thesis committee: Dr. H. Caesar, TU Delft, supervisor
Dr.ir. Y.B. Eisma, TU Delft, second examiner
W. van der Geest, DevOps, supervisor
A. Steenbeek, DevOps, supervisor



Hybrid Geometry-VLM Evaluation of Surveillance Norms on a Patrol Robot

Gijs Lutje Beerenbroek¹ Willem van der Geest² Anne Steenbeek³ Holger Caesar⁴

A patrol robot drives the same route repeatedly; which of the changes between one visit and the next a supervisor should see depends on the surroundings, earlier moments and local rules. We present a hybrid geometry-VLM pipeline for evaluating operational surveillance norms. LiDAR change detection and person tracking feed a spatio-temporal evidence graph of changed objects, people and site areas; the graph indexes what is passed to a vision-language model. The model characterises what it is shown and judges it against the site’s norms, which range from explicit through relational to implicit. Developed at one site and evaluated at another, with reportable and non-reportable situations staged, it reaches an F_1 of 0.70 at a precision of 0.66. A camera-only alternative nearly matches its recall but at an alert precision of 0.08; withdrawing the relational evidence costs 0.17 of F_1 . What a design must supply differs across the three norm categories.

I. Introduction

Airports, ports, power stations and military bases are guarded continuously. Since October 2024, European Union member states must ensure that operators of critical infrastructure protect their premises, and the legislation names perimeter monitoring among the measures expected of them [13]. This work concerns the surveillance of such a site.

The task falls to human operators. But operators watching camera feeds miss more as their exposure lengthens [10]. The obvious remedy is an automatic system that raises an alarm, but that can introduce the opposite failure. Operators who receive too many false alerts learn to discount all of them [9, 36]. A useful system has to fit between the two. It cannot ask a person to look at everything, and it cannot report so much that its reports stop being read.

This work was set by the Dutch Ministry of Defence, which asked what it would take for a patrol robot to take on part of that task. It provided RoboGuard, a wheeled platform carrying LiDAR, cameras and odometry. The sites concerned are guarded by the Royal Netherlands Marechaussee, which differ in layout, in use and in the rules that apply locally. Training a separate model for each of them is not a practical answer. The aim is to keep one

mechanism fixed and let each site state what matters there. These sites are also in use while they are guarded. The people the robot meets on its route are therefore part of the problem rather than noise to be removed. What they are doing, and how they stand in relation to the objects around them, is often what decides whether anything is wrong at all. The conditions under which the recordings were made, and the limits on what the system does with the people in them, are set out in Appendix M.

What counts as wrong is not a property of an event but of the expectations held at the site. A guard works to a set of them, running from written instructions to judgements that resist any complete definition. These expectations are separated by the kind of fact that decides them, into *explicit*, *relational* and *implicit* norms, defined with the site’s own examples in Section IV-B. Enumerating every way a bag may become unattended does not eliminate the need for judgement; it merely shifts that judgement into the rule set.

Two capabilities make the task newly approachable. A mobile robot can revisit a route and record what it sees, so one visit can be set against another [22, 28]. Vision-language models can describe open-ended scene content without a detector trained for every object and situation [29, 44]. The second matters because an implicit norm offers no predicate to check. If a judgement cannot be expressed as an explicit rule, a model must make that judgement from the available evidence. Training one per site is not practical, so we use a pretrained model. Some relevant context, however, is not contained in the image, such as where a restricted area ends or who had previously been standing beside a bag. The open problem is therefore determining what evidence a judgement requires when the relevant fact may be visible in an image, distributed over time, or defined by site policy.

This leads to the question the thesis answers: *how can a patrol-robot pipeline evaluate a broad set of operational surveillance norms, and how must that design change as the norms move from explicit to implicit?* We make four contributions:

1. We show how geometric change detection, person tracking and vision-language reasoning can be combined in one pipeline on a patrol robot to evaluate site-specific surveillance norms without training a separate model for each site (Section V-A, Section V-B).
2. We provide a spatio-temporal evidence graph that supplies the spatial, temporal and person-object context that is not always recoverable from a single image. It

¹MSc. student, Cognitive Robotics.

²Senior Data Scientist and Technical Lead at DevOps.

³Robotics lead at DevOps.

⁴Assistant Professor, Intelligent Vehicles Group.

evaluates no norm itself (Section V-C).

3. We develop an evaluation protocol for surveillance report-worthiness that stages non-reportable situations so precision is measured rather than assumed, scores alerts against norm-aware ground truth, and treats abstention separately (Section IV-B, Section IV-C).
4. We show experimentally that, in this evaluation, explicit, relational and implicit norms make different demands on the design, in the evidence, the wording and the architecture they require (Table 4).

The pipeline, evidence assembly, prompts and judge scaffolding were fixed on the development site; the norms and zones were supplied for each site, and the geometry operating point was fitted on the evaluation site (Appendix L.1). The scenes are staged and the system is evaluated offline. Within that scope the evaluation measures event recall and alert precision against a camera-only alternative, and shows that comparable recall can be turned into a smaller and more precise operator-facing alert stream.

II. Related Work

Several lines of work bear on the decision our pipeline has to make. Change detection locates where a scene differs, and at most says what changed. Scene graphs hold relations, but only the ones somebody encoded. Anomaly detection scores how unusual a situation is. Unusual is not the same as report-worthy. The normative systems reviewed here decide report-worthiness, but only for situations whose conditions were written in advance. What is missing between them is a system that does three things at once. It recovers the change geometrically. It supplies the relations no frame carries. And it leaves the judgement of whether the norm covers the situation to a model reading the image.

A. Change detection, captioning, and tracking

Change detection locates where a scene differs between two visits. It is the starting point of our pipeline. The closest precedent is a security patrol robot that compared a current patrol against a reference model and flagged where the two differed [2]. That is our problem without the layers that interpret and judge the change. Later work detects changes directly in the LiDAR point clouds of a mobile robot [22], or without labels at all, by comparing renderings of a reconstruction of the scene [14]. Other work also generates semantic descriptions of 3D scene changes from paired point-cloud scans [27,31]. All of these report that something has changed, and increasingly what it is. None decides whether the change is worth reporting. That decision is the subject of this thesis.

People contaminate the map by leaving a ghost trail. Methods exist that erase dynamic traces from an accumulated cloud, by testing visibility across range images [21] or by an occupancy ratio per bin [26]. Both assume one consistent accumulation. Two separately recorded patrols

do not give that. Hence we remove the points at the track instead, before either map is built.

B. Scene graphs and structured state

A scene graph describes a scene as objects and the relations between them. Such graphs are often kept as a lasting three-dimensional map [15], and extended to hold scenes that change while a robot works in them [16]. Relations can then be reasoned about directly on the graph, for instance by scoring which of them are anomalous [24]. But a graph-based decision is limited by what the representation retains, and such graphs are incomplete [25]. Enumerating in advance every relation that might matter is infeasible. Representation granularity therefore fixes what a system can reach, before any reasoning happens. We use the graph only as an index, and make the judgement from the camera image, which carries the relations directly (Section III-C).

C. Anomaly detection and rule-based monitoring

Anomaly detection learns what a normal scene looks like and flags whatever departs from it [30,32]. For surveillance, this framing does not match our objective. A patrol meets a wide and open-ended set of scenes, so a stable “normal” is hard to establish for any of them. A rare change is often harmless; a common one can matter. An anomaly score is also hard to act on: what counts as a high one depends on the algorithm that produced it and on the data it was scaled against [19]. A separate line of work evaluates scenes against rules specified in advance [7]. Its limitation is the rule space: the conditions must be defined before deployment. This is precisely where much of surveillance sits: in softer, unwritten expectations [5]. Closest to our aim is semantic anomaly detection, which brings a language model’s general world knowledge to bear on whether a situation is wrong in meaning rather than in statistics [11]. More generally, using strong language models for open-ended evaluation is an established methodology [43]. We build on these ideas by judging each change for report-worthiness against the site-specific operational norms, from clear written rules to unwritten expectations. Semantic anomaly detection is also the closest example of how we measure. That work drives a simulated car past ordinary stop signs and traffic lights and counts the false alarms [11]. We stage relational conditions as matched pairs, one reportable and one non-reportable, and count which of the two it reports.

D. Norms in agent systems

Normative multi-agent systems study how artificial agents follow norms [6]. A norm there is an explicit rule together with the exact conditions that decide whether a situation breaks it. A designer fixes those conditions by hand, before the agent is deployed [1]. Deciding that a particular situation counts as breaking a norm is what Searle calls a “counts-as” judgement [34]. Fixing that judgement in advance is the difficult step.

Applied to surveillance, that approach means specifying

in advance the conditions under which a bag counts as unattended. The norm itself is simple: a bag must not be left unattended. The situations that can lead to that state are not. Its owner may walk away, the vehicle beside it may leave, or nobody may have been near it in the first place. Writing a separate condition for every possibility is not practical, and in a security setting the relevant case may be precisely the one that was not anticipated. We therefore keep the norm as a statement about the required state and make the counts-as judgement when the situation is evaluated. The judge receives the evidence recovered from the scene, including measured relations, and decides whether the situation falls under the norm. This changes what the norm text needs to specify. Rather than listing every possible mechanism, it must provide the cues that distinguish a violation from routine activity. How such instructions are written is known to affect model behaviour. Meaning-preserving reformulations change it [33], and so does how explicitly the evaluation criteria are stated [18]. Those studies measure the effect through agreement on a benchmark. We instead test how norm wording affects the operational decision and where additional specification stops helping.

None of these lines establishes what evidence a counts-as judgement requires. In Searle’s work, whether a situation counts as a violation depends on its context, and that context is part of the condition. A norm that depends on a relation to something outside the object therefore cannot be evaluated from the object alone. When the necessary evidence is visible, what a model can extract depends on how the view is framed [37, 41]. A common alternative is to describe the scene in words first, but such descriptions can be ambiguous or underspecified [11]. We therefore test whether different norm categories require different kinds of evidence, and report those differences in the evaluation and discussion.

III. Method

The pipeline reads two recordings of the same route, a reference patrol and the patrol under evaluation, and returns a report-or-ignore verdict on each change and each person it finds, together with the norms that verdict cites. Two streams run over the current recording, one for static object change and one for the people present during the patrol, and both pass through the same three layers. The geometry layer establishes what has changed and who is present, working directly on the LiDAR point cloud. The semantics layer describes what each flagged location contains, factually, from the camera evidence. The norm layer decides whether that is report-worthy against the site’s norm set. The two streams are described first, then the graph that fuses them and settles what the judge is shown, then the judgement itself.

A. Static change detection

The static stream detects objects that have appeared or disappeared between two patrols. The reference patrol is

a prior recording of the route, taken as the state in which no norm is violated. The current patrol is the one under evaluation. Each patrol is accumulated into a world map: every LiDAR point is placed in a common frame through the robot’s odometry, and the result is voxelised. The two maps are then overlaid. A voxel is either occupied in both, neither, or exactly one, and only the last case marks a change. A voxel is considered *removed* if it only appears in the reference patrol, and *added* if it only appears in the current patrol. The occupancy test is a nearest-neighbour comparison within a distance T to absorb slight residual registration error. Changed voxels are grouped into candidate clusters with DBSCAN [12]. An observability gate then drops the clusters that reflect coverage rather than change. What is left is what we call *change clusters*. Every threshold named in this section is given with its purpose in Appendix A.

B. The dynamic stream: person detection, tracking, and stripping

The dynamic stream recovers the people present during a patrol. It does so for two purposes: their points are taken out of the point clouds before the map is built, and each person becomes a unit the norms can be evaluated on. A unit is what a verdict can be attached to, a changed object or a person. The static stream assumes a world that changes only between patrols. A real patrol shares the site with people, and people matter twice. Left in the point cloud, a walking person accumulates into the current map as occupied space. The change detector then reports a ghost trail of clusters along that person’s path. Attended to, the same person is a subject of the norm set. Whether someone loiters, enters a restricted area, or leaves an object behind cannot be asked until the system knows who was present and where they went.

People are detected directly in the LiDAR. Each scan passes through CenterPoint [40], a voxel-based 3D detector pretrained on the nuScenes autonomous-driving dataset [8], and the detections are linked into tracks by SORT-style association [4, 23]. Track fragments separated by longer occlusions are then re-linked when their temporal and spatial separation is physically plausible [17, 42]. Before map accumulation, points within each confirmed person’s footprint are removed from the LiDAR scans. For scans where the detector does not fire, the person’s footprint is interpolated from neighbouring detections and the points inside it are removed, preventing a ghost trail from forming (Figure 6). The tracks themselves are kept and enter the evidence graph as the dynamic stream. Appendix B.1 gives the association and re-linking, the confirmation rule and the stripping procedure, and Appendix A their parameters.

C. The evidence graph

The evidence graph fuses the static and the dynamic stream into one spatio-temporal structure, and serves as an index. One further source of information is injected there: the site’s map, a set of area polygons drawn once by hand. It

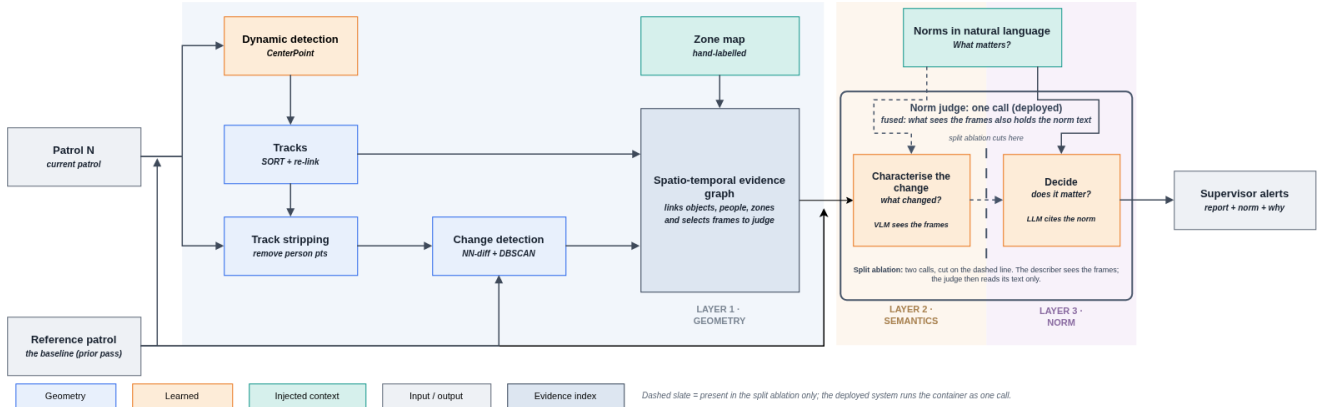


Figure 1. Overview of the pipeline. A single patrol recording feeds two streams, static object change and dynamic people. Both are lifted through the same three layers and fused in one spatio-temporal evidence graph, which a single norm judge reads to emit report-or-ignore alerts. Geometry comes first because it is a cheap, illumination-independent filter, so the costly image-level reasoning is applied only where something has moved.

is not perceived, because an area may be marked on the ground and still not be readable from a frame. Where it ends, and which rule it carries, are set by site policy. The nodes are the tracked people, the change clusters and the injected zones, and four typed edges are measured between them: how close a person came to a change and for how long, how long a unit lay inside a zone, how long two people stayed within range of each other, and whether a removed cluster pairs with an added one. Figure 7 shows one such structure being built from a single frame. One edge type also changes the unit itself: a pairing edge links a nearby removed and added cluster when they are likely to represent the same object before and after it moved. The pair is then treated as a single unit and judged once rather than twice. The gate on each edge type is given in Appendix A. The whole graph for one patrol, with the measurements its edges hold, is shown in Figure 8.

Enumerating every relation that might matter is infeasible, so the graph’s job is selection: it settles what the judge is shown. Each unit and each edge gets one view: a bird’s-eye rendering, or a camera frame chosen from every pose and camera the patrol recorded by a cost over centring, range, occlusion and cut-off at the frame edge. A vision-language model handed a wide camera frame often fails to attend to a small, non-salient subject in it [37, 41], and a change region is typically exactly that. The target is therefore marked with a rectangle projected from its own voxels. What the judge receives for one unit is that unit’s own evidence together with its one-hop neighbourhood, in pictures and in numbers. Appendix E gives the cost in full, why the frame is left intact rather than cropped, and the contents of the package.

The pictures carry more than the edges measure. Take a bicycle left in front of a door. The door never changed, so it is never a node, and “blocks the door” was never an edge. But the relation is there in the frame, and Figure 10 is one such alert as the system produced it. A zone boundary

or the passage of time must therefore be measured and supplied as an edge, while a relation to the unchanged surroundings is left to the norm layer to read from the image.

D. Interpretation and norm evaluation

Each unit is judged in a single call to one vision-language model, `claude-sonnet-4-6` [3]. The call runs through the public API with default sampling. The call receives the evidence package together with the norm set. It returns a factual characterisation of the unit, a report-or-ignore decision, any norms it considers violated, and a short explanation. Describing and judging therefore happen together. Splitting them is a design alternative: a first call, the *describer*, reports what it sees, and a second, the *judge*, applies the norms to that description. The evaluation compares two split configurations: one in which the describer receives the norm set and one in which it does not.

A verdict is a set of cited norms rather than a single label, because one unit can break several norms at once. Each cited norm is attributed to the unit that norm is about. That attribution is the anchor. The alert stream $\mathcal{A}$ is the set of distinct norm-and-anchor pairs (n, a) . Hence an operator sees one alert per distinct claim about the world.

Where the evidence does not settle whether a norm is violated, the judge returns that norm as uncertain rather than firm. If the system must either make a firm decision or remain silent, uncertainty is forced into one of those two outcomes, both of which can be problematic in operational surveillance [20, 36]. Abstention is a verdict, and Section IV-C scores it on its own terms: it never counts as a correct report.

The norm set is written to suit that arrangement. A norm states a required state of the world: a bag must not be left unattended, an area must remain clear. The mechanism that brought the state about is read off the evidence at judgement

time rather than written into the rule. Abandonment is the clearest case. No norm in the set mentions departure. But the temporal edge records that a person stood beside the object and has since left, and that the object remained. The judge reads unattended from that. The full norm set of each site is given in Appendix F, and the system prompt in Appendix N.

IV. Experimental setup

A. Sites, robot, and data

The recordings come from a RoboGuard patrol robot, an outdoor ground vehicle. It carries a 128-beam Ouster LiDAR turning at 20 Hz with a configured range of 30 m, 4 RGB cameras covering nearly the full circle around it, and an inertial navigation system reporting odometry in a local east-north-up frame.

The site is the Koning Willem III Kazerne (KW3), a training facility of the Royal Netherlands Marechaussee. It is representative of a site the robot is intended for. The robot drove autonomously between waypoints, past an office building, greenery and a car park. The route was chosen because objects could be staged along it. One reference and two patrols were recorded in June 2026, each patrol scored against the reference, and both were staged before any norm result existed. The pipeline was developed on an earlier site and its mechanisms brought to this one unchanged, with the geometry operating point refitted here (Appendix L.1).

Every change along these routes was staged deliberately. Ground truth was annotated afterwards in three layers (Appendix G), all of it by the author, who is a certified security officer.

B. Norm set and scenario staging

The KW3 set holds 12 reportable norms and a non-reportable class that routine activity falls into. The norm set was developed with personnel working at KW3. Full texts in Appendix F.

Norms are sorted by the kind of fact that decides them, into three categories plus the non-reportable class (Figure 2). An *explicit norm* turns on a property of a single entity and is settled by a lookup on that entity or its state: a fire extinguisher missing from its bracket. A *relational norm* turns on a relation to a second term the entity does not itself carry, an access route, an area, an owner, an earlier moment: a bag with nobody attending it. An *implicit norm* turns on no crisp predicate at all and is settled by reading the situation as a whole: disorder. The judge is never told which category a norm belongs to. The categories still predict what each kind of norm needs supplied.

Several norms carry an exception bound to a type of area: an unattended bag in the break area is routine, the same bag elsewhere is not. Each relational variable is staged as a pair, one reportable take and one not, differing only in that variable. A bag left alone against the same bag with its owner beside it. A person inside the restricted

arm	P	R	F_1
patrol 1	0.63 ± 0.09	0.69 ± 0.04	0.66 ± 0.06
patrol 2	0.70 ± 0.05	0.78 ± 0.03	0.74 ± 0.02
combined	0.66 ± 0.02	0.74 ± 0.04	0.70 ± 0.02

Table 1. KW3 held-out norm-layer result: event-level recall, alert-level precision, mean $\pm$ one standard deviation over the repeats.

area against one on the walkway past it. The routes were staged far denser in norm-relevant change than a patrol would ordinarily meet: 15 and 17 reportable events over the two patrols. 10 explicit, 9 relational and 13 implicit (Figure 2). A further 48 staged changes are non-reportable by design.

C. Metrics and ground-truth protocol

Two scores are kept distinct throughout, because they measure different layers. The *geometry* score measures localisation alone: a change cluster is a true positive when it falls inside a ground-truth change box. The *norm* score measures report-worthiness over the alerts that leave the pipeline (Section III-D). Recall is measured at the event level and precision at the alert level, using the same true-positive count: the first alert that binds a reportable event is a true positive, any subsequent alert binding the same event is *absorbed*, and all other alerts are false positives, including one whose cited norm matches no candidate event at all. Abstention is scored separately, as correct, missed, or an over-claim that is also a false positive, and never credits recall. Appendix G gives the three labelling layers.

Because the VLM calls are stochastic, run-to-run variation is measured explicitly. Each arm is run three times on identical evidence. Judge-derived rates are reported as the mean and standard deviation across these runs. With only three runs, the standard deviation is descriptive; it is not a confidence interval and no significance tests are performed. Across repeated runs, up to five verdicts per patrol differ. The number of changed verdicts also depends on the norm wording (Section V-D).

V. Results

A. Main result on the held-out site

On the held-out site the pipeline reaches an F_1 of 0.70. Over both patrols the pipeline reaches a precision of 0.66 ± 0.02 , a recall of 0.74 ± 0.04 and an F_1 of 0.70 ± 0.02 (Table 1). On the median run that is 23 of 32 reportable events caught against 11 false positives and 9 misses. In addition, 5 duplicate alerts were absorbed. Patrol 2 leads on recall and F_1 in every repeat, but at its widest the gap is 1.4 events, too small to attribute to the site.

Underneath that number is what the geometry layer handed the judge. At its operating point that layer localised 44/50 staged events against 16 false positives, a recall of 0.88 at a precision of 0.73, an F_1 of 0.80. The

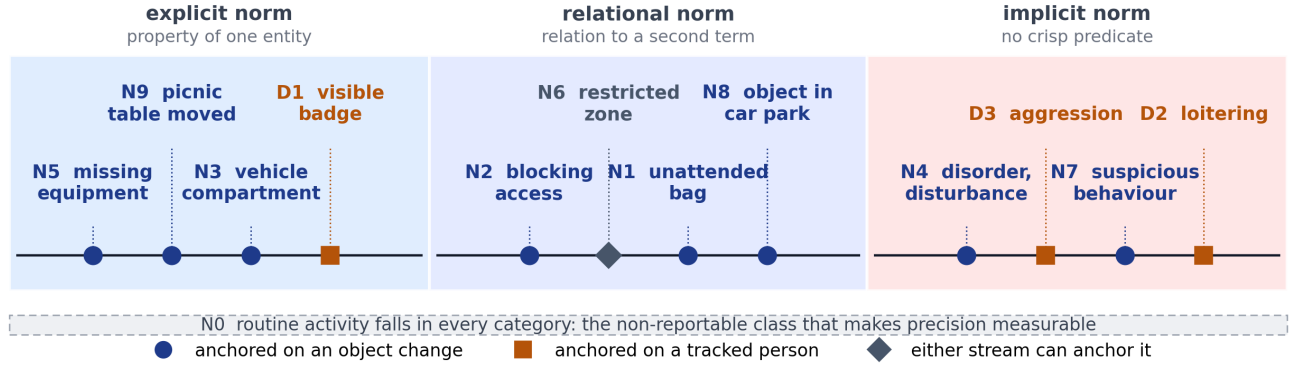


Figure 2. The KW3 norm set in the three categories it was written to span. Each holds four norms, and position within a band carries no meaning. Marker shape gives the stream that anchors a norm, and N0 is the class routine activity falls into. Staged reportable events, both patrols combined: 10 explicit, 9 relational, 13 implicit, 32 in total.

recall matters more than precision. A false positive is a cluster the judge inspects and dismisses. It turns out only 1 of the 11 false norm fire sits on a cluster that matched no staged change. A false negative is unrecoverable, because the norm layer can only subtract. 31 of the 32 reportable events cleared the filter; the one that did not, a bag set down beside a route in patrol 1, is the only miss the norm layer never had the chance to decide. The 7 events missed in every run have five distinct causes. Appendix H attributes each of them, gives the abstention calibration and takes the disorder events apart; Appendix I reproduces three reported alerts and one missed event as the system produced them.

Recall drops only for implicit norms. Explicit norms reach 0.90 recall and relational norms 0.85, both above the overall recall in Table 2. Implicit norms reach 0.54. Much of this drop is driven by the dwell norm D2, which is never firmly reported across its 3 staged events.

Precision does not follow recall. Explicit norms are claimed almost exactly once per staged event, 27/33, so 0.82. Relational norms are claimed far more freely, 23/42, so 0.55, the lowest in the table despite their high recall. Implicit norms sit between at 0.66.

B. VLM baseline

The obvious alternative to the pipeline is a camera and a norm list. The baseline is exactly that: one vision-language call per viewpoint, seeing the reference and the current frame, with no geometry, tracking or relational graph. Its firings carry no anchor, so matches to ground truth were labelled by hand, which lends it a localisation it does not produce. Two of its dynamic hits, a loitering and an aggression event, are defined over time and cannot honestly be earned from one frame. It was run once, so it carries no spread. Its recall is an upper bound for the two reasons above. Appendix G gives the sampling and the scoring.

One event apart on the full staged set. Both systems are scored on all 32 reportable events, including the one change the geometry layer produced no cluster for, which

the baseline caught and the pipeline could not. On the counts run the pipeline reports 23 of them and the baseline 22, R 0.69. Aggregate recall separates little, and is hence not where the pipeline’s value lies.

Per category they separate in opposite directions (Table 2). On explicit norms the pipeline leads, 0.90 against 0.70. On relational norms the pipeline leads by the widest margin, 0.85 against 0.56. On implicit norms the baseline leads, 0.77 against 0.54, and that lead is one norm. It reads disorder 7 times out of 7 against the pipeline’s 4. Disorder is the one implicit norm that stands whole in a single view and needs no second term, so describing every viewpoint gives the baseline more chances at it. The relational lead is the opposite, and all of it sits in the 3 zone-bound events, the same ones that fail in every repeat once the injected map is withdrawn. What the pipeline buys on this category is the second term the predicate requires. Splitting by whether a single view is sufficient shows where the difference lies. The systems tie on the 23 single-view events (18 each) and differ by only one event on the remaining 9 cases (5 vs. 4), so this is only a weak proof.

The main separation is precision. At event level the baseline scores P 0.16 against 0.66, while its operator-facing precision falls to 0.08. It is beaten on precision in every category, including the one it wins on recall: its implicit alerts run at 0.22 against the pipeline’s 0.66. Firing at every viewpoint produces 267 reports to catch 22 events, with one change reported up to 18 times. The pipeline produces 39 alerts to catch 23 events. The baseline’s two precision figures differ because its firings carry no anchor: the annotator deduplicated them by hand, which a deployed version could not. The pipeline’s anchor does that itself, absorbing 5 duplicates. Appendix H gives the event-level error analysis.

C. Evidence ablations

Each arm here removes one route by which evidence reaches the judge and predicts what should break. The reference is the headline configuration of Table 1 with

norm category	events	pipeline			primed split			camera-only		
		R	P	F_1	R	P	F_1	R	P	F_1
explicit norm	10	0.90 ± 0.10	0.82 ± 0.02	0.86 ± 0.05	0.90 ± 0.10	0.68 ± 0.09	0.78 ± 0.10	0.70	0.22	0.33
relational norm	9	0.85 ± 0.06	0.55 ± 0.02	0.67 ± 0.03	0.85 ± 0.06	0.45 ± 0.02	0.59 ± 0.03	0.56	0.09	0.15
implicit norm	13	0.54 ± 0.00	0.66 ± 0.04	0.59 ± 0.01	0.74 ± 0.04	0.62 ± 0.06	0.68 ± 0.05	0.77	0.22	0.34

Table 2. Event recall, alert precision and F_1 per norm category, over all 32 staged events; the events column is per repeat. Pipeline is the fused configuration, primed split the recall end of the same pre-site trade-off; the baseline was run once, so only it carries no spread. A miss belongs to the event’s own norm and an alert is charged to the norm it cites, which is why the two rates have different denominators.

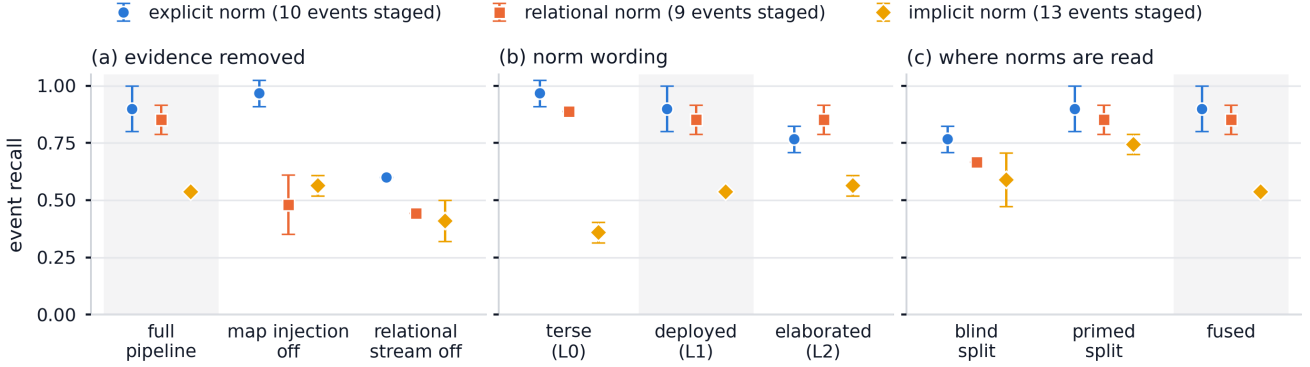


Figure 3. Event recall per norm category under each design lever, KW3 combined patrols; bars are one standard deviation. The shaded column is the same headline configuration in every panel. (a) evidence removed; (b) norm wording; (c) where the norms are read.

nothing removed. Aggregate results for all arms are given in Appendix J.

No relational evidence: every category falls. The person and site-map streams are removed at graph build, leaving the judge little more than a list of isolated changes. Recall falls by 0.26 and F_1 by 0.17, well beyond the run-to-run variation of either arm (Figure 3a). Person-anchored norms occur in all three categories, so removing this evidence affects all three.

No map injection: the zone-bound reports disappear. A finer cut hides only the zone-containment context, keeping every other relational cue, and so isolates the injected map. Recall on relational norms falls from 0.85 ± 0.06 to 0.48 ± 0.13 , 23/27 satisfied against 13/27, with no overlap between the spreads, while the other two categories stay within their spreads (Figure 3a). Every restricted-zone and car-park report disappears once the containment edge is hidden, at each repeat. KW3 exercises these norms lightly, one restricted-zone and two car-park events out of 32, so a mechanism that fails completely still costs only a couple of events.

D. Norm wording

Each norm is written at three levels of detail. Terse (L0) states the rule and its scope. Deployed (L1) adds the cues that separate a violation from a routine situation, and is the wording used in every other experiment. Elaborated (L2) spells those cues out further. Scope and exceptions are held constant across the three, so only the cues change (Appendix K). The evidence and the evaluation labels are

fixed. So is the rest of the prompt, which already gives the site’s zone rules and when to abstain, so this lever measures what the norm text adds on top of that scaffold. The elaborated wording was written after the system had been run on this site.

Two sets per norm. Write V_n for the situations that truly violate norm n , fixed by the ground truth, and $D_{n,r}$ for those the judge reports under it at wording r . Their overlap is a correct report, $D_{n,r}$ outside V_n a false one, and V_n outside $D_{n,r}$ a miss, so the precision and recall of Section IV-C are read off the two sets. Wording moves $D_{n,r}$ only. It can move it differently for different norms.

Precision rises at every step: 0.46, 0.66, 0.72. The system is not simply reporting less. True alerts hold while false ones fall (Table 3), and precision rises in every category.

Recall does not follow. It moves 0.70, 0.74, 0.71: up, then back. The categories move in opposite directions (Figure 3b). Explicit recall falls from 0.97 to 0.77, implicit recall rises from 0.36 to 0.56, and relational recall hardly moves. In both directions the terse and elaborated spreads are disjoint. No category peaks at the deployed wording, although the aggregate does: over the first step the implicit gain is large and the explicit loss is small, and over the second the implicit gain stops while the explicit loss continues. F_1 therefore rises and then flattens, 0.56, 0.70, 0.71. The elaborated spread, ± 0.06 , contains the deployed value, so three repeats do not establish the elaborated wording as the better one.

norm category	terse	deployed	elaborated
explicit	32/16	27/6	23/5
relational	24/51	23/19	23/13
implicit	11/11	21/11	22/9
all	67/78	71/36	68/27

Table 3. True and false alerts, given as TP/FP, per norm category at each wording level, pooled over the 3 repeats and both patrols; absorbed duplicates are excluded. The rates these counts underlie are means over the same repeats, so the two are not addable.

The categories reach that by three different routes (Table 3). Relational norms shrink what lies outside V_n and keep the overlap, so precision rises and recall barely moves. Implicit norms enlarge the overlap while what lies outside stays near constant, so both rates rise. Explicit norms lose on both sides, faster outside V_n than inside. The reported sets are therefore not nested. Going from the terse to the elaborated wording, 28 false reports and 2 true ones leave $D_{n,r}$, while 4 true reports and 0 false ones enter it. Elaboration is not the terse wording filtered: it changes which situations are reported, not only how many. Each moving category is largely one norm: disorder for implicit, whose recall runs 0.29, 0.57, 0.62, and the badge for explicit.

The explicit loss is evidential. Badge recall falls 0.92, 0.75, 0.42, while the other explicit norms stay at ceiling in every run. The elaborated badge rule also states when the available view cannot settle the question. Over the 6 runs per level, firm badge citations fall 19, 10, 5, while abstentions rise 56, 66, 72. On 7 occasions the same unit is reported at the deployed wording and abstained at the elaborated one in the same repeat. The judge is not ruling the violation out. It declines to certify a badge it cannot resolve, and under this scoring that abstention is a miss.

Wording cannot supply an observation the evidence lacks. The disorder gains came from cues the frames already supported. The dwell norm shows the other side: across its 3 staged events it is firmly reported at none of the three levels, because no duration in the evidence can be compared against a threshold the norm never names. A more detailed badge rule cannot make an unreadable badge legible either.

E. Norm placement

The last lever is which stage of the pipeline reads the norm text. The *headline* configuration fuses characterisation and judgement in one call that sees the frames. The *primed split* makes two calls, a norm-aware describer that sees the frames and a judge that sees only its description. The *blind split* withholds the norm text from the describer and supplies it only at judgement. Figure 1 marks where the split cuts. Crossed with the wording ladder this gives nine configurations on byte-identical evidence up to and including the graph. The wording pattern of Section V-D

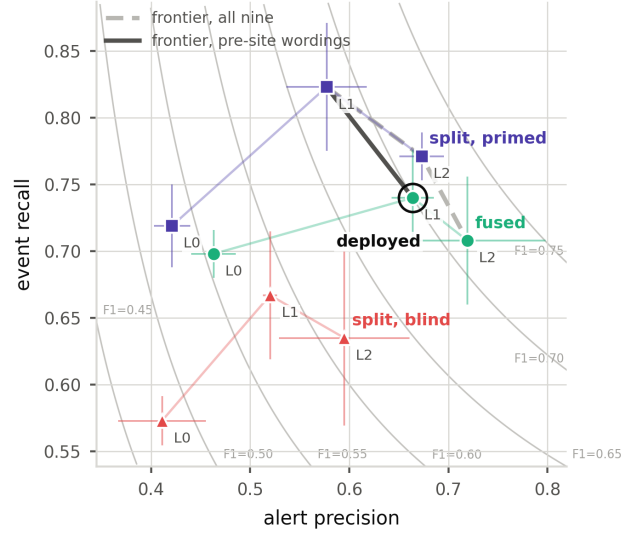


Figure 4. The nine configurations with iso- F_1 contours in grey. Each trajectory is one placement of norm knowledge, walked from the terse to the elaborated wording; whiskers are one standard deviation on both axes.

holds at each placement: precision rises with detail, and recall peaks at the deployed wording (Figure 4).

Priming dominates, while fusion trades recall for precision. Giving the describer the norm text improves recall at every wording level in Figure 4, with no overlap between the recall spreads, and never at the cost of precision, which rises as well at the deployed and elaborated levels. Moving from the primed split to the fused call produces a smaller trade-off: precision rises and recall falls at every level, while the F_1 spreads overlap. The choice between them is therefore an operating point, not a quality ordering. The fused configuration is reported as the headline because it is the one carried unchanged from the development site. No other configuration beats these three on both precision and recall at once: elaborated fused (P 0.72, R 0.71), elaborated primed split (P 0.67, R 0.77), and deployed primed split (P 0.58, R 0.82). The first two use elaborated wording written with this site’s observed failure modes in view, so they show what additional specification can achieve after seeing the site. The deployed wording was itself revised after the first runs here (Appendix L.1), so the comparison is one of degree rather than of wording written blind against wording written after. All nine configurations are reported in Table 8.

The two steps reach different categories (Figure 3c). Blinding the describer costs explicit norms and relational norms, whose predicates name something the describer has to have resolved first. Fusing the calls costs the third: implicit recall is 0.54 fused against 0.74 primed. This is the only comparison in which the fused architecture is the worst of the three, and it is the same cell on which the camera-only baseline led.

The fused call produces fewer additional norm cita-

tions. A judge reading a description can rule out a norm only from facts preserved in that description, while a judge holding the frames can also use visual evidence that was not stated. This buys precision, since every norm a verdict cites is charged separately. Both arms name a first norm about equally well, 0.69 fused against 0.67 primed; the primed split then adds 5.3 further citations per pass against 1.7 fused, at a precision of only 0.23. A wrong denial can also cost recall. Once a unit is described as upright it is never subsequently reported as disorder, on 0 of 31 judgements across the grid. The fused call commits to the object’s condition less often, stating it on 0.33 of units against 0.79 for the primed split. On one tipped bin the fused call recorded it upright and ruled out three norms by name, losing the event. Fusing changes two things at once, however: the judge gains direct visual access and description and judgement become one call, so the experiment does not separate their effects.

VI. Discussion

What a norm asks of the design. Read by norm category, the results answer the second half of the research question, how the design must change as the norms move from explicit to implicit. Across the evaluated norms, the three categories expose different requirements of the design (Table 4). The opposing category-level recall trends reflect changes in the overlap between $D_{n,r}$ and V_n (Section V-D), not differences in the size of V_n .

For explicit norms, the recall loss under elaboration is concentrated in the badge norm, where reporting depends on whether the visual evidence resolves the required observable. For implicit norms, the increase in disorder recall is consistent with added cues helping the judge recognise the intended violation. For relational norms, wording improves precision but changes recall little; a deciding relation must still be available in the image or supplied through other evidence.

The available evidence also limits which violations can be established. A norm cannot be evaluated reliably when its deciding fact is neither observable in the supplied images nor represented in the other evidence. The KW3 norm set contains one such case. The dwell norm D2 names no threshold and nothing in the evidence supplies one (Appendix H); beyond that, the only duration reaching the judge is the length of the tracked segment, and that is bounded by the robot’s own pass. In the tested configurations, D2 is never firmly reported at any wording level or under any architecture. The tested changes to wording and norm placement do not overcome this limitation.

The badge norm shows a different limitation. Abstinence occurs only for this explicit norm, and not for any implicit one. What a norm needs specified and what it needs observed are different demands.

The different effects of wording, evidence removal and norm placement support these different design requirements. The evidence for this pattern is limited: each

category contains only a few norms and staged events (Section IV-B). The main support is therefore that different levers affect different categories, and that this pattern repeats across architectures.

The pipeline. It was measured on a site it was never developed on (Section IV-A). The recall weakness predominantly stems from the layers above the change detector. The disorder events it loses were all detected as change clusters; what prevented their reporting was view selection and interpretation (Appendix H). Splitting characterisation from judgement recovers most of it (Section V-E), and elaborating the norm text raises the same rate at every rung (Section V-D). Neither supplies a second look, which is how the baseline succeeds here (Appendix H). Localisation on this site is in-sample. The fold-out and cross-site transfer in Appendix L.3 show that the same geometry pipeline remains usable across splits and sites, but its operating point does not transfer unchanged.

The evidence graph. The ablations support the need for the relational evidence it carries, rather than the graph representation itself. Removing the relational layer costs 0.17 of F_1 , well beyond the run-to-run spread, and affects every category because person-anchored norms occur in all three. Hiding only the injected map leaves the other categories unchanged, while every restricted-zone and car-park report disappears. This selective failure supports the need to supply relations that the image cannot provide directly. What neither ablation shows is that a graph is the right container for that evidence: both vary what reaches the judge, and neither varies the structure that holds it.

The evaluation protocol. Staging non-reportable situations measures precision instead of assuming it, abstinence is represented as a verdict rather than forced into a report-or-ignore choice, and localisation and report-worthiness are scored apart.

Several limitations bound what these numbers show. The ground truth was written by one person, who also designed the pipeline and staged the scenes, so there is no second annotator or agreement measure. The labels derive from the shoot script rather than the system output, but the norms, scenes and scoring are still not independent of each other. The development-site figures provide a baseline for the held-out result, not evidence of generalisation in their own right (Appendix L.2). The judge is stochastic, with run-to-run variation of up to about five verdicts per patrol, so every arm is repeated 3 times. The staged patrols also contain an unusually high density of norm-relevant changes. On an ordinary patrol, reportable events will likely be rarer, so the same number of false alerts would result in lower alert-level precision. The violations were staged in the open, so the measured recall says little about cases that are subtle or deliberately concealed. The categories also mix two differences: how specific the rule is, and whether judging it requires only the detected entity or also a relation to something else. Because these vary together in this dataset, we cannot tell which of them causes

norm kind	what the design must supply	shown by
explicit norm	a view good enough to resolve the observable, and a describer that names it	explicit recall is unchanged by fusion and falls with elaboration; it also falls when the describer is norm-blind (Figure 3b,c)
relational norm	the second term the norm names, an area, a route, an owner, an earlier moment: free from the image where it is visible, measured and injected where it is not	removing the person and zone streams costs every category, hiding the map alone silences the zone norms alone (Figure 3a); the widest lead over a camera-only system (Table 2)
implicit norm	detailed interpretive cues, <i>and</i> for disorder a stage that states the object’s condition	the only category whose recall rises with wording detail, and the only one whose recall falls under fusion (Figure 3b,c)

Table 4. What each kind of norm requires of the design, read across the evaluation experiments, with the experiment that shows each requirement.

the observed category differences.

What the design is built not to notice. The change filter sets the perceptual boundary, since a candidate exists only where occupancy differs from the reference. An oil stain or a light left burning produces nothing to evaluate, and no improvement to the judge recovers it. Range and occlusion narrow that boundary further, as Appendix H sets out.

The limits so far are perceptual; beyond them lie the scope limits. Authoring a norm set is human labour this work neither eliminated nor measured, and the two sites needed different sets. A surveillance norm encodes what matters at its site, so a norm set is site-specific by construction. It was not held out here: the deployed wording was revised after the first runs at the site (Appendix L.1). Its influence is measured instead, by the wording ladder of Section V-D. The reference map is built once and treated as current, which is benign with both recordings made the same afternoon but not for a reference that ages. Illumination, weather and season were excluded. Nothing ran online, so latency and the supervisor’s response to the stream are both untested.

VII. Conclusion

This thesis set out to decide what on a patrol is worth reporting, not merely to detect that something changed. It does so with a hybrid geometry–VLM pipeline. LiDAR change detection and person tracking feed one evidence graph, which links what changed to the people and areas around it and selects the images the judge is shown. A vision-language model then characterises each unit and evaluates it against the site’s norms in one call. On a site staged and recorded before any norm result existed, the system reaches an F_1 of 0.70 at the norm layer. The geometry operating point underneath that figure was tuned to the same site. Two comparisons locate where that performance comes from. Removing the relational layer reduces aggregate F_1 by 0.17, with recall declining in every norm category, while removing only the site map causes the area-bound norms to fail and leaves the others unchanged. A naive camera-only model given the same recordings and

the same norms comes within one event of the pipeline’s recall at an operator-level precision of 0.08 against 0.66. What the design buys is therefore comparable event recall with far fewer alerts and higher alert precision.

The pipeline produces a deduplicated stream of alerts. Each alert identifies the violated norm and the entity it concerns, and carries the evidence used to reach that decision.

The design requirements differ across the three norm categories. Explicit norms require the relevant observable to be localised, resolved and described. Implicit norms benefit from more detailed cues and from a visual characterisation that takes the norm into account. Relational norms additionally require the second term of the relation, taken from the image where visible and supplied as measured evidence where it is not. These requirements constrain the design, but the experiments do not show that the evidence graph used here is the only way to meet them. More generally, a norm can only be judged when the deciding fact is available to the system.

The resulting system supports a supervisor rather than replacing one: it reduces a patrol to a small set of evidence-backed alerts and leaves unresolved cases visible. Future work should separate the properties that are combined in the present norm categories, and test cases in which the deciding evidence is less clearly observable or extends over longer periods. Bringing the system to a new site needs a norm set, a zone map and a reference patrol.

References

- [1] Huib Aldewereld, Frank Dignum, Andrés García-Camino, Pablo Noriega, Juan A. Rodríguez-Aguilar, and Carles Sierra. Operationalisation of norms for electronic institutions. In *Coordination, Organizations, Institutions, and Norms in Agent Systems II (COIN 2006)*, volume 4386 of *Lecture Notes in Computer Science*, pages 163–176. Springer, 2007. 2
- [2] Henrik Andreasson, Martin Magnusson, and Achim J. Lilienthal. Has something changed here? Autonomous difference detection for security patrol robots. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3429–3435, 2007. 2
- [3] Anthropic. Claude (sonnet 4.6). <https://www.anthropic.com/claude>, 2026. Large language / vision-language model; used as the change interpreter and norm judge. 4
- [4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Uprocroft. Simple online and realtime tracking. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016. 3
- [5] Cristina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge, 2006. 2
- [6] Guido Boella, Leendert van der Torre, and Harko Verhagen. Introduction to normative multiagent systems. *Computational and Mathematical Organization Theory*, 12(2–3):71–79, 2006. 2
- [7] Lukas Brunke, Yanni Zhang, Ralf Römer, Jack Naimier, Nikola Staykov, Siqi Zhou, and Angela P. Schoellig. Semantically Safe Robot Manipulation: From Semantic Scene Understanding to Motion Safeguards. *IEEE Robotics and Automation Letters*, 10(5):4810–4817, May 2025. 2
- [8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 3
- [9] Maria Cvach. *Monitor Alarm Fatigue : An Integrative Review*. *Biomedical Instrumentation & Technology*, 46(4):268–277, July 2012. 1
- [10] Fiona M. Donald, Craig H. M. Donald, and Andrew Thatcher. Work exposure and vigilance decrements in closed circuit television surveillance. *Applied Ergonomics*, 47:220–228, 2015. 1
- [11] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa Nesnas, and Marco Pavone. Semantic Anomaly Detection with Large Language Models. *Autonomous Robots*, 47(8):1035–1055, 2023. 2, 3
- [12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996. 3
- [13] European Parliament and Council of the European Union. Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC. Official Journal of the European Union, L 333, 27.12.2022, pp. 164–198, 2022. 1
- [14] Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald Dansereau, Niko Sünderhauf, and Dimity Miller. Multi-view pose-agnostic change localization with zero labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2
- [15] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Concept-Graphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028, 2024. 2
- [16] Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation. *IEEE Robotics and Automation Letters*, 9(10):8298–8305, Oct. 2024. arXiv:2403.08605 [cs]. 2
- [17] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 788–801, 2008. 3
- [18] Jessica Huynh, Alfredo Gomez, Athiya Deviyani, Renee Shelby, Jeffrey P. Bigham, and Fernando Diaz. Quantifying the statistical effect of rubric modifications on human-automator agreement. arXiv:2605.06283, 2026. 3
- [19] Félix Iglesias Vázquez, Henrique O. Marques, Arthur Zimek, and Tanja Zseby. What do anomaly scores actually mean? Dynamic characteristics beyond accuracy. *Data Mining and Knowledge Discovery*, 39(1):2, Jan. 2025. 2
- [20] Fatemeh Jalalvand, Mohan Baruwat Chhetri, Surya Nepal, and Cécile Paris. Alert prioritisation in security operations centres: A systematic survey on criteria and methods. *ACM Computing Surveys*, 57(2):1–36, 2024. 4
- [21] Giseop Kim and Ayoung Kim. Remove, then revert: Static point cloud map construction using multiresolution range images. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2
- [22] Alexander Krawciw, Jordy Sehn, and Timothy D. Barfoot. Change of scenery: Unsupervised LiDAR change detection for mobile robots. In *2024 Conference on Robots and Vision (CRV)*, 2024. 1, 2
- [23] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 3, 14
- [24] Yu-Chen Lai, Motoharu Sonogashira, Itthisak Phueaksri, and Yasutomo Kawanishi. Scene-Specific Anomalous Relationship Detection Using Scene Graph Summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3995–4003, 2025. 2
- [25] Xingchen Li, Long Chen, Jian Shao, Shaoning Xiao, Songyang Zhang, and Jun Xiao. Rethinking the evaluation of unbiased scene graph generation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 2
- [26] Hyungtae Lim, Sungwon Hwang, and Hyun Myung. ERA-SOR: Egocentric ratio of pseudo occupancy-based dynamic

- object removal for static 3D point cloud map building. *IEEE Robotics and Automation Letters*, 2021. 2
- [27] Ruiping Liu, Junwei Zheng, Yufan Chen, Zirui Wang, Kunyu Peng, Kailun Yang, Jiaming Zhang, Marc Pollefeys, and Rainer Stiefelhagen. Situat3dchange: Situated 3D change understanding dataset for multimodal large language model. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. 2
- [28] Wandeng Mao, Liang Jiang, Shanfeng Liu, Shengzhe Xi, and Hua Bao. A long-term localization and mapping system for autonomous inspection robots in large-scale environments using 3D LiDAR sensors. *PLOS ONE*, 20(7):e0328169, 2025. 1
- [29] OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. 1
- [30] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):1–38, Mar. 2022. arXiv:2007.02500 [cs]. 2
- [31] Yue Qiu, Shintaro Yamamoto, Ryosuke Yamada, Ryota Suzuki, Hirokatsu Kataoka, Kenji Iwata, and Yutaka Satoh. 3D change localization and captioning from dynamic scans of indoor scenes. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1176–1185, 2023. 2
- [32] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. A Survey of Single-Scene Video Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2293–2312, 2022. 2
- [33] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. 3
- [34] John R. Searle. *The Construction of Social Reality*. Free Press, New York, 1995. 2
- [35] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11953–11963, 2023. 17, 21
- [36] Shahroz Tariq, Mohan Baruwat Chhetri, Surya Nepal, and Cécile Paris. Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Computing Surveys*, 57(9):1–38, 2025. 1, 4
- [37] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094, 2024. 3, 4
- [38] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V, 2023. arXiv preprint. 21
- [39] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. 21
- [40] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3
- [41] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 3, 4
- [42] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 3
- [43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. 2
- [44] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*, 2024. 1

A. Pipeline variables

Every threshold the body defers to is collected in Table 5 at its KW3 value, with a note on where that value came from. Both sites run the same mechanism set.

Table 5. Pipeline variables at KW3, the held-out site, with the value the reported runs used and where it came from. A star marks a value a sweep set. Where Maaldrift, the development site, uses a different value, the note gives it.

Quantity	Value	Provenance
<i>World map construction</i>		
Both patrols are accumulated identically, so that a difference between the two maps is a difference in the world rather than in how it was measured. The min-hits terms set how often a voxel must be struck before it counts as occupied. The range-aware terms relax that bar with distance, since returns thin out as $1/d^2$ and a flat bar would empty the far field.		
δ , voxel edge length (m)	0.1	fixed
ρ , min-hits ratio of scans	0.1	set once and not swept; the maps are more sensitive to it than to any other term here
d_0 , range-aware min-hits knee (m)	4	calibrated on the second patrol
p , range-aware exponent	2	not fitted; the exponent that matches the $1/d^2$ falloff
<i>Change detection</i>		
A voxel occupied in exactly one map is a candidate change, and these decide which candidates survive. T absorbs the registration error between two separately driven patrols, and the ground floor discards candidates below the ground plane before they are clustered. E, S and V separate a cluster from speckle. COV, ASYM, FAR and ZMIN are the observability gates of Section III-A. COV and ASYM keep only clusters both patrols passed at comparable range, so that a difference in coverage is not read as a change. FAR drops distant removals: at long range a one-sided observation manufactures a removal, while the changes actually staged far out are additions. ZMIN excludes clusters floating above the ground.		
T, change distance threshold (m)	0.15*	set jointly with S
ground floor for change candidates (m)	0*	drops sub-ground returns
E, DBSCAN ε (m)	0.2*	
S, DBSCAN min-samples	10*	the false-positive lever; 8 at Maaldrift
V, min voxels per cluster	10*	12 at Maaldrift
COV, coverage radius from path (m)	25	scaled to sensor range; 15 at Maaldrift
ASYM, co-visibility asymmetry bound (m)	1.5*	0.75 at Maaldrift
ZMIN, max cluster base height (m)	0.5*	floating-base gate
elongation ratio	5*	thin wall strips
elongation min extent (m)	0.3	paired with the ratio
FAR, far-field filter radius (m)	10*	drops far removals only
<i>Person detection and tracking</i>		
The chain is detect, associate, re-link, confirm. The score floors trade people missed against structures mistaken for people, as does T_{ref} , which discards any detection sitting on structure the reference map already holds. D_{assoc} and G_{scan} bound scan-to-scan association, and a gap longer than G_{scan} ends a fragment. G, D and v_{max} are the three bounds of Equation (1), which decide when two fragments are one person. L_{min} is what separates a person who stood still from an isolated false detection.		
detection score floor	0.3	a lower floor adds false tracks
track association score floor	0.3	0.5 at Maaldrift
T_{ref} , reference-map exclusion radius (m)	0.18	static structure is not a person
D_{assoc} , association radius (m)	1.5	fixed
G_{scan} , max missed scans before a fragment ends	20	fixed
G, re-link time window (s)	14	covers a 12 s occlusion dropout
D, max re-link distance (m)	4	Equation (1)
v_{max} , max implied re-link velocity (m/s)	3.5	plausibility gate
L_{min} , super-track confirmation duration (s)	3	confirmation floor
<i>Evidence graph</i>		
The graph holds the two streams and the injected zone map. The radii and the dwell bound decide when a measured relation is written as an edge at all, and are set so that an edge means the two units were genuinely together rather than momentarily near. GAP is the one gate that changes the unit count rather than the evidence: inside it, a removed and an added cluster are one object that moved, and are judged once.		
R_{prox} , proximity edge radius (m)	2	fixed
R_{tt} , track-track edge radius (m)	6.5	fixed
Δt_{min} , minimum edge dwell (s)	0.5	fixed
GAP, pairing bbox gap (m)	0.3	fuses a removed/added pair into one unit
<i>View selection and marking</i>		
The four weights and d_{max} are the terms of Equation (2), the cost that picks which recorded moment is shown for a unit or an edge: how centred the target is, how far the robot was, how much of the line of sight is blocked, and how much of the target falls inside the frame. The two pixel floors keep the drawn rectangle legible when the target is small, which is the case the model otherwise overlooks.		
α , centredness weight	1	Equation (2)
β , range weight	0.5	Equation (2)
γ , occlusion weight	1.5	Equation (2)
w , in-frame coverage weight	1	Equation (2)
d_{max} , view-scoring range limit (m)	12	Equation (2)

Table 5. Pipeline variables at KW3 (*continued*).

Quantity	Value	Provenance
bbox padding (px)	40	legibility floor
bbox minimum side (px)	80	legibility floor

B. Mechanism details

B.1. Person detection, tracking, and stripping

Each LiDAR scan passes through CenterPoint, and its person detections are placed in the world frame through the robot’s odometry. Association then runs scan by scan: a Hungarian assignment [23] on world-frame distance matches detections to the open tracks, and a track survives a short run of scans on which the detector does not fire. A longer occlusion breaks a path into fragments, which a second stage re-links. A fragment ending at position p_i and time t_i is chained to a later one starting at p_j, t_j when

$$t_j - t_i \leq G \wedge \|p_j - p_i\| \leq D \wedge \frac{\|p_j - p_i\|}{t_j - t_i} \leq v_{\max}, \quad (1)$$

that is, when the gap is short, the jump is small, and the implied velocity is physically plausible. This does not guarantee exactly one track per person. It collapses the many short fragments a frame-to-frame tracker produces into a few coherent trajectories. A track is confirmed once it has lasted a minimum duration or produced a stationary burst of observations in one place.

Stripping deletes the points inside a confirmed track’s footprint from every scan it spans, and on the scans where the detector did not fire the footprint position is interpolated from the detections on either side, for the reason Appendix C shows. The cleaned clouds are the ones accumulated into the world map. Every bound named here is a row of Table 5.

B.2. View selection and marking

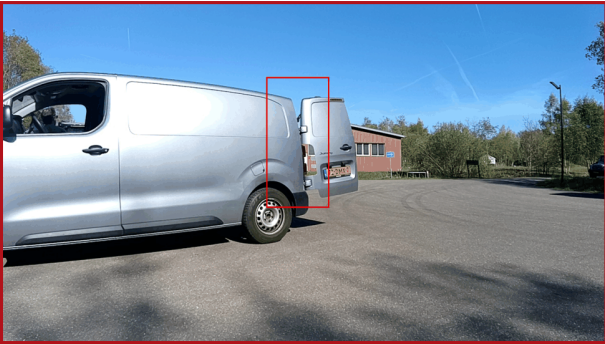
Every pair of robot pose and camera recorded within range of a target is a candidate view. Each candidate is scored by the cost

$$s = \alpha c + \beta \frac{d}{d_{\max}} + \gamma o + w(1 - f), \quad (2)$$

where c is the offset of the projection from the image centre, d the range, o the occluded fraction and f the fraction of the target’s projected extent inside the image. The occluded fraction is the share of rays from the robot to points sampled across the target’s bounding box that are stopped by occupied map voxels, marched through a horizontal slice of the accumulated map. Sampling the box rather than its centre matters, since a single ray can pass through a gap while most of the object is hidden (Figure 5). The last term penalises a view that cuts the target off at the frame edge. The chosen view is the arg min over candidates. Where two targets must appear together the cost is summed over both, and a moment qualifies only if each of them scores at all. The reference view is chosen by the same cost over the reference patrol, restricted to the camera the current view used, so that the two are comparable. The occlusion term changes what is reported and not only what is seen: on the change of Figure 5, from the view it rejects the describer reported a second van parked behind the actual van and the change was filed as routine, while from the view it chooses the describer reported the open rear door and N3 fired.

The rectangle that marks a change is drawn around the cluster’s own voxels projected into the image. It is padded and held to a minimum size, so a small object remains legible.

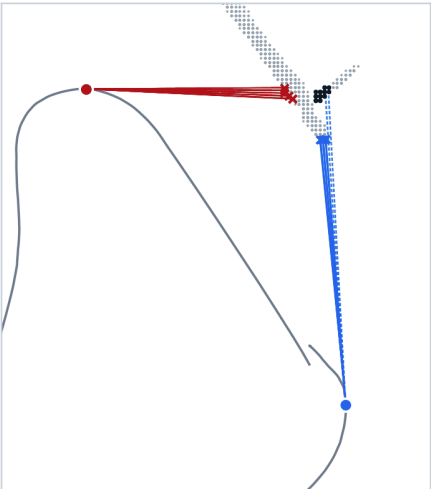
change cluster 5 · development site



rejected best view on centredness and range alone



chosen best view once occlusion is scored



plan view of the same two candidates

- ray blocked, x marks where
- - - ray reaching the change
- the change itself
- accumulated map, height band

	camera	range	occluded	score
rejected	back	5.6 m	100%	1.80
chosen	front	7.4 m	60%	1.24

Figure 5. The effect of the occlusion term. The left view is the pick with γ set to zero, the right the pick with γ at its set value, and the plan view gives the rays cast from each of the two poses.

C. Stripping the gap frames

The gaps the detector leaves inside a confirmed track are frequent and short, so a strip applied only where it fired would leave most of the trail behind. Figure 6 shows both kinds of sweep along one person’s path.

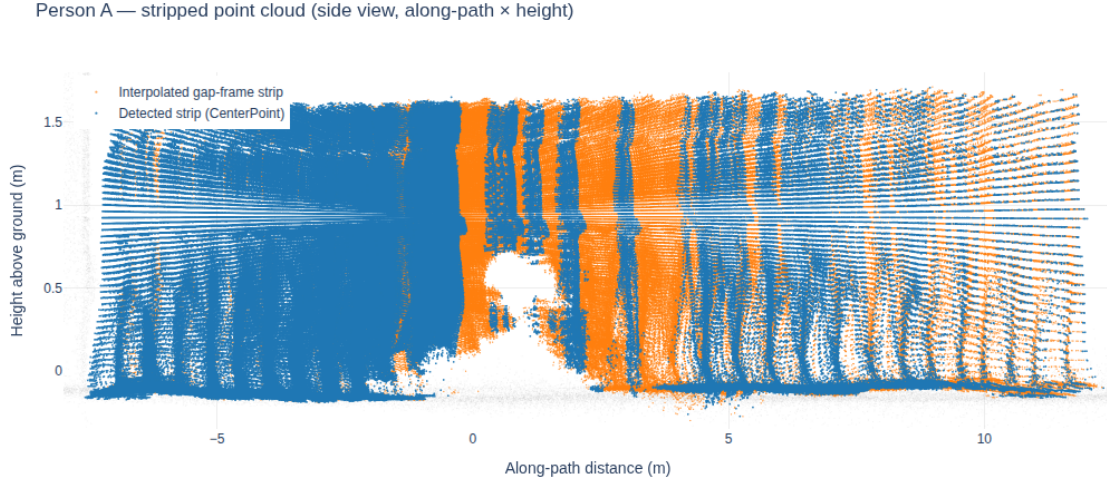


Figure 6. The points removed for one person along their own path, under the second stripping condition of Section III-B. Each near-vertical stripe is one LiDAR sweep. Blue stripes are sweeps on which the detector fired, whose points are removed directly. Orange stripes are sweeps on which it did not, whose strip position is interpolated from the detections on either side.

D. Evidence-graph topology

Figure 7 sketches how the nodes and edges of Section III-C are formed, in the notation Figure 8 uses.

The graph carries no semantics. The change cluster in Figure 7 is in fact a bag, and the graph records it as added; the object label is assigned later, by the vision-language model at the interpretation stage. The zone is the one element that comes from neither stream. It is injected from the site map.

Figure 8 gives one patrol’s graph in full, from the development site, chosen because a whole graph still fits legibly on a page there. It has three tiers. The injected zones sit at the top and are context nodes, which never receive a verdict. The tracked people sit in the middle. The unit nodes sit below, and each one receives exactly one verdict. Most units, in the band along the bottom of the figure, have no relations at all and are judged on their own evidence.

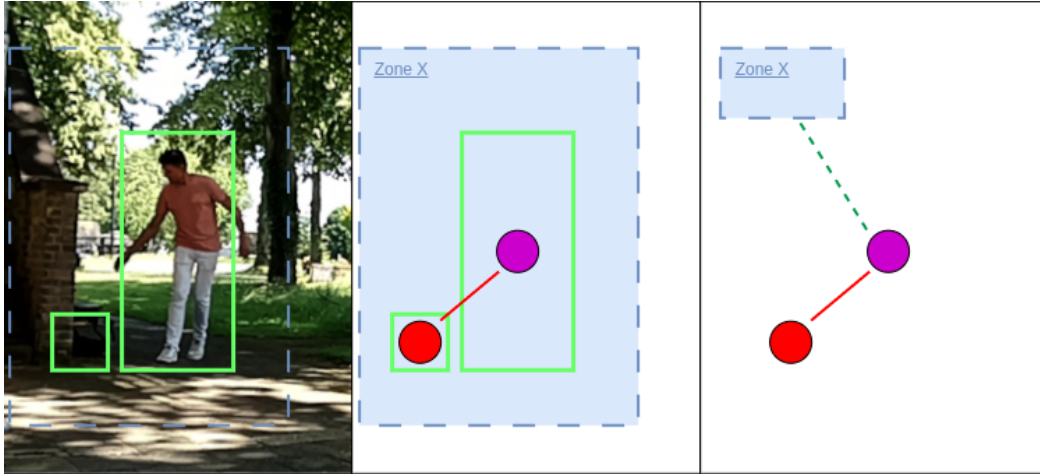


Figure 7. One star of the graph being formed, from the frame it starts in. The three panels are one scene drawn three ways, not three stages the method runs in that order. Left, a patrol frame with what the two streams recovered from it, boxed in green: a tracked person, from the pedestrian detector, and a cluster the LiDAR comparison marked as added. Both are computed in the point cloud and projected into the image for display, and neither is detected in the image. The dashed region is a zone. Middle, the same scene with the image dropped, each detection collapsing to a node, the person purple and the change cluster red. Right, what the judge is handed, nodes and typed edges alone. The edges are drawn as relations only; the quantities they hold, a separation and the times a relation began and ended, are not shown.

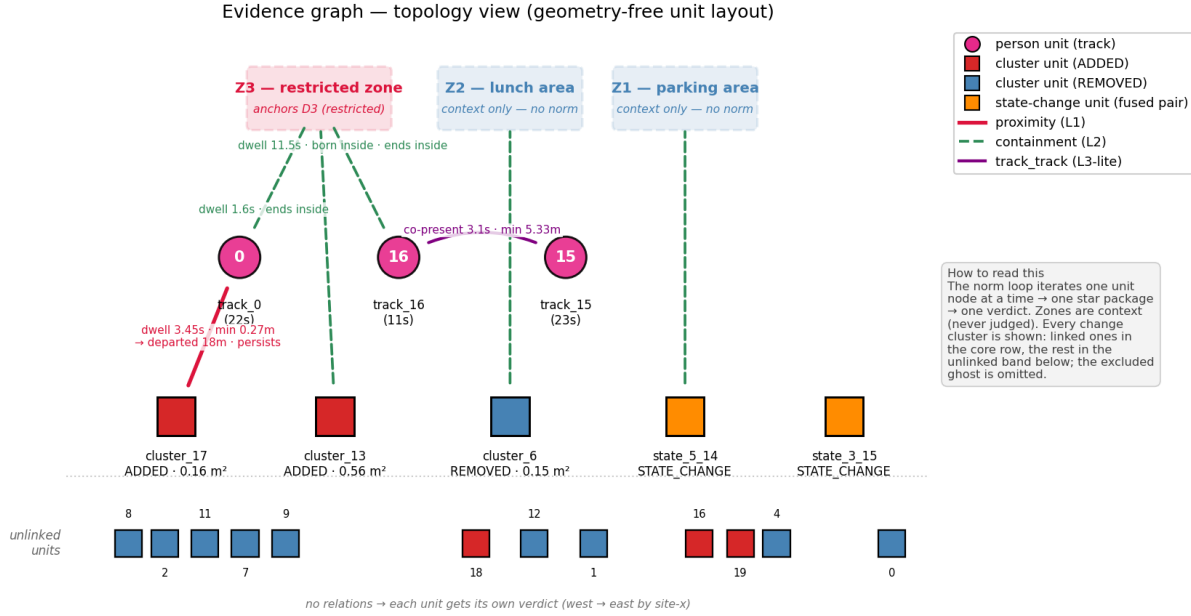


Figure 8. One patrol’s evidence graph in full, in the three tiers described above. Node colour gives the change type. An orange node is a removed cluster and an added cluster that a pairing edge has already fused into one state-change unit. Edges are annotated with the quantities they hold: how close a person came to a change and for how long, whether they then departed and whether the object remained, how long a unit lay inside a zone, and how long two people were within range of each other.

E. Evidence assembly

Evidence assembly turns a unit and its relations into the images the judge sees. The graph has already settled which units and relations are shown. Two failures follow from the attention problem of Section III-C: the view may not see the change at all, because something stands between the robot and it, or it sees the change while the model attends elsewhere. Geometry addresses both: the change’s position in three dimensions and the camera poses are known throughout the patrol.

What has to be shown differs by unit and by relation, and the graph fixes the target and the time window in each case. A change asks for a view of the current patrol and one of the reference. A person’s track asks for three views spread across the span of their presence (Figure 12). A proximity edge asks for a single frame holding both parties. A containment edge asks instead for a bird’s-eye rendering of the unit inside its zone (Figure 11). It is the only case rendered that way, because a zone’s extent and the rule it carries come from the injected map, which a photo can miss.

Choosing the frame is then one operation in every case: the chosen view is the one minimising the cost of Appendix B.2.

Marking comes second. The rectangle of Appendix B.2 directs the model to the target inside the chosen view [35]. The frame is left intact rather than cropped or blurred, because our norms turn on the surroundings. Whether an object blocks a door or sits inside a restricted area is read off the neighbouring objects and their state (Appendix I).

The package the judge receives, for one unit, is that unit’s own evidence together with its one-hop neighbourhood in the graph. Each is given in pictures and in numbers: the frame or bird’s-eye rendering the unit or edge asked for, and the geometry behind it, a size, a height, a separation, the times a relation began and ended. The package stops at one step. A relation two steps away is about some other unit, not this one. One package is shown as the judge receives it in Appendix N, and four worked cases in Appendix I.

F. Norm sets

Below are the norms, written in the deployed rung; for the terse and elaborated rung, see Appendix K. Each site has its own set, and no identifier is compared across sites. The category of each norm is the one shown in Figure 2.

F.1. Test site (KW3)

- N1. Unattended bag** (relational norm). Report an unattended bag or backpack left with no owner nearby in an operational area. This covers bags and backpacks only, not boxes, crates, or equipment. A bag left in the picnic or break area is routine and is not reported.
- N2. Blocking access** (relational norm). Report an object placed in front of or blocking an entryway or access route, a door, gate, passage, or ramp, such that it would hinder entry or passage. Judge hindrance of the route, not the presence of an

object near it: an object set aside in open space, or one that a person can easily pass or walk around, is not blocking and is not reported.

- N3.** *Open vehicle compartment* (explicit norm). Report an open compartment, door, or hatch on a parked vehicle when no one is attending it. If a person is at or beside the vehicle, loading or using it, the open compartment is routine and is not reported.
- N4.** *Disorder / disturbance* (implicit norm). Report signs of disorder or disturbance: an object toppled, overturned, knocked over, or scattered out of its proper state. The cue is the disturbed state, not mere presence: an object that has simply appeared or been repositioned but sits upright and intact is not disorder unless it is clearly out of order in relation to its surroundings.
- N5.** *Missing safety equipment* (explicit norm). Report safety equipment, in particular a hand-held fire extinguisher, missing from its mounting or usual place. The cue is a conspicuously empty bracket or stand where the equipment used to be.
- N6.** *In restricted zone* (relational norm). Report a person or object located inside the restricted zone, the area behind the cones. The cones or markers that delineate the boundary are not themselves reportable as objects in the zone.
- N7.** *Suspicious behaviour / attempted access* (implicit norm). Report behaviour or traces suggesting someone is or was attempting unauthorized access, or otherwise behaving suspiciously. Judge the intent the scene implies as a whole. Signs include a person climbing at a building, objects positioned as a climbing aid, a concealed identity, or clothing unfitting the setting. Equipment merely present or in use for ordinary maintenance does not qualify on its own.
- N8.** *Non-vehicle in car park* (relational norm). Report a non-vehicle object placed or left in a parking area, which is for vehicles. This concerns objects; ordinary use of the car park by vehicles is not reportable.
- N9.** *Picnic table moved* (explicit norm). Report a picnic table moved, displaced, or removed from its usual position, in any zone. This concerns the table itself as a fixed asset, distinct from personal items left near it.
- D1.** *Badge / identifiability* (explicit norm). Every person in an operational area is expected to wear a visible identification badge, on a lanyard at the neck or a cord at the waist. Report a person with a concealed identity, a mask or balaclava, or no visible badge where one is expected. If badge visibility cannot be determined from the available view, because the person is too far to resolve it or faces away the whole time, abstain and request a closer frame rather than guess.
- D2.** *Loitering / dwell* (implicit norm). Report a person loitering, that is seemingly doing nothing, leaning, or sitting for an extended period in an operational area. Dwelling in the picnic or break area is routine and is not reported. Brief pauses and ordinary transit are not loitering.
- D3.** *Aggression / pushing* (implicit norm). Report aggressive physical interaction between people: pushing, shoving, grappling, or fighting. Ordinary close contact such as greeting, talking, or handing something over is not aggression.
- N0.** *Routine (non-reportable)* (non-reportable). Routine operational activity is not reported: vehicles arriving or leaving, bins or boxes repositioned during logistics, personal items removed, and people moving through or resting in permitted areas. When a change has an ordinary explanation and no reportable norm is met, presume routine.

F.2. Development site (Maaldrift)

- N1.** (relational norm). Report an unattended bag or backpack left with no owner nearby in an operational area. This covers bags and backpacks only, not boxes, crates, or equipment. A bag left in the picnic or break area is routine and is not reported.
- N2.** (relational norm). Report an object placed in front of or blocking an entryway or access route, a door, gate, passage, or ramp, such that it would hinder entry or passage. Judge hindrance of the route, not the presence of an object near it: an object set aside in open space, or one that a person can easily pass or walk around, is not blocking and is not reported.
- N3.** (explicit norm). Report an open compartment, door, or hatch on a parked vehicle when no one is attending it. If a person is at or beside the vehicle, loading or using it, the open compartment is routine and is not reported.
- N4.** (implicit norm). Report signs of disorder or disturbance: an object toppled, overturned, knocked over, or scattered out of its proper state. The cue is the disturbed state, not mere presence: an object that has simply appeared or been repositioned but sits upright and intact is not disorder unless it is clearly out of order in relation to its surroundings.
- N0.** (non-reportable). Routine operational activity is not reported: vehicles arriving or leaving, bins or boxes repositioned during logistics, personal items removed, and people moving through or resting in permitted areas. When a change has an ordinary explanation and no reportable norm is met, presume routine.

F.3. Never-fire probe

Six norms added to the KW3 set for the negative control of Appendix L.4. No staged scene satisfies any of them, so every alert citing one is a false positive by construction. Each turns on a predicate that is absent at the site, so refusing it takes a judgement rather than a missing word. The norms above are unchanged in this set and carry their deployed (L1) text.

- N11.** *Vehicle damage* (explicit norm). Report a vehicle that shows damage: a broken or shattered window, a smashed light, a deflated tyre, or a dented or scraped panel. An open door, boot or hatch is a vehicle state and not damage, and a dirty or weathered vehicle is not damaged.
- N12.** *Work equipment in the break area* (relational norm). Report tools, work equipment, or building materials left in the picnic or break area: a ladder, a toolbox, a bucket, cabling, or timber standing in a space meant for breaks. The tables and benches themselves, and personal items such as bags, are not reported.
- N13.** *Vandalism* (implicit norm). Report vandalism: graffiti, defacement, breakage, or other deliberate damage to buildings, signage, fencing, or site equipment. An object that has been knocked over, moved, or left out of place but is undamaged is not vandalism.
- D4.** *Weapon carried* (explicit norm). Report a person openly carrying a weapon in their hands: a firearm, a knife, or a club. Tools, sports equipment and ordinary carried items such as bags, boxes or a ladder are not weapons, and equipment worn as part of a uniform or duty kit is not reported. If an object a person carries cannot be identified, it is not reported as a weapon.
- D5.** *Tampering with a vehicle* (relational norm). Report a person tampering with a parked vehicle: trying or forcing a door or boot, reaching into a compartment they have opened, or taking something from inside it. A person walking past a vehicle, standing beside it, or loading it in the ordinary way is not tampering.
- D6.** *Recording the site* (implicit norm). Report a person recording or documenting the site: photographing or filming buildings, fences, gates or security equipment, or sketching or writing down details of the layout. A person handling a phone or another item without directing it at the site is not reported.

G. Ground-truth protocol

Ground truth is annotated in three layers, each in a separate browser labeller. All of it was annotated by the author after the recordings, from the shoot script and the recorded frames. Figure 9 shows the annotated world in overview for both held-out patrols: the reference map, the two routes, the person tracks, the zone polygons and the change clusters the geometry layer kept.

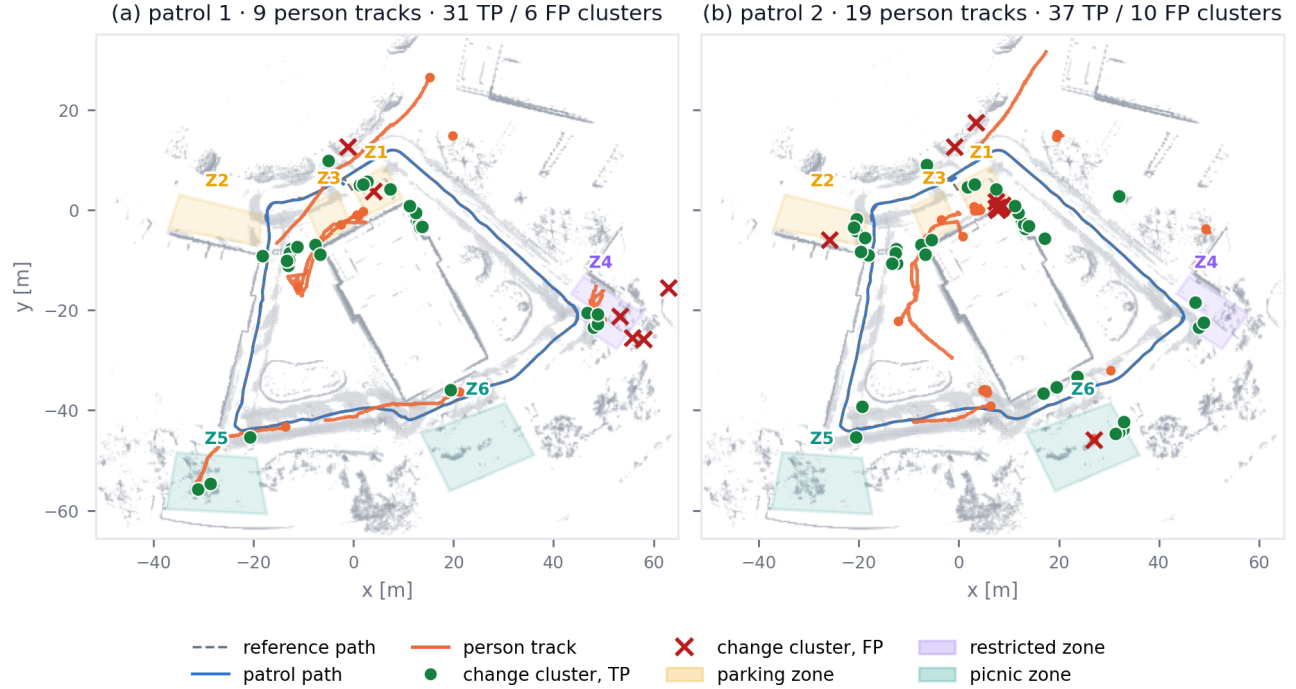


Figure 9. The held-out site in overview, one panel per patrol. The grey shading is the reference world map, rendered as point density. The dashed line is the reference route and the solid blue line the patrol route. Orange polylines are the confirmed person tracks, each with a dot at its first observation. Discs and crosses are the change clusters the geometry layer kept at the locked operating point, marked true or false positive against the localisation ground truth: at cluster level 31/6 on patrol 1 and 37/10 on patrol 2, and the 16 false positives here are the ones the geometry headline counts. The tinted polygons are the semantic zones supplied per site, parking (Z1 to Z3), restricted (Z4) and picnic (Z5, Z6).

Localisation. Each staged change is drawn as an oriented rectangle with a height band over the reference map. The box is placed on the object, never on a candidate. An object that moved or changed state is recorded as one event carrying two boxes, its before and its after position, and either box credits it. A second hit on the same event is absorbed rather than counted twice.

Report-worthiness. Each event is labelled with the norm or norms it violates. A norm is marked *uncertain* on an event when the evidence available from the route cannot settle it. Scoring is at the event level throughout: an event unions the boxes and the norms of the rows that constitute it.

People. People are annotated as presence polygons carrying a person identifier. A person norm is credited when a track enters the polygon the norm is anchored to. The polygons describe the staged reality, not what the perception layer saw.

The camera-only baseline. The baseline of Section V-B fires independently at poses sampled every 1.5 m along the route, and its firings are then scored through the accounting above, like-for-like on metric code, model and norms. Its relational deficit is entirely zone-bound: on the 3 zone-bound events it scores 1 against the pipeline’s 3, and the other 6 tie at 4 each.

H. Error analysis

Every deterministic miss, by cause. The 7 events missed in all 3 runs are attributed in Table 6.

cause	events	reachable by
the norm names no threshold, and nothing in the evidence supplies one	3	neither
the change was never localised by the geometry layer	1	neither
the view was selected before the change existed	1	the frame selector
an exclusion clause belonging to another norm was applied	1	the norm text
the change was seen clearly and read as unchanged	1	neither

Table 6. The 7 events missed in all 3 runs, by cause. The first row is the dwell norm D2, which is never cited at all; the second is the unattended bag of Section V-A. The last three are disorder under N4, and are also the 3 events on which the camera-only baseline of Section V-B succeeds where the pipeline does not. Figure 13 reproduces the last of them.

Firing at every viewpoint gives the baseline later moments, where the pipeline’s frame selector had already settled on a view from before the change, and whole scenes, where the pipeline’s evidence was cropped to one vacated position. On the third neither system had an advantage, and only the pipeline read a toppled bin as upright. A further 4 events are caught in some runs and missed in others: 3 badge events and the metal ramp under N2.

Abstention calibration. Abstention was designed for one norm. All 14 uncertain marks are the badge rule D1, which needs a frontal view that range and pose routinely deny. Scored by Appendix G, the judge marks it uncertain on 11 of them, stays silent on 3, and commits firmly on 0. An over-claim reaches the supervisor as a violation the evidence does not support and is charged a false positive; a missed abstention is silence about a norm the ground truth marks unsettled. This evaluation produced the second kind only.

The boundary no run reaches. The same threshold that defines a candidate sets a size floor at the voxel edge (Table 5). Range compounds it, because people are detected dependably over a shorter distance than the map reaches, which leaves an outer band where they are neither tracked nor stripped: lost as norm subjects, and smeared into the map as spurious change. Occlusion breaks identity. A person passing behind a vehicle leaves no detections to associate, so one person becomes two tracks or none, and anything defined over duration loses its subject.

I. Worked alerts

Four cases are reproduced as the system produced them, three that were reported and one that was not. Each card carries the camera views the norm layer was given, with the projected change region drawn on them. Underneath are what the geometry layer measured, the model’s own characterisation of the scene, its verdict and the norm it cited, and the ground-truth event. The true-positive and missed labels come from the scorer that produced Table 1. Characterisation and verdict come from a single model response, since the reported arm fuses the two calls.

Two stronger ways of forcing attention were considered and rejected for the reason Appendix E gives. Cropping to the change discards the surroundings outright [35]. Blurring outside the box keeps their layout but not their legibility, and is reported to direct a contrastive model better than a drawn marking does [39]. Marking is corroborated on a generative model too [38].



reference patrol · left camera, 3.0 m



current patrol · left camera, 2.9 m

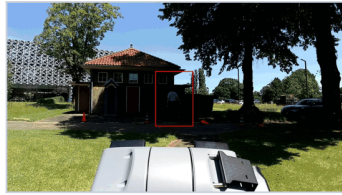
GEOMETRY	change cluster 30, added; 1.04 m ² footprint over 78 voxels; chosen view 24% occluded.
SEMANTICS <i>high confidence</i>	“bicycle”. A bicycle has been added in front of a large wooden double door between the reference and current patrol. The reference frame shows the doorstep area clear; the current frame shows a black upright bicycle parked on the step directly in front of the door.
NORM <i>report</i> <i>medium priority</i>	A bicycle has been parked directly in front of the double door entrance on the landing step, occupying the access path to the doors across nearly their full width. This constitutes an object placed in front of an entryway that would hinder entry or passage, meeting the threshold for N2. Unlike an object set aside in open space, this bicycle is centrally positioned on the doorstep and would require removal or movement to access the doors.

Ground truth: event 8, “Bike (blocking door)”, reportable under N2 blocking access; counted as a true positive in the reported run. The characterisation and the verdict are two fields of a single model response.

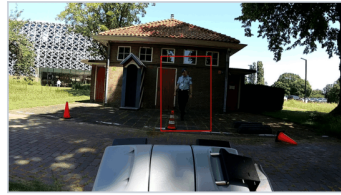
Figure 10. A reported alert, and the case the design of Section III-C is built around. The geometry layer flags the bicycle, which is the only thing that changed. The door did not change, is not a node, and no edge says “blocks the door”; the relation that makes the change reportable exists nowhere in the graph. It is nonetheless plain in the selected view, and the judge decides from the image. The judge cites the access norm and gives its reason in the norm’s own terms, hindrance of the route rather than proximity to it.

track 10 · patrol 1

N6 in restricted zone



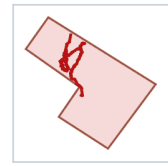
early · front camera, 11.0 m, t=142 s



mid · front camera, 7.1 m, t=161 s



late · front camera, 5.5 m, t=162 s



graph · inside Z4, 48 s

GEOMETRY person track 10, tracked 48 s over 287 detections; 6.8 m displacement at 0.14 m/s.

SEMANTICS *high confidence* Moving slowly (~ 0.1 m/s) over 48 seconds within the restricted zone; appears to be patrolling or inspecting the area in front of building 42, spending time near traffic cones. Upright, walking slowly; in mid and late frames standing/pausing near a traffic cone in front of a building entrance. No visible bag or equipment carried in hand; a lanyard/badge visible on chest in mid and late frames.

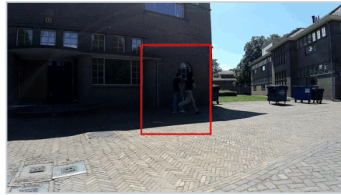
NORM *report* *low priority* The zone containment edge confirms this person was inside zone Z4 (restricted area) for the entire 48-second observation window. Under N6, any person inside the restricted zone is reportable. A visible badge/lanyard is present in the mid and late frames, so D1 is not violated. The linked static clusters (traffic cones) are attributable to their own cases.

Ground truth: event P5, "Person E (difficult). Proceeds to knock over", reportable under N6 in restricted zone; counted as a true positive in the reported run. The characterisation and the verdict are two fields of a single model response.

Figure 11. The opposite half of the same design. A person is inside the restricted area, but no photograph shows an administrative boundary. The containment edge supplies it as a measurement, forty-eight seconds inside the polygon, and the fourth panel is the bird's-eye rendering that edge selected. The verdict cites that edge by name. The three camera views span the person's presence rather than repeating one moment, and the model uses them to resolve the badge question separately.

track 0 · patrol 1

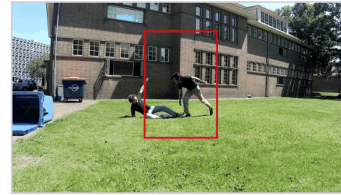
D3 aggression / pushing



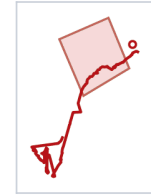
early · left camera, 8.5 m, t=3 s



mid · left camera, 16.1 m, t=18 s



late · left camera, 6.7 m, t=35 s



graph · inside Z3, 3 s

GEOMETRY person track 0, tracked 47 s over 196 detections; 14.9 m displacement at 0.32 m/s.

SEMANTICS
high confidence Initially walking through the compound near the building entrance and bin area; in the late frame engaged in what appears to be a physical altercation — one person is on the ground on the grass while track_0 stands over them in an aggressive posture, reaching/pushing toward them. Upright and mobile in early/mid frames; in the late frame the person is standing over another person who is on the ground, leaning forward aggressively with arm extended. No clearly visible carried items.

NORM
report
high priority The late frame shows track_0 standing over a downed track_5 in an aggressive forward-lunge posture with arm extended, consistent with pushing or striking; this constitutes physical aggression reportable under D3. Badge visibility cannot be resolved from the available frames (backlit, partially turned away), so D1 remains uncertain.

Ground truth: event P1, “one person pushing another to the ground” (names replaced), reportable under D3 aggression / pushing; counted as a true positive in the reported run. The characterisation and the verdict are two fields of a single model response.

Figure 12. An alert that only the temporal span makes available. The person walks through the compound for most of their track, and nothing in the earlier views is reportable; the behaviour that triggers the norm occurs late. Because a person’s evidence is three views across the span of their presence rather than a single best view, the late frame is in the package, and the verdict rests on it. The same alert also shows the abstention mechanism working beside a firm verdict: the badge norm is left uncertain, with the reason given, while the aggression norm is cited firmly.

cluster 27 · patrol 2

missed · GT N4 disorder / disturbance



reference patrol · left camera, 3.8 m



current patrol · left camera, 3.8 m

GEOMETRY change cluster 27, added; 0.36 m² footprint over 132 voxels; chosen view 0% occluded.

SEMANTICS “large blue container / crate (lidded, rectangular)”. Added: a large blue rectangular lidded container (approximately pallet-box/bulk-bin size) has appeared in the current patrol at the location previously occupied by a wheeled blue waste bin. The wheeled bin that was present in the reference frame is no longer at this exact spot.
medium
confidence

NORM The added object is a large lidded bulk container, not a bag/backpack (N1 inapplicable), not blocking a clearly defined entryway or gate (N2: it sits in an open paved yard area alongside other bins where passage around it is feasible), and is not a non-vehicle object in a designated parking area (N8). It sits upright and intact, so N4 does not apply. This appears to be a routine repositioning of logistics/waste-handling equipment in an operational courtyard, consistent with N0.
ignore
N0 cited
event missed

Ground truth: event 6, “Tipped over bin”, reportable under N4 disorder / disturbance. The change was detected, the view was chosen and the box was drawn on it; what failed was the reading of the scene. Counted as a missed event in the reported run.

Figure 13. A missed event, included because of how far the pipeline gets before it fails. The change is detected, the view is chosen at under four metres with nothing in the way, and the region is drawn on a toppled container that occupies most of the box. What fails is the reading: the model describes the container as upright and intact and concludes that the disorder norm does not apply. It is one of the three deterministic disorder misses attributed in Appendix H.

J. Every arm, aggregate rates

The body reports each experiment by the quantity its argument turns on, which for the evidence ablations is a per-category recall and for the placement study a pair of rates. This section collects the aggregate precision, recall and F_1 of every configuration run on KW3 on one basis. Every cell is on the deduplicated alert stream and the event-level recall of Section IV-C. Table 7 gives the arms that remove evidence; Table 8 gives the nine cells that vary the design instead.

arm	P	R	F_1
deployed pipeline	0.66 ± 0.02	0.74 ± 0.04	0.70 ± 0.02
map injection off	0.67 ± 0.03	0.67 ± 0.05	0.67 ± 0.03
relational stream off	0.60 ± 0.02	0.48 ± 0.04	0.53 ± 0.02
camera-only baseline	0.16	0.69	0.26

Table 7. The evidence arms of Section V-C at aggregate level, with the camera-only control of Section V-B for scale. Each cell is a mean over 3 repeats and the $\pm$ is the sample standard deviation over them. The deployed row is the configuration of Table 1; the baseline was run once, so only it carries no spread.

where the norms are read	wording	P	R	F_1
fused describer and judge	terse	0.46 ± 0.02	0.70 ± 0.02	0.56 ± 0.01
	deployed	0.66 ± 0.02	0.74 ± 0.04	0.70 ± 0.02
	elaborated	0.72 ± 0.08	0.71 ± 0.05	0.71 ± 0.06
split, norm-primed describer	terse	0.42 ± 0.02	0.72 ± 0.03	0.53 ± 0.02
	deployed	0.58 ± 0.04	0.82 ± 0.05	0.68 ± 0.04
	elaborated	0.67 ± 0.02	0.77 ± 0.02	0.72 ± 0.02
split, norm-blind describer	terse	0.41 ± 0.04	0.57 ± 0.02	0.48 ± 0.03
	deployed	0.52 ± 0.01	0.67 ± 0.05	0.58 ± 0.02
	elaborated	0.59 ± 0.07	0.64 ± 0.07	0.61 ± 0.02

Table 8. The wording $\times$ placement cube, all three rates, each a mean over 3 repeats with the sample standard deviation after the $\pm$. Rows vary where the norm text is read, the inner rows how fully it is written; the deployed system is the bold row. These are the nine cells Figure 4 plots as a precision–recall trade. The split cells are on the paired basis, so each samples describer and judge variance together.

K. Norm-wording rungs

The wording arm of Section V-D rewrites every norm at three levels of cue detail, holding each norm’s scope and its exceptions constant and leaving the worked examples empty at every level, so that the only variable is how fully the discriminative cues are spelled out. The middle rung is the deployed text of Appendix F. The two norms that drive the redistribution reported there are given in full at all three levels.

N4, Disorder / disturbance (implicit norm).

terse (L0) Report signs of disorder or disturbance in the patrol area.

deployed (L1) Report signs of disorder or disturbance: an object toppled, overturned, knocked over, or scattered out of its proper state. The cue is the disturbed state, not mere presence: an object that has simply appeared or been repositioned but sits upright and intact is not disorder unless it is clearly out of order in relation to its surroundings.

elaborated (L2) Report signs of disorder or disturbance, meaning an object left out of its proper resting state through some disturbance: a container on its side, a bin overturned or knocked over, barriers or cones knocked down and scattered, items strewn across a route. The decisive cue is the disturbed state, not presence or a small move: an object toppled or overturned is disorder, while an object that has appeared or shifted but sits upright and intact is orderly. Do not treat an ordinary object merely placed beside a path as disorderly unless it is clearly out of order relative to its surroundings.

D1, Badge / identifiability (explicit norm).

terse (L0) Report a person with no visible identification badge, or abstain if it cannot be determined.

deployed (L1) Every person in an operational area is expected to wear a visible identification badge, on a lanyard at the neck or a cord at the waist. Report a person with a concealed identity, a mask or balaclava, or no visible badge where one is expected. If badge visibility cannot be determined from the available view, because the person is too far to resolve it or faces away the whole time, abstain and request a closer frame rather than guess.

elaborated (L2) Every person in an operational area is expected to carry a visible identification badge, on a lanyard around the neck or a cord at the waist. Report a person whose identity is concealed by a mask or balaclava, or who has no visible badge where one is expected. The badge is small and resolvable only from a reasonably close, roughly frontal view, so if the person is too far to resolve it or faces away throughout the available frames, do not guess no badge: abstain, uncertain about D1, and request a closer or frontal frame. The norm applies to every person, so most distant or back-turned people resolve to abstain, not to a violation; report only a clearly concealed face or a clearly badge-less person at close frontal range.

L. Generalisation

This appendix defines the scope of the transfer evaluation, reports development-site performance, tests geometry parameters across data splits and sites, and checks for false reports under additional norms with no staged violations.

L.1. Scope of the transfer claim

The pipeline was developed at Maaldrift and evaluated at KW3. The evidence assembly, semantic-layer prompts and judge scaffolding remained unchanged for KW3. The norms and zone definitions were supplied separately for each site, and the

geometry operating point was tuned using KW3 geometry ground truth. The KW3 results therefore measure how the pipeline transfers with site-specific configuration and geometric tuning. The Maaldrift results were obtained on the development data, where the pipeline and norm wording were developed.

L.2. The development site

On the Maaldrift development bag the fused norm layer reaches a precision of 0.89, a recall of 0.89 and an F_1 of 0.89, catching 8/9 reportable events with a single false positive. These figures are evaluated using the same scoring procedure as the held-out site, so the two are directly comparable. For localisation only, the Maaldrift geometry layer achieves P 0.78, R 0.91, F_1 0.84, recalling 10/11 events. The sweep reported a higher score when the configuration was chosen; the figure given here is the one the detection pipeline itself produces. One of the alerts these figures are built from is reproduced in Figure 14, scored by the development-set scorer and excluded from the reported results.

cluster 17 · development site

N1 unattended bag, D1 electronic device



reference patrol · right camera, 2.5 m



current patrol · right camera, 2.0 m



graph · track 0 at 0.27 m, t=16 s

GEOMETRY change cluster 17, added; 0.16 m² footprint over 41 voxels; chosen view 0% occluded.

SEMANTICS “bag”. A dark-colored backpack has been placed on the ground in the marked area, which was empty pavement in the reference image. The backpack is clearly visible standing upright on its base; the person seen in the reference image is no longer present in the current frame.
high confidence

NORM The person (track_0) co-located with the backpack, then departed 18.08m away while the object persists, satisfying the abandonment criterion for N1 (unattended bag on open pavement). Additionally, track_0 is carrying a handheld flat rectangular device consistent with a tablet or remote controller while walking the operational area, triggering D1.
report
high priority

Ground truth: event 5, “bag”, reportable under N1 unattended bag, and a true positive on the development site. That site is in sample: its norms were written knowing what it contains, it has no row in any held-out table, and the numbers reported elsewhere in this document are not computed over it. The card is here for the wording of the verdict, which is stated in the quantities the proximity edge measured and is the only verdict at either site of which that is true. It is not evidence that the edge was needed: with the relational layer switched off, the same unit still reports N1 from the two frames alone, and the characterisation above, a separate call that is shown no edge and no norm, has already recorded that the person in the reference view is gone from the current one. This arm runs characterisation and verdict as two calls.

Figure 14. The mechanism the unattended-bag norm was designed around, shown working, on the development site. A person and a bag are one proximity edge apart, and the edge measures what no photograph states: closest approach of 0.27 m, then a departure of 18.08 m with the object still in place. The person in the reference patrol is the author, stationary, operating the robot. The card’s second citation, D1, belongs to an earlier development-site norm set and has no counterpart in Appendix F; that citation is reproduced as the run produced it and is scored nowhere in this document.

L.3. Geometry: folds and cross-site transfer

Geometry is in-sample on KW3 at the operating point, so its generalisation is measured two ways. The fold-out selects the operating point on one patrol and scores it blind on the other. The cross-site transfer moves a configuration unchanged between the two sites. The fold-out falls below the in-sample operating point, and recall is what it costs (Table 9; the achievable envelope and these operating points are plotted in Figure 15). The two F_1 values are close, and the in-sample point is the one reported: every extra event it recovers is one more event on which the layers above geometry can be measured (Section V-A). Cross-site transfer costs precision and leaves recall largely intact. Maaldrift’s values on the KW3 bag still find 42/50 events, two short of the 44/50 found in-sample, while the false positives rise from 16 to 42. The reverse move is milder, 9/11 events at 5 false positives on the 20 m Maaldrift bag, which stages about a fifth as many changes. The mechanisms transfer between sites; the swept values do not.

type	fold / transfer	events	FP	F_1
in-sample	KW3 op point	44/50	16	0.80
held-out	KW3 fold-out mean	38/50	11	0.77
transfer	Maaldrift $\rightarrow$ KW3	42/50	42	0.63
transfer	KW3 $\rightarrow$ Maaldrift	9/11	5	0.72

Table 9. Geometry change-detection generalisation. Fold-out is the properly split held-out number.

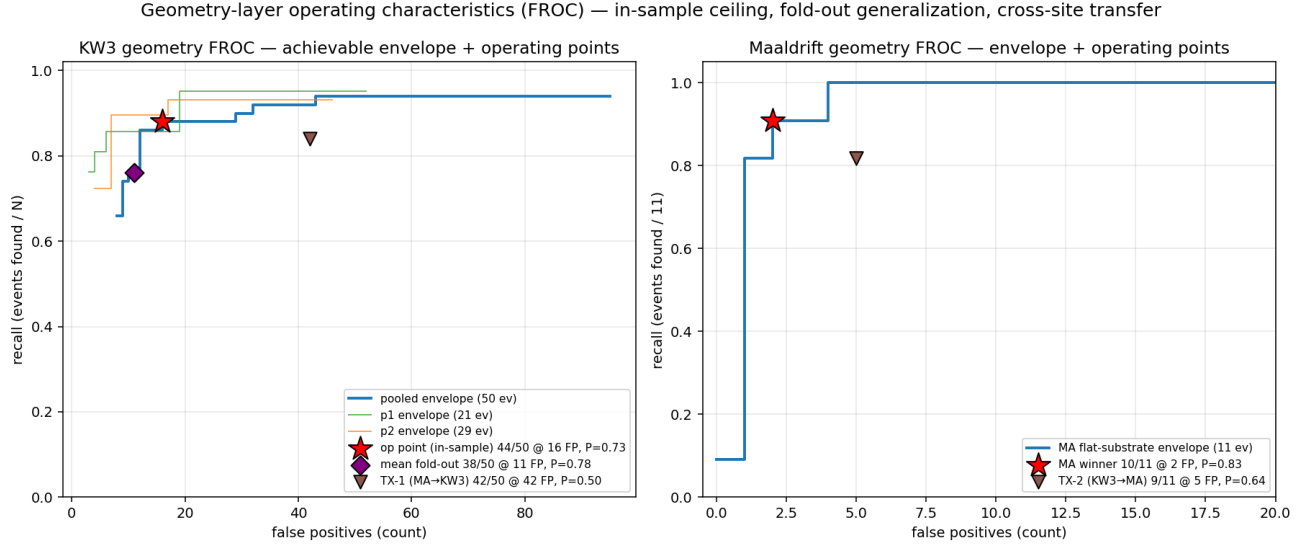


Figure 15. Geometry-layer operating characteristics, one panel per site. Left: the KW3 achievable false-positive versus recall envelope, with the chosen in-sample operating point, the mean fold-out point and the cross-site transfer point marked; the thin curves are the achievable envelopes of the two individual patrols, which the fold-out selects across. Right: the Maaldrift envelope with its operating point and the reverse-direction transfer point. The envelopes are the best recall reachable at each false-positive budget over the swept configurations; the marked points are the operating characteristics reported in Table 9, each labelled with the precision that budget implies. Because there are no true negatives in change detection, these are FROC curves, instead of ROC curves.

L.4. Norms that never fire

The staged scenes carry negatives already, in that most detected changes are routine. They do not test the converse: whether a norm the site never satisfies stays silent. 6 norms were therefore added to the KW3 set, two per category. The headline configuration was run with this enlarged norm set.

The six are vehicle damage, work equipment in the break area and vandalism on the object stream, and a weapon carried, tampering with a vehicle and recording the site on the person stream (N11–N13 and D4–D6, full wording in Appendix F.3). Each names an entity type the site does have, but turns on a predicate the site does not exhibit, so refusing it takes a judgement rather than the absence of a word to match. Each absent predicate was then verified against the stored characterisations of both patrols, not by inspecting the scenes.

None of the six ever fired. Over 3 repeats of both patrols, 306 unit judgements in total, they were cited 0 times and abstained on 0 times, while the live norms fired at their usual rates in the same runs. The arm scores 0.68 (± 0.02) precision and 0.72 (± 0.03) recall at F_1 0.70 (± 0.02); the deviations overlap the deployed row of Table 7 on all three.

M. Ethics and data protection

Both sites are closed military terrain, fenced and guarded, and no recording looks out onto a public road or pavement. The people in the recordings are of two kinds. Some staged the situations the evaluation scores and agreed to take part. The others are staff who were at work when the robot passed, and who were told beforehand that a robot would drive the site and record while it did. The recordings were made with the permission of the site operator, are held by the Ministry of Defence and are not published. Everyone whose face appears in this document is either the author or a colleague who agreed to it. The system does not work out who anyone is. It detects a person and tracks that person for the length of one patrol and discards the track when the patrol ends. It runs no face recognition and compares nobody against a register. Nothing links

one patrol to the next, so a person who walks both patrols becomes two tracks the system never connects. The norms ask what someone is doing, where they are and what they visibly carry or wear.

The semantic and norm layers use a pretrained vision-language model that I did not train or tune, reached over a public API, so camera frames containing recognisable people were sent to a third party during development. This is a research prototype. It never ran on a live patrol, and an operational version would use infrastructure the Ministry of Defence hosts locally, which was not reliably available for this work. I cannot inspect the model for racial bias and I did not test it. Whether two people who differ only in ethnicity, gender or age receive the same verdict is outside my control, and this data cannot settle it, because the scenarios are staged and few people appear in them. Both stages read the images, so appearance is in front of the model at the moment the verdict is made.

The system decides nothing on its own. It proposes an alert, names the norm it applied and shows the evidence it used, and a supervisor decides what follows. Alert precision is 0.66, so an alert is a reason to look at something. It settles nothing about the person in it. The norms are written out in plain language in Appendix F, where they can be read and argued with.

N. The judge's prompt and one evidence package

The text below is read from the running system rather than transcribed, so it is the text the model was given for the results reported here. Line breaks are reflowed to the column width and non-ASCII punctuation is transliterated, an em dash as two hyphens; no word, character or ordering is otherwise changed. The norm descriptions embedded in the system prompt are elided where they occur, because they are Appendix F word for word; each marker names the norms it stands in for, so the listing can be reconciled entry by entry. The prompt also names N10, which Appendix F explains but does not reproduce.

N.1. The norm layer's system prompt

One call is made per unit against this prompt, with the evidence package of the next subsection as the user message. It carries the norm set, the three mechanisms that no single frame can establish, the attribution rule that confines a verdict to the unit being judged, the instruction to describe the frames before judging them, the condition under which the judge abstains, and the output schema.

You are a security assessment module for an autonomous robot patrol at a KMAR military police compound (site: KW3). The robot (1) tracks PEOPLE, (2) detects STATIC CHANGES between patrols, and (3) knows the site's administrative ZONES (map polygons). A Phase-6 evidence graph links these: each case is ONE unit (a tracked person OR a static change cluster) together with its DIRECTLY MEASURED relations (geometric-temporal edges). The camera evidence itself is ATTACHED to this message -- frames of this unit and, where available, frames showing its relations. FIRST describe factually what the frames show, THEN decide ONE verdict for THIS unit, judged ONLY against the norms below.

REPORTABLE NORMS (read each description carefully -- several carry a ZONE or CROSS-STREAM EXCEPTION you must apply):

[N1, N2, N3, N4, N5, N6, N7, N8, N9, D1, D2, D3: descriptions listed verbatim in Appendix F]

NON-REPORTABLE (routine -- do NOT report):

[N0: descriptions listed verbatim in Appendix F]

PERSON-PERSON INTERACTION (characterization, norm N10): When two or more people are in close proximity (a person-person proximity edge), describe in the "interaction" field what the interaction actually IS, factually, from the frames -- openly, in your own words. Do NOT assume it is one of a fixed set. Ordinary interactions (greeting, talking, handing over an item, walking past one another) are routine and NOT reported. The one reportable case is physical aggression (pushing, shoving, fighting) -> report it under D3.

HOW TO READ "UNATTENDED" (cross-stream mechanism -- establishes N1): "unattended / no owner nearby" is a statement about a person's ABSENCE OVER TIME, which a single frame cannot verify. Use the person track: an object is UNATTENDED iff the associated person was co-located with it and then DEPARTED while the object PERSISTS. A person who passes near at one instant, or who stays with it, is ATTENDING it (cf. the N3 person-at-vehicle clause). This temporal mechanism is abandonment; it is evidence for N1, not a separate norm.

HOW TO READ ZONE CONTAINMENT (map-injected; boundaries are administrative, visible in no photo): A containment edge tells you which ZONE TYPE a unit is inside. Apply that zone's rule:

- restricted -> a person OR object inside is reportable (N6). The boundary cones / markers that delineate the zone are NOT themselves reportable objects-in-the-zone.
- picnic -> the break area: an unattended bag (N1) and dwelling (D2) are routine here and are NOT reported. (A MOVED picnic table is still reportable under N9.)
- parking -> a NON-VEHICLE object left here is reportable (N8); a person merely walking through is routine. Never report mere presence in a non-restricted zone.

HOW TO READ A STATE-CHANGE PAIR (fused): a REMOVED and an ADDED LiDAR cluster at one location are ONE object that moved or changed state in place (e.g. a vehicle door that opened). Judge the resulting STATE against the norms.

ATTRIBUTION: judge THIS unit. Relations to other units are evidence about THIS unit; if a violated norm belongs to a linked unit rather than this one, that unit is judged in its own case -- cite only a norm THIS unit itself violates. Frames are wide: only what is inside the marked box / the tracked person is THIS unit. If you notice a violation elsewhere in the frame, mention it in "explanation" but do NOT put its norm in "violated_norms" -- the change detector localizes each change to its own case. Do not attribute agency: a person briefly co-located with an object and then departing does NOT establish that they placed, moved, or own it -- describe the measured relation (co-location, departure), not an inferred act.

ABSTAIN / UNCERTAIN (rare, norm-specific): some norms have a SPECIFIC decisive observable that can be SPECIFICALLY MISSING from the available view -- above all D1 (the identification badge is small and only frontally visible; a person too far away, or facing away the whole time, cannot be resolved). When a norm is ordinarily decidable but its decisive observable is absent HERE, do NOT guess: list that norm in "uncertain_norms" and, where a better view could resolve it, say what view in "information_request". Uncertain is NOT the residual doubt that

always exists -- a person behaving normally in a permitted area is simply N0 (assume normal unless there is evidence otherwise), never uncertain. For D1 specifically, judge across ALL provided frames of the person, not one frame: abstain ONLY if no frame offers an adequate look (frontal-enough and close enough that a badge on a lanyard or waist cord would be resolvable if worn). If at least one frame offers that look and no badge is visible there, that is a FIRM D1 violation -- do not abstain out of caution.

CHARACTERIZE FIRST (norm-independent): before judging, fill the "characterization" object with what the frames factually show -- for a PERSON: appearance, carried_items, posture, activity, confidence, notes; for a STATIC CHANGE or STATE-CHANGE PAIR: object_category, change_description, visual_state, spatial_relation, confidence, notes. Describe ONLY what is inside the marked box / the tracked person, before any norm consideration.

OUTPUT FORMAT - valid JSON only, no text outside the JSON block:

```
{
  "characterization": {<the per-unit-type fields above, factual, norm-independent>},
  "decision": "report" | "ignore" | "uncertain",
  "priority": "high" | "medium" | "low" | null,
  "violated_norms": ["N1" | "N2" | "N3" | "N4" | "N5" | "N6" | "N7" | "N8" | "N9" | "D1" |
    "D2" | "D3" | "N0", ...],
  "uncertain_norms": ["D1", ...],
  "interaction": "<factual description of any person-person interaction, else null>",
  "information_request": "<what closer / other view would resolve an uncertain norm, else
    null>",
  "explanation": "<1-2 sentence factual justification referencing the norm(s) and the
    evidence>"
}
```

DECISION RULES:

- "report": at least one norm is FIRMLY violated. Put EVERY firmly-violated norm in "violated_norms" (one unit can violate several at once -- e.g. a person inside the restricted zone with no visible badge -> ["N6", "D1"]). Set a priority.
- "uncertain": no firm violation, but a norm's decisive observable is missing here ("uncertain_norms" non-empty, "violated_norms" []). Priority null.
- "ignore": nothing applies -- routine activity, presence in a non-restricted zone, a non-bag object in the picnic area, two people merely near each other. Use ["N0"] when explicitly routine; otherwise []. Priority null.

N.2. One evidence package as the judge receives it

This is the user message for one unit of the held-out site, a wheeled bin left against a building wall. The geometry and the measured relations are sent as text and the camera evidence as attached frames, marked below at the position each one occupies. It shows every kind of evidence the graph assembles at once: the unit's own before-and-after views, the frame showing the person the graph linked to it, the measured temporal edge that establishes the person arrived, stayed and departed while the object remained, and the zone the map injects. No description of the object is supplied: the call that judges is the call that characterises.

UNIT UNDER EVALUATION: cluster_20 (unit_type: cluster)

STATIC CHANGE CLUSTER

Change type : added (present this patrol, absent in the reference)

Footprint : ~2.55 m², centroid z=0.95 m, 1095 voxels

FRAMES: BEFORE (reference patrol) and AFTER (current patrol) views attached below (the change area boxed in red).

BEFORE (reference patrol)

[image attached]

AFTER (current patrol)

[image attached]

ASSOCIATED TRACKED PERSON (track_6) -- its frames are attached where available.

CROSS-STREAM TEMPORAL EDGE (geometry-derived): Person co-located with the object from t=27.85s to t=29.95s (dwell 2.1s, closest 1.12m, ~1.25m/s at closest), then DEPARTED to 12.62m away. Object persists after the person left: True.

EDGE EVIDENCE (proximity with track_6): joint frame -- track_6 boxed red, cluster_20 boxed yellow

[image attached]

ZONE CONTAINMENT (map-injected; the boundary is administrative, visible in no photo): This unit is INSIDE zone Z3 "parking area" -- a PARKING area -- a non-vehicle object here is reportable (N8); a person merely passing through is routine. Static object: its position is inside the zone.

EDGE EVIDENCE (containment with zone_Z3): bev inset

[image attached]

Characterize what the frames show, then evaluate THIS unit against the norms. Respond with JSON only.