WHAT VIDEO ANOMALY PROBES MEASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

An above-chance probe score on frozen video features is often read as evidence that the representation captures anomaly. We argue that this reading needs an evaluation contract, and we audit the supervised probe with four low-dimensional controls—scene complexity, video identity, position in video, and detected-agent counts—under scene-grouped folds, paired rows and video-clustered uncertainty, on a frozen V-JEPA 2 encoder over UCF-Crime and NWPU-Campus. On UCF-Crime, a linear probe given only the first encoder window of each clip predicts whether the clip will contain an anomaly at video-level AUROC 0.899; this is a video-level diagnostic, not evidence of temporal anticipation. On NWPU-Campus, a nine-dimensional count vector is a strong comparator under the diagnostic: with each arm's better of two heads, no tested representation arm is superior after paired Holm correction, and the count arm is the highest selected arm in every one of 25 stored seeds. That ordering is conditional on the readout: under a fixed linear head the count arm sits at chance and no arm is distinguishable from it, so we report paired tests within each head family rather than a single ranking. The seed sweep jointly varies fold assignment and optimisation and moves arms by up to 0.186 AUROC, widest for the complexity control. Under the benchmark's own normal-only protocol the same counts reach at most 60.6 micro-AUC against a published 76.9, whereas the frozen state reaches 59.1 micro-AUC and an exploratory 67.2 macro-AUC in the prespecified estimator cell: a control can dominate a supervised diagnostic without being a competitive detector. Two analytical results—that a prediction-error residual is a dispersion functional, and that a self-normalised pre-hoc hazard measures the model's own tail mass—motivate the complexity control and explain why no unsupervised anticipation arm appears. We release the protocol and every artefact, and limit the conclusion to measurement design: an anomaly score should state its target, information cutoff, controls, grouping, readout selection, aggregation and uncertainty before it is read as evidence about a representation.

1 INTRODUCTION

Video anomaly scores are often read as evidence that a representation contains anomaly semantics. The usual instrument is a probe: freeze the representation, fit a small supervised head, report an AUROC. The probe is useful, but its estimand is the predictability of the labels under a particular split, readout and aggregation. It does not by itself identify event understanding, anticipation, or deployable detection.

We study that gap with a benchmark-audit protocol. The protocol adds four low-dimensional controls—scene complexity, video identity, position in a video, and detected-agent counts—groups folds by scene, compares arms on identical rows, and reports paired, video-clustered uncertainty with multiplicity control. We apply it to a frozen V-JEPA 2 encoder (Assran et al., 2025) on UCF-Crime (Sultani et al., 2018) and NWPU-Campus (Cao et al., 2023). The controls are diagnostic comparators, not proven non-semantic causes: object presence and scene occupancy can themselves be part of an anomaly definition, and a representation that fails to beat a control may still carry information the control does not.

Four findings survive the audit. First, on UCF-Crime a linear probe given only the first encoder window of each clip predicts whether the clip will contain an anomaly at video-level AUROC 0.899; that is a video-level diagnostic and does not establish temporal anticipation (Section 4.1). Second,

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

on NWPU-Campus a nine-dimensional count vector is a strong comparator for the tested features: with each arm's better of two heads, no representation arm is superior after paired Holm correction, and counts are the highest selected arm in every one of 25 stored seeds. Third, that ordering is conditional on the readout: under a fixed linear head the count arm sits at chance, so Section 4.2 reports paired tests within each head family rather than one ranking. Fourth, the stored seed sweep—which jointly changes fold assignment, head initialisation and minibatch order—moves arms by up to 0.186 AUROC, widest for the three-dimensional complexity control.

The benchmark's own protocol changes the picture. With normal-only training and the published frame-level evaluation, the count baseline reaches at most 60.6 micro-AUC against a published 76.9, while the frozen state reaches 59.1 micro-AUC and 67.2 macro-AUC in the prespecified per-scene-kNN cell; the macro comparison is exploratory. A control can dominate a supervised diagnostic without being a competitive detector under the deployment protocol.

Two analytical results sit behind the protocol (Appendix C). A prediction-error residual is a dispersion functional of the conditional law, which is why a scene-complexity control belongs in the protocol; and a self-normalised pre-hoc hazard on model-drawn futures measures the normal model's own tail mass rather than the future, one constant across context and horizon for location-scale models, which is why no unsupervised anticipation arm appears in this study. The remaining controls are empirical, learned from our own failed anticipation attempts (Appendix E).

Our contributions are methodological and bounded: (i) a reusable control protocol for diagnostic probes on video anomaly corpora, (ii) matched evidence that simple controls, readout selection and aggregation materially change the conclusion on two surveillance corpora, and (iii) an artefact-backed reporting discipline that separates diagnostic probing from native anomaly detection and from event anticipation. We do not claim that the tested representation lacks anomaly information, that either corpus is unusable, or that other video anomaly methods share these failure modes.

2 Related work and scope

Semi-supervised video anomaly benchmarks train on normal video only and evaluate frame-level scores (Cao et al., 2023; Luo et al., 2017); weakly supervised systems fit a ranking head on video-level labels of the training split and evaluate on held-out clips (Sultani et al., 2018; Tian et al., 2021; Wu et al., 2020). Our supervised probe is neither. It is a diagnostic fit on labelled evaluation videos under cross-validation, in the sense of the representation-probing literature (Hewitt & Liang, 2019), and it asks what a frozen feature makes linearly or shallowly extractable under a chosen split. A detector must fit without the evaluation anomalies and operate at a specified time. The two protocols answer different questions, and Section 4.4 is where the difference is measured.

Comparing a probe with label-free controls follows the control-task view of probing (Hewitt & Liang, 2019); shortcut learning and dataset bias are the general concern (Geirhos et al., 2020; Torralba & Efros, 2011). Grouped cross-validation, cluster bootstrap (Cameron et al., 2008), step-down multiplicity control (Holm, 1979) and the study of evaluation variance (Bouthillier et al., 2021) are standard tools; our contribution is a concrete set of controls attached to the exact rows, folds and artefacts of video anomaly probes. Song & Lee (2026) show that pooled frame-level metrics in weakly supervised detection mostly measure video-level ranking; our pre-event diagnostic and the count comparison are complementary case studies, not a claim of priority over that observation, and Rashidi (2026) audit the benchmarks themselves.

NWPU-Campus is scene-dependent, which makes scene grouping and occupancy controls especially relevant (Cao et al., 2023; 2025). Published methods are reported only as context for the native protocol; a literature number is not a paired baseline in our experiment. The scope is one frozen video encoder, two surveillance corpora, and the controls of Section 3. Language-mediated detectors (Wu et al., 2024; Zanella et al., 2024), score-based predictors (Zhang et al., 2025), open-set benchmarks with disjoint anomaly types across splits (Acsintoae et al., 2022) and multimodal corpora (Wu et al., 2020) are where the protocol should be run next; whether they share these floors is not evidence supplied by this paper.

3 EVALUATION PROTOCOL

Estimand and controls. We ask whether a learned feature gives predictive performance beyond simple, label-free comparators under a specified diagnostic probe. A result that does not exceed a control is reported as "no demonstrated improvement over that control"; it is not read as identifying a mechanism. The four controls are: *scene complexity* (3-d: mean token variance, mean patch standard deviation and mean token norm of the encoder output); *video identity* (a video-level diagnostic from a pre-event window, Section 4.1); *position in video* (1-d: the normalised index of the slice within its video); and *object counts* (9-d: per-slice detections of six COCO classes—person, bicycle, car, motorcycle, bus, truck—one "other" bin for every remaining class, plus the number of observations and the number of distinct tracks). Counts may carry genuine event information on a scene-dependent corpus; their value here is that they are simple and auditable. We call a control a *floor* when a score is read against it. Appendix C gives the reason the complexity control is in the protocol; the other three are empirical, and Appendix E records how they were found.

Data and representation. We use the published UCF-Crime and NWPU-Campus splits and frame masks; neither corpus is re-cut. The representation is a frozen V-JEPA 2 ViT-L encoder with full-frame 256×256 inputs and no fine-tuning or domain adaptation. One forward pass spans 32 raw frames, of which 16 are fed to the encoder at temporal stride 2; with tubelet size 2 and patch size 16 this yields 8 temporal positions $\times 256$ spatial patches. A *slice* is one temporal position and therefore covers 4 raw frames (0.16 s at 25 fps); a scored row is a slice with its preceding slices as context inside the same window, and its label is the maximum of the corpus's frame mask over its raw frames. Agent features come from a COCO-pretrained YOLO11x detector with ByteTrack and YOLO11x-pose, with tracking reset between videos (Appendix A). The 34-d *agent behaviour* arm is seventeen per-observation statistics—box height, width, area and aspect, box centre (x, y), speed, height-scaled speed, acceleration, turning angle, detector confidence, keypoint confidence, keypoint x and y spread, ankle and wrist height, and torso length—each aggregated by max and by mean over the observations in the slice. A slice with no detection has zero counts and zero observations; its behaviour statistics are undefined and are imputed with the column mean over all rows, which is transductive and is recorded as such.

Folds and readouts. Probe folds are grouped by scene on NWPU-Campus (43 scenes, $K = 5$) and by video on UCF-Crime, where every clip is its own camera. Every arm is scored on identical rows. Two heads are fit per arm, a linear head and a two-layer MLP, both with the same epochs and seed. We report the *raw* out-of-fold AUROC and a *complexity-adjusted* AUROC in which out-of-fold scores are residualised on the 3-d complexity basis by least squares, cross-fitted by video; the adjustment is label-free, but it can remove signal that is correlated with complexity as well as nuisance variation, so it is a sensitivity, not a semantic purification. Standardisation and the PCA projections of the token and agent-ROI arms are fit on all rows and are labelled transductive. Two comparisons are reported. The *selected-head* comparison takes each arm's better of its two heads on the same out-of-fold predictions that are then bootstrapped; that selection is applied to the controls as well as to the representation arms, and the reported intervals condition on it. The *fixed-head* comparison pairs every arm against the count control under the same head, with no selection, in two separate families.

Uncertainty. Paired differences use identical rows, a video-clustered bootstrap and Holm–Bonferroni step-down at family-wise $\alpha = 0.01$ within each declared family; per-comparison intervals are shown beside Holm-adjusted p-values. Video-clustered intervals describe variation among videos conditional on the observed scene set; they do not establish new-camera generalisation; a scene-clustered version, which we have not run, would generally be wider. The seed sweep redraws the fold assignment, head initialisation and minibatch order together, so it is reported as joint evaluation randomness, not as a fold-only variance estimate.

Native protocol. The normal-only analysis fits density estimators on the 305 normal training videos and evaluates the published frame-level micro- and macro-AUC on the 242 test videos, with no anomaly label at fit time. The count control and the representation are run through the same grid of 5 estimators; the representation's primary cell was fixed in advance as the 128-d state with the estimator that was best on counts (per-scene kNN), micro-AUC was the planned metric, and macro-

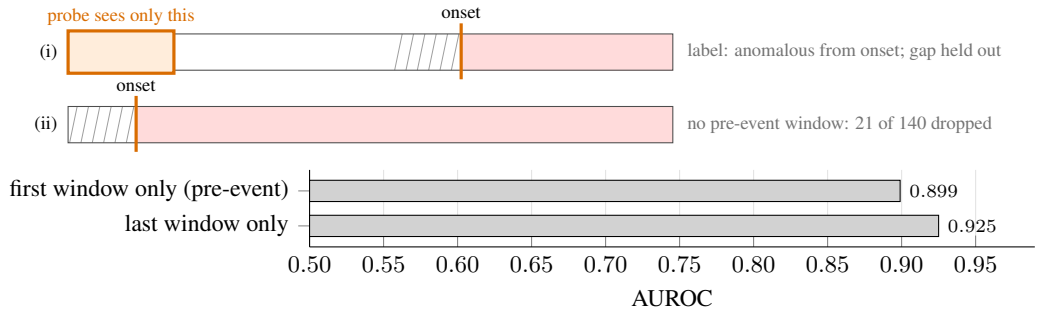

Figure 1: **Top:** how the first-window arm is cut. Every clip, normal or anomalous, contributes its first encoder window (bold); the label is whether the clip contains an anomaly anywhere. A 5.0 s gap before the onset (hatched) is held out, and an anomalous clip whose onset falls inside that gap has no pre-event window and is dropped rather than truncated—row (ii)—which removes 21 of 140 anomalous clips. Durations are schematic. **Bottom:** video-level AUROC of a linear probe on that single window, and, for completeness, on the last window of every clip. The axis starts at the ranking null of one half. Whatever the first-window bar measures, it is not anticipation, because at that point there is nothing yet to anticipate; the last-window bar is not a provenance control, because the event or its aftermath can still be in view.

AUC is exploratory. Anticipation is a different estimand again: it needs a specified event onset and an availability cutoff, and no arm in this paper is offered as an anticipation result.

4 Results

4.1 Pre-event video discrimination is not temporal localisation

We fit a linear probe inside the labelled UCF-Crime test split, K-fold over videos, to predict whether a clip contains an anomaly anywhere, from a single encoder window. The *first-window* arm uses the first 32-frame window of the file for every clip, normal or anomalous, so the anchor is the same for both classes; anomalous clips whose onset falls within 5.0 s of the start are dropped rather than truncated, which removes 21 of 140 anomalous clips and leaves 269 (Figure 1). It reaches video-level AUROC 0.899. The endpoint is video-level: it asks whether the clip is *eventually* labelled anomalous, not whether a score can rank the event inside the clip, and at the first window there is nothing yet to anticipate. The cue may be camera, framing, compression, scene type or source footage as well as any genuine precursor; we do not decompose it. A *last-window* arm, the final window of every clip, reaches 0.925 on 290 clips; it is shown for completeness and is not a provenance control, because the event or its aftermath can still be visible there.

We do not set this number beside a frame-level probe that sees the whole clip, or quote a share of that probe's lift explained by the first window: the two probes are fit on different row populations with different post-processing, so their ratio is not a matched estimand. What the first-window result supports is that a supervised probe can separate anomalous from normal *clips* before the event. What it takes to claim anticipation is a within-video design: near-onset against far-from-onset windows of the same clip, an availability cutoff at the end of the encoder window, and source-grouped folds. Appendix E.4 runs that design on both corpora, and on UCF-Crime the within-video arm falls below its position floor while the between-video arm clears it: the between-video signal is the video identity measured here.

4.2 Object counts are a strong comparator for NWPU-Campus probe scores

The primary NWPU-Campus partition scores 94,232 labelled slices from 242 videos over 43 scene folds. Table 1 gives both heads for every arm, raw and complexity-adjusted; Table 2 is the selected-head ordering, and Figure 2 its paired comparison against counts.

Table 1: Both heads for every arm in the primary partition, raw and complexity-adjusted out-of-fold AUROC. This is the table the selected-head ordering of Table 2 is drawn from. Under the linear head the count control and position sit at chance; under the MLP they lead.

Arm	linear raw	linear adj.	MLP raw	MLP adj.
agent counts (9-d)	0.496	0.490	0.677	0.686
position (1-d)	0.487	0.483	0.638	0.636
tokens frame-global (128-d)	0.583	0.571	0.539	0.523
behaviour + counts	0.522	0.513	0.557	0.556
pooled state (128-d)	0.559	0.554	0.519	0.508
agent behaviour (34-d)	0.508	0.499	0.548	0.552
ROI + counts	0.482	0.465	0.486	0.478
agent-ROI appearance (128-d)	0.488	0.474	0.449	0.433
cx only (3-d)	0.494	0.472	0.484	0.443

Table 2: Selected-head ordering in the primary partition (seed 1750412224). Each arm's head is the better of the two in Table 1 on the same out-of-fold predictions that are then bootstrapped, for controls and representation arms alike, so intervals condition on that selection. † marks an interval that includes chance, i.e. lower limit ≤ 0.5 before rounding; the frame-global token row's lower limit rounds to 0.500 but exceeds chance before rounding, so it is not marked.

Arm	Head	AUROC	99 % CI
agent counts (9-d)	mlp	0.686	[0.611, 0.742]
position (1-d)	mlp	0.636	[0.596, 0.678]
tokens frame-global (128-d)	linear	0.571	[0.500, 0.630]
behaviour + counts	mlp	$0.556^{\dagger}$	[0.480, 0.627]
pooled state (128-d)	linear	$0.554^{\dagger}$	[0.462, 0.646]
agent behaviour (34-d)	mlp	$0.552^{\dagger}$	[0.464, 0.620]
ROI + counts	mlp	$0.478^{\dagger}$	[0.365, 0.589]
agent-ROI appearance (128-d)	linear	$0.474^{\dagger}$	[0.365, 0.582]
cx only (3-d)	linear	$0.472^{\dagger}$	[0.410, 0.541]

Selected-head comparison. With each arm's better head, the count arm reaches 0.686 adjusted AUROC and position is the next strongest. Of 8 arms paired against counts on identical rows, 7 have Holm-corrected intervals excluding zero in favour of counts and none is superior; position is the one arm not distinguishable from counts. The selection step favoured the controls here as much as the representation arms: counts and position both selected the MLP, the pooled state and frame-global tokens the linear head.

Fixed-head comparison. The ordering above depends on the readout. Under a fixed linear head the count control's adjusted AUROC is 0.490, interval [0.397, 0.589], and the pooled state and frame-global tokens sit above it in point estimate. Pairing every arm against the count arm *under the same head*, with no selection and Holm within each family of 8 comparisons (Table 3): under the MLP, 0 arms beat counts and 7 lose; under the linear head, 0 beat counts and 0 lose, with the pooled-state difference at +0.064 [-0.026, +0.163] and the token difference at +0.080 [-0.023, +0.183]. A linear head on nine raw counts can express only a score monotone in each count, whereas the representation arms give the same head 128 whitened dimensions, so the linear family is not a like-for-like test of the features; it is the reason the paper reports paired families rather than one ranking, and the reason "counts are a strong comparator" is stated for the tested readouts and not as a property of the corpus.

None of this proves that the representation uses counting, that counts are non-semantic, or that a different readout could not recover complementary information. The concatenation arms (behaviour + counts, ROI + counts) score below counts alone under both heads, which is a statement about the optimisation of these shallow heads on these inputs, not evidence that the added coordinates are noise. Token pooling (mean, max, learned attention) is a readout sensitivity in Appendix F.

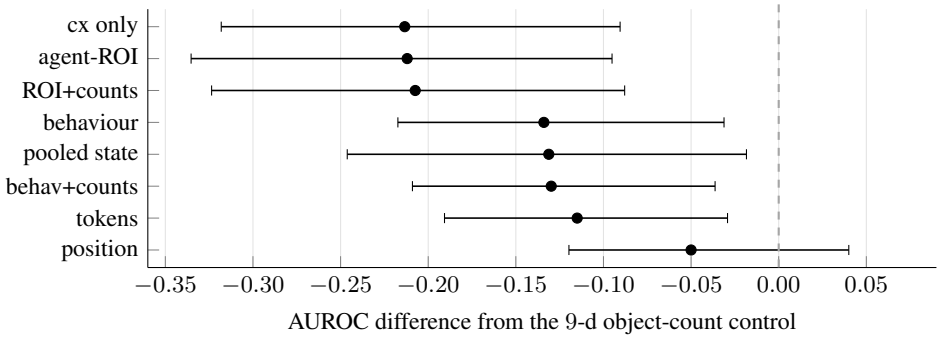

Figure 2: Every arm against the object-count control, each under its *selected* head (Table 2), paired on identical rows in one fold partition, video-clustered bootstrap, 99 % intervals, Holm–Bonferroni at family-wise $\alpha = 0.01$. The dashed line is no difference from counting. Under this readout policy nothing lies to the right of it: 7 of 8 arms lose with the corrected interval excluding zero, and position-in-video—the one exception—is not distinguishable from counting rather than better than it. The same comparison with the head held fixed is Table 3: under the MLP it reproduces this picture, under the linear head no arm is distinguishable from counts.

Table 3: Paired difference from the count control under the *same* head, identical rows, video-clustered 99 % intervals, Holm within each head family. No head selection enters this table. Rows are ordered by effect size within each family, so Holm-adjusted p-values need not be monotone down the column.

Arm (same head as counts)	AUROC	Δ vs counts	99 % CI	p_{Holm}	
linear head: counts 0.490 [0.397, 0.589]					
ROI + counts	0.465	-0.025	[-0.108, +0.071]	1.0000	n.s.
cx only (3-d)	0.472	-0.018	[-0.139, +0.094]	1.0000	n.s.
agent-ROI appearance (128-d)	0.474	-0.017	[-0.103, +0.082]	1.0000	n.s.
position (1-d)	0.483	-0.007	[-0.131, +0.116]	1.0000	n.s.
agent behaviour (34-d)	0.499	+0.008	[-0.078, +0.106]	1.0000	n.s.
behaviour + counts	0.513	+0.022	[-0.065, +0.118]	1.0000	n.s.
pooled state (128-d)	0.554	+0.064	[-0.026, +0.163]	0.4585	n.s.
tokens frame-global (128-d)	0.571	+0.080	[-0.023, +0.183]	0.3920	n.s.
mlp head: counts 0.686 [0.611, 0.742]					
agent-ROI appearance (128-d)	0.433	-0.253	[-0.376, -0.132]	0.0020	loses
cx only (3-d)	0.443	-0.243	[-0.343, -0.135]	0.0020	loses
ROI + counts	0.478	-0.207	[-0.324, -0.088]	0.0020	loses
pooled state (128-d)	0.508	-0.177	[-0.288, -0.062]	0.0020	loses
tokens frame-global (128-d)	0.523	-0.163	[-0.255, -0.062]	0.0020	loses
agent behaviour (34-d)	0.552	-0.134	[-0.217, -0.031]	0.0020	loses
behaviour + counts	0.556	-0.130	[-0.209, -0.036]	0.0020	loses
position (1-d)	0.636	-0.050	[-0.120, +0.040]	0.1565	n.s.

4.3 Probe rankings vary with joint evaluation randomness

Across 25 stored partitions the count arm ranges from 0.662 to 0.711 under its selected head and is the highest selected-head arm in every partition (Figure 3, Table 5). The widest range is 0.186, for cx only (3-d); the learned arms move by roughly a tenth of an AUROC. Because the seed also re-draws head initialisation and minibatch order, and because the head is re-selected per arm in every partition, this is joint split, optimisation and selection variation, not a partition-only estimate. The practical conclusion is modest: an absolute probe AUROC is incomplete without its split, readout policy and seed, and paired differences on identical rows are the quantity to quote.

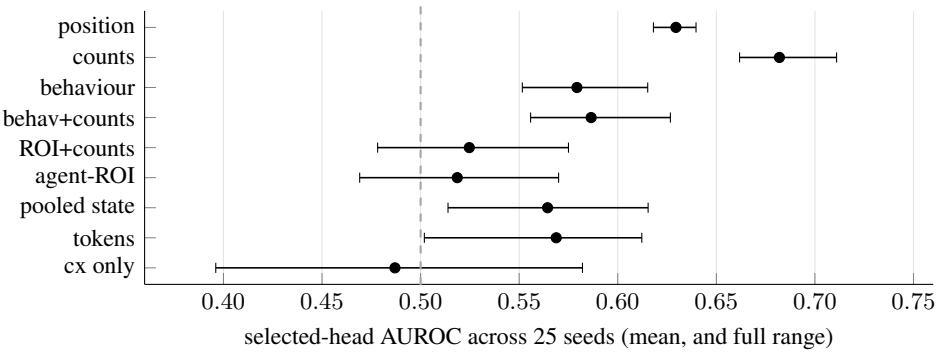

Figure 3: Each arm's selected-head AUROC across 25 seeds; the seed re-draws the fold partition, head initialisation and minibatch order together, and the head is re-selected per arm in every seed, so the spread is joint split, optimisation and selection variation. Features and rows are identical throughout. Arms are ordered by spread, widest at the bottom; the dashed line is chance. The two most stable arms are the two lowest-dimensional controls—position and object counts—and the widest spread, 0.186, belongs to the three-dimensional complexity control; the learned arms move by roughly a tenth of an AUROC. This is why every claim in the paper is a paired difference on identical rows rather than an absolute value.

4.4 The native protocol changes the conclusion

Under NWPU-Campus's normal-only protocol the same 9-d counts, fit on the 305 normal training videos, reach 54.1 micro-AUC with a per-scene Gaussian and at most 60.6 with per-scene kNN across 5 estimators (Appendix G); the lowest published entry is 61.9 (MemAE) and the highest 76.9 (Zhang et al., 2025). A control that dominates the supervised diagnostic is not thereby a competitive detector.

The frozen state, through the same estimator grid, reaches 59.1 micro-AUC and 67.2 macro-AUC in the prespecified cell (128-d, per-scene kNN). That cell is also the maximum over the 15-cell representation grid of Table 7, so its point estimate is upward-biased by coincidence with the selection and we read the paired differences rather than the level. Against count kNN the paired micro difference is -1.5 [-15.1, +14.2] and the macro difference +8.4 [+1.9, +15.0]; micro was the planned metric and macro is exploratory, and the primary cell's macro exceeds all 11 floor cells of the grid and sits -10.9 relative to the published ADSM macro. A representation can carry within-video ranking signal that a pooled micro metric does not show. A probe comparator and a native detector answer different questions, and Figure 4 is the same feature answering both.

5 Discussion and limitations

The evidence supports a narrow conclusion. On two surveillance corpora, simple controls, the choice of readout and the choice of aggregation materially change what a frozen-feature probe appears to measure. Object counts are a strong comparator inside the supervised diagnostic under the heads that favour them and a weak detector under the benchmark's own protocol; a representation that loses to counts under one readout and one pooled metric shows within-video signal under another. These are properties of the evaluation contract, not a verdict that the representation is empty or that the benchmarks are useless.

What the controls are not. They are not causal decompositions. A feature that does not outperform counts can carry distinct information; a count can be genuine anomaly evidence on a scene-dependent corpus; and residualising on complexity removes label-correlated signal as readily as nuisance variation (the linear-residualisation remark of Appendix D requires the signal to be uncorrelated with the proxies, which is not verifiable here). Stronger attribution requires nested incremental models—counts against counts-plus-representation with a readout that can ignore added coordinates—which the stored concatenation arms are not.

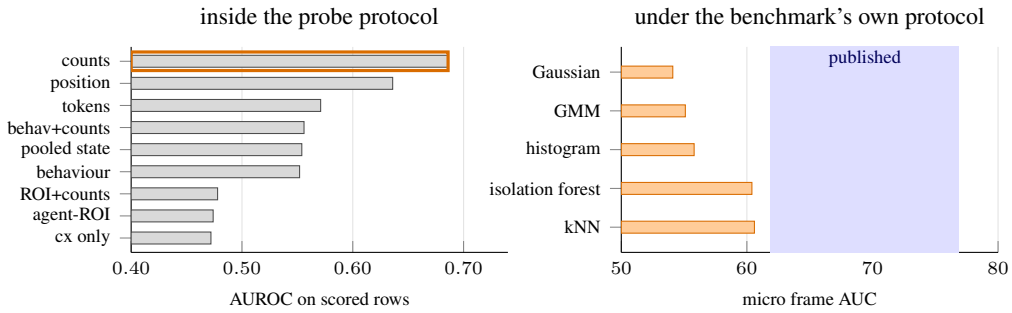

Figure 4: The same 9-dimensional count vector, scored two ways, with opposite verdicts. **Left:** inside the probe protocol of Section 4.2—a supervised head cross-validated within the test split, each arm under its selected head (Table 2)—counting (outlined) ranks above every learned arm. **Right:** under NWPU-Campus's own semi-supervised protocol, fit on normal video only, the same features under 5 density estimators reach at best 60.6—below the lowest published entry (MemAE, 61.9) and far below the best (76.9); the shaded band spans the published range. *The two axes are different quantities on different row sets and must not be read across.* What matters is where counting sits within each panel. The published figures are other authors' runs, and are given per row with their sources in Appendix B rather than plotted here.

What the intervals do not cover. The selected-head ordering conditions on a choice made on the evaluated predictions, for controls and representation arms alike; the fixed-head families remove the choice but pair a nine-dimensional monotone readout against whitened 128-d inputs. Residualisation is cross-fitted by video and PCA is fit on all rows, both transductive within a scene. Bootstrap intervals are video-clustered and conditional on the observed scene set; a scene-clustered version would be wider, and we have not run it. The seed sweep combines split, optimisation and selection randomness. Detector coverage is uneven across videos (Appendix H), which handicaps the count control rather than favouring it. The UCF-Crime endpoint is a video-level diagnostic; source-footage duplication across clips is a plausible part of it and is not measured here.

Anticipation. No arm in this paper is offered as an anticipation result. The unsupervised hazards of Appendix E are pinned by construction; the supervised anticipation probes of Appendix E.4 are inconclusive at their power on NWPU-Campus and measure video identity on UCF-Crime. A subsequent harness with causal next-window targets, persistence baselines and nested fitting on NWPU-Campus found no pooled single-window arm beyond its label-free floors; it is not part of this paper's evidence, and low power there is not evidence that precursors do not exist. A defensible anticipation claim needs a human-adjudicated event onset, an availability cutoff at the end of the encoder window, same-clip near-versus-far negatives, and alert, lead-time and false-alert metrics rather than a pooled AUROC.

Scope. One frozen encoder, two corpora, shallow heads, one detector. Whether language-mediated detectors, score-based predictors or open-set benchmarks (Wu et al., 2024; Zanella et al., 2024; Zhang et al., 2025; Acsintoae et al., 2022) share these floors is the natural next run of the protocol, not a conclusion of this paper.

Recommendation. Every reported anomaly result should specify its target, label access, information cutoff, grouping unit, controls, readout-selection rule, aggregation metric and uncertainty cluster. Without those fields, a number is not a stable measurement of a learned representation.

AI USE STATEMENT

(This section is required and does not count toward the page limit.)

We used a large language model (Anthropic's Claude, via Claude Code) throughout this work. Of the tasks the ICLR policy places under **required disclosure**, we used it to: implement methods (most of the analysis code—the probe harness, the floors, the bootstrap and correction machinery,

the counting baseline); design and give feedback on experiments (proposing arms, controls and pre-registered decision criteria, recorded in a versioned plan before each run); propose and refine hypotheses; help develop the conceptual framework (the controls and the evaluation contract); formulate evaluation claims and assist in checking the analytical arguments of Appendices C and D; generate synthetic data sets (the numerical checks of those identities); clean and reformat datasets (the per-corpus metadata and split manifests); and interpret results. We did not use generative AI for translation or for qualitative or thematic data analysis, which are not applicable to this work. Of the tasks under **recommended disclosure**, we used it to create and edit software code, create the TikZ and pgfplots figures, draft and edit prose, summarise and identify literature, check bibliography entries against their sources, and format references. We did not use generative AI to produce, alter or impute any experimental measurement; every number in this paper is computed by the released code from the released artefacts.

We built three mechanisms specifically because AI assistance makes transcription errors easy to introduce and hard to see, and we report them here because they are what makes the disclosure checkable rather than merely stated. First, no number is typed into the manuscript: each is generated from a committed JSON artefact by a script that stamps the artefact path and JSON key beside the macro it defines, and a build check fails if a literal decimal appears in the body. Second, a build check fails if any cited work lacks a note recording how its bibliographic details were verified. Third, every experiment records its resolved random seed into its own artefact, after we found that an ordering we had been quoting was assembled across three different fold partitions.

That last point deserves stating plainly. Several errors in this work were introduced with AI assistance and several were found with it. We do not present the tooling above as making AI-assisted research safe; we present it as the minimum we found necessary to check our own work.

We take responsibility for the final content of this work, including all text, claims and artefacts, whether or not they were produced with the aid of generative AI.

Reproducibility statement

(Recommended; does not count toward the page limit.)

The supplementary material contains the protocol, every JSON artefact behind every number in this paper, and the code that turns one into the other. It is organised in three tiers, stated by effort. Tier 1 regenerates every number in the paper from the shipped artefacts in seconds, with no GPU, no dataset and no network; this is the tier that answers whether the tables reproduce the shipped summaries—it cannot establish that the measurements behind them are valid—and we expect it to be the one most readers run. Tier 2 reproduces all probe arms, the paired tests and the 25-partition sweep from cached features on a single GPU. Tier 3 is feature extraction and encoding, which requires the corpora; we cannot redistribute NWPU-Campus or UCF-Crime video, and we document the download and the exact preprocessing instead. Every script takes a seed and records the resolved value into its output, which Section 3 argues is necessary rather than merely good practice on this corpus.

Ethics statement

(Recommended; does not count toward the page limit.)

This work uses three public surveillance corpora (NWPU-Campus and UCF-Crime in the main text, ShanghaiTech in Appendix E.1), under their original terms, and introduces no new recording of people. We add no capability for identifying or tracking individuals: the agent features we compute are geometric and pose-based, deliberately class-agnostic, and are used here as an experimental control rather than as a proposed system.

The direction of the work is toward reducing overclaiming about deployed anomaly detection. Surveillance systems are procured on reported benchmark numbers, and we show that on two standard corpora a supervised probe score can be matched by low-dimensional controls and can change verdict with the readout, the aggregation and the evaluation protocol. We think the honest reading is that a probe AUROC on these benchmarks should not be taken as evidence of deployable capability, and we would rather that conclusion be available to the people who buy and deploy such systems than not.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Congqi Cao, Hanwen Zhang, Yue Lu, Peng Wang, and Yanning Zhang. Scene-dependent prediction in latent space for video anomaly detection and anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):224–239, 2025. doi: 10.1109/TPAMI.2024.3461718.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

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

Glenn Jocher and Jing Qiu. Ultralytics YOLO11, 2024. URL `https://github.com/ultralytics/ultralytics`.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6536–6545, 2018. doi: 10.1109/CVPR.2018.00684.

Yang Liu, Hongjin Wang, Zepu Wang, Xiaoguang Zhu, Jing Liu, Peng Sun, Rui Tang, Jianwei Du, Victor C. M. Leung, and Liang Song. CRCL: Causal representation consistency learning for anomaly detection in surveillance videos. *arXiv preprint arXiv:2503.18808*, 2025.

Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 341–349, 2017. doi: 10.1109/ICCV.2017.45.

Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15425–15434, 2021.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Mohammadreza Rashidi. Benchmark AUC is not deployable reliability: A cross-dataset audit of off-the-shelf features for surveillance video anomaly detection. *arXiv preprint arXiv:2606.29506*, 2026.

Inpyo Song and Jangwon Lee. Frame-level evaluation in weakly supervised video anomaly detection mostly measures video-level ranking. *arXiv preprint arXiv:2608.21854*, 2026.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision – ECCV 2020*, pp. 322–339, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hanwen Zhang, Congqi Cao, Qinyi Lv, Lingtong Min, and Yanning Zhang. Autoregressive denoising score matching is a good video anomaly detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *Computer Vision – ECCV 2022*, 2022.

A IMPLEMENTATION DETAIL

The setup of Section 3 in full, and four things a reader should know before quoting an interval from this paper.

What the intervals do not cover. (i) *Head selection.* In the selected-head comparison each arm's number is the better of two heads on the same out-of-fold predictions that are then bootstrapped, so it is a maximum over two heads and its interval carries neither the variance of that choice nor that of refitting. The selection is applied to every arm, controls included: counts and position selected the MLP, the pooled state and the token arm the linear head, so the choice does not work uniformly in one direction. The fixed-head families of Table 3 remove the choice. (ii) *Residualisation.* The complexity adjustment is a three-coefficient least-squares fit of score on complexity, with no label, cross-fitted over folds split by video rather than by scene. A held-out scene's scores are therefore adjusted using the score–complexity relation of other videos, some from the same scene: unlabelled transduction, which a strictly nested version would fit by scene. The adjustment also removes any label-related signal that is linear in the proxies, which is why raw scores are reported beside it. (iii) *Multiplicity.* Holm is applied to the bootstrap p-values; the $99\,\%$ intervals beside them are per-comparison and not adjusted for the family. Holm-adjusted p-values are a pure function of the raw ones, and a version of the producing script returned them without Holm's running maximum; the stored values were recomputed from the raw ones (`fix_holm_pvalues.py`, recorded in each artefact touched): 3 cells of the main paired comparison moved by at most 0.0005, and 11 cells in all across Appendices F and E.3 by at most 0.027, with no reject decision affected. Tables sorted by effect size therefore show adjusted p-values that are not monotone down the column; they are monotone in raw-p order. (iv) *The sweep.* The seed sets the fold partition, head initialisation and minibatch order together, and the head is re-selected per arm in every seed, so Figure 3 and Table 5 measure partition, training and selection randomness jointly. Spreads there are computed on unrounded values, so a displayed range can differ from the difference of its displayed endpoints by one unit in the last place.

Corpora. NWPU-Campus (Cao et al., 2023) supplies 305 normal-only training videos and 242 test videos over 43 scenes; we score 94,232 labelled slices from the test split. On UCF-Crime (Sultani et al., 2018) we use the 290-video labelled test split. Neither corpus is re-cut: we use the published splits and frame masks.

Encoder and slices. The representation is a frozen V-JEPA 2 ViT-L encoder (Assran et al., 2025, 300M parameters), used with no fine-tuning or domain adaptation. Frames are resized full-frame to 256×256 with no crop, since both corpora label whole frames and a centre crop would discard the periphery where a precursor first becomes visible. One forward pass spans 32 consecutive raw frames; every second frame is kept, so the encoder receives 16 frames, and with tubelet size 2 and patch size 16 this gives 8 temporal positions $\times\, 256$ spatial patches $= 2048$ tokens of dimension 1024. A *slice* is one temporal position and therefore corresponds to 4 raw frames, $0.16\,\mathrm{s}$ at $25\,\mathrm{fps}$. Windows are encoded contiguously and non-overlapping, and each video's frame rate is measured rather than assumed. A slice's label is the maximum of the corpus's own frame mask over its raw frames; slices with no mask are dropped, and each scored row is a slice with four preceding slices of context inside its window. Because the encoder attends across the whole window, no slice's features exist before the window's last raw frame; every timing statement in Appendix E uses that frame as the availability time.

Agent features. Agents come from an off-the-shelf COCO-pretrained detector—YOLO11x (Jocher & Qiu, 2024) at $1280\,\mathrm{px}$ with ByteTrack (Zhang et al., 2022), plus YOLO11x-pose for person keypoints—with the tracker reset between videos so track identities cannot leak across scenes. We store only what the detector saw (class, box, confidence, keypoints, track id); kinematics are derived downstream. Speed is the displacement of the box centre between consecutive observations of the same track, per frame, in frame-normalised units and again divided by box height; acceleration is the change in that speed, and turning angle the change in heading. Pose statistics are keypoint confidence, the standard deviation of keypoint x and y positions, ankle and wrist height, and torso length. The nine count features are the per-slice number of detections of person, bicycle, car, motorcycle, bus and truck, the number of detections of any other class, the total number of observations and the

Table 4: Selected-head paired difference in AUROC against the object-count control, on identical rows. Video-clustered bootstrap, 99 % CI, Holm–Bonferroni at family-wise $\alpha = 0.01$. Rows are ordered by effect size, so Holm-adjusted p-values need not be monotone down the column. Both arms in each pair use their selected head (Table 2); the fixed-head version is Table 3.

Arm − counts	Δ	99 % CI	p_{Holm}	
cx only (3-d)	-0.213	[-0.318, -0.091]	0.0020	loses
agent-ROI appearance (128-d)	-0.212	[-0.335, -0.095]	0.0020	loses
ROI + counts	-0.207	[-0.324, -0.088]	0.0020	loses
agent behaviour (34-d)	-0.134	[-0.217, -0.031]	0.0025	loses
pooled state (128-d)	-0.131	[-0.246, -0.018]	0.0030	loses
behaviour + counts	-0.130	[-0.209, -0.036]	0.0025	loses
tokens frame-global (128-d)	-0.115	[-0.191, -0.029]	0.0030	loses
position (1-d)	-0.050	[-0.120, +0.040]	0.1565	n.s.

Table 5: Selected-head AUROC across 25 seeds, ordered by spread. The seed re-draws the fold partition, head initialisation and minibatch order together, and the head is re-selected per arm in every seed; features and rows are identical. Spreads are computed before rounding.

Arm	Mean	Range over partitions	Spread
position (1-d)	0.630	[0.618, 0.640]	0.022
agent counts (9-d)	0.682	[0.662, 0.711]	0.049
agent behaviour (34-d)	0.579	[0.552, 0.615]	0.064
behaviour + counts	0.586	[0.556, 0.627]	0.071
ROI + counts	0.525	[0.478, 0.575]	0.097
agent-ROI appearance (128-d)	0.519	[0.469, 0.570]	0.101
pooled state (128-d)	0.564	[0.514, 0.615]	0.102
tokens frame-global (128-d)	0.569	[0.502, 0.612]	0.110
cx only (3-d)	0.487	[0.396, 0.582]	0.186

number of distinct tracks. A slice with no detection has all counts zero; its behaviour statistics are undefined and are imputed with the column mean over all scored rows before standardisation.

B TABLES BEHIND THE MAIN-TEXT FIGURES

Figures 2 and 3 plot the first two of these; the values are given here so that the exact intervals and corrected p-values can be read off, and the leaderboard of Section 4.4 is reproduced in full rather than summarised in prose.

Table 6: NWPU-Campus under the benchmark's own normal-only protocol. Published figures are as reported by Zhang et al. (2025) and Liu et al. (2025); rows without a macro value are reported micro-only in those sources, and published rows are other authors' runs, not paired experiments. Our rows are the weakest and the strongest normal-only count estimator (Appendix G) and the representation's prespecified cell (Table 7; * macro exploratory). Counting is the strongest arm inside the probe protocol of Table 2 and below every published entry outside it.

Method	Micro-AUC	Macro-AUC	Provenance
ADSM (2025) (Zhang et al., 2025)	76.9	78.1	published
SSAE (TPAMI 2025) (Cao et al., 2025)	75.6	76.9	published
CRCL (2025) (Liu et al., 2025)	69.7	–	published
FBAE (CVPR 2023) (Cao et al., 2023)	68.2	–	published
AMMC-Net (Cai et al., 2021)	64.5	–	published
MPN (Lv et al., 2021)	64.4	–	published
HF2-VAD (Liu et al., 2021)	63.7	–	published
MNAD (Park et al., 2020)	62.5	–	published
OG-Net (Zaheer et al., 2020)	62.5	–	published
MemAE (Gong et al., 2019)	61.9	–	published
counts, per-scene Gaussian (ours)	54.1	57.7	9-d, normal-only
counts, per-scene kNN (ours)	60.6	58.8	9-d, normal-only
frozen state, per-scene kNN (ours)	59.1	67.2*	128-d, normal-only

Table 7: The full normal-only grid behind Section 4.4: micro / macro frame AUC for the frozen state at three PCA dimensions under 5 per-scene estimators, beside the same estimators on the 9-d counts and the 3-d complexity features, and position alone. Artefact `nwpu_campus_x2c_vjepa_semisup.json`, seed 1969585517; hyperparameters are those tuned for counts and were not re-tuned for the representation. * marks the prespecified primary cell (128-d, per-scene kNN, the estimator that was best on counts). It is also the maximum of the 15 representation cells, so its level is upward-biased by that coincidence; the paired differences in Section 4.4 are the quantity to read. Macro rises with the retained dimension under every estimator except the histogram.

Estimator	state d=16	state d=32	state d=128	counts 9-d	cx 3-d
Gaussian	53.6 / 60.6	55.0 / 64.4	56.1 / 65.7	54.1 / 57.7	48.5 / 49.0
GMM	58.3 / 61.5	55.4 / 63.9	56.0 / 65.9	55.1 / 58.5	48.9 / 54.7
histogram	54.4 / 57.0	51.1 / 52.1	50.8 / 51.9	55.8 / 55.9	49.1 / 49.8
isolation forest	47.0 / 59.0	50.1 / 62.8	54.1 / 64.8	60.4 / 59.6	46.9 / 56.0
kNN	54.7 / 62.7	56.8 / 65.0	59.1 / 67.2*	60.6 / 58.8	49.7 / 56.8
position-in-video alone, no estimator: 54.6 / 53.9					

C WHAT PREDICTION-ERROR AND SELF-NORMALISED SCORES MEASURE

Two analytical results sit behind the protocol of Section 3, and this is where they are stated properly. The first says that the quantity a prediction-error detector actually measures is a property of the conditional *spread*, not of the event; it is the reason the scene-complexity control is in the protocol. The second says that the standard fix for that problem—subtracting the model's own entropy—removes the confound and the signal at the same time; it is an impossibility argument for one class of unsupervised anticipation score, and the reason no such arm appears in the study. Neither result supplies the other three controls (video identity, position, counts), which are empirical and were found the way Appendix E records. Proofs are in Appendix D; both are short.

Setting. A frozen encoder maps a context window to x and a future window to y. A density model $p_N(y \mid x)$ is fit on normal video only. Two families of score are in use: *post-hoc* scores, which see the observed future, and *pre-hoc* scores, which sample plausible futures from the model and ask whether the normal regime covers them. Anticipation requires the second.

C.1 THE RESIDUAL IS A DISPERSION FUNCTIONAL

Proposition 1 (Residual = dispersion functional)**.** *Let $\hat{y} = \mathbb{E}[y \mid x]$ be the L_2-optimal point forecast. Then $\mathbb{E}[\|y - \hat{y}\| \mid x]$ is a dispersion functional of $p(y \mid x)$. If $p(\cdot \mid x)$ belongs to a location-scale family with scale $\sigma(x)$, the residual is* proportional *to $\sigma(x)$.*

The consequence is not subtle. A prediction-error detector ranks windows by conditional spread. On surveillance video, conditional spread is dominated by how much is moving in the frame, so such a detector ranks *busy* scenes above quiet ones whether or not anything anomalous occurs. The canonical instance is the future-frame baseline of Liu et al. (2018). This is a statement about the estimand, not about any particular architecture. It is also the reason the scene-complexity floor is in the protocol: if the estimand is dispersion, a control built from frame-level activity is the right thing to ask a score to beat.

We state it as motivation rather than as a finding about any published system. Whether the residual of a *latent*-space predictor still tracks motion—a target encoder is trained to discard unpredictable detail, so it need not—is an empirical question, and the cx-only arm of Table 2 is where we measure it rather than assume it.

C.2 SELF-NORMALISATION, AND THE TRAP IT SETS

The natural correction is to subtract the model's own conditional entropy. The resulting score is the typicality statistic of Nalisnick et al. (2019), proposed there for out-of-distribution detection; we are not introducing it, and Proposition 2 records the property of it that matters for anticipation.

Proposition 2 (Self-normalisation)**.** *For any conditional density p_N and every x, $\mathbb{E}_{y \sim p_N(\cdot \mid x)}[-\log p_N(y \mid x)] = H[p_N(\cdot \mid x)]$ exactly. Hence the self-normalised score*

$$S_{\mathrm{cal}}(x, y) \;=\; -\log p_N(y \mid x) \;-\; H[p_N(\cdot \mid x)] \tag{1}$$

satisfies $\mathbb{E}_{y \sim p_N(\cdot \mid x)}[S_{\mathrm{cal}}(x, y)] = 0$ for every x, regardless of the conditional's spread.

Proposition 2 is assumption-free: no distributional premise, no location-scale assumption, no appeal to how well p_N is fit. It removes the confound of Proposition 1 *identically*, and the cancellation is not specific to a scalar scale: for any affine conditional $y = \mu(x) + A(x)\varepsilon$ with invertible $A(x)$ the volume term $\log|\det A(x)|$ appears in both $-\log p_N$ and $H[p_N]$ and cancels (Appendix D), so a full-covariance head does not escape it. A crowded plaza has high conditional entropy and high raw negative log-likelihood, and the two cancel; a person running across an empty walkway is predictable in the L_1 sense yet sits far in the tail of p_N, and S_{cal} fires.

That same identity is what makes the score useless for anticipation. Define the pre-hoc hazard at horizon k,

$$h_k(x) \;=\; \mathbb{P}_{\,y \sim p_N(\cdot \mid x,\, k)}[\, S(x, y) > \tau \,], \tag{2}$$

with τ a conformal quantile of S computed on *observed* normal futures. This is the standard construction: sample where the world might go, and count how much of that mass the normal regime rejects.

Corollary 1 (The hazard trap)**.** *Take $S = S_{\mathrm{cal}}$ and draw the futures in equation 2 from p_N itself. Then $h_k(x)$ is a functional of the* shape *of the model's own conditional $p_N(\cdot \mid x, k)$ and of nothing else: it is the normal model's self-rejection rate, and neither the realised future nor the true law q enters it. In particular, if $p_N(\cdot \mid x, k)$ is location-scale—$y = \mu(x,k) + \sigma(x,k)\,\varepsilon$ with $\varepsilon \sim p_0$ fixed—then S_{cal} is pivotal, with the same model-side law for every x and k, so $h_k(x) \equiv h$ is one constant for every context and every horizon, and the AUROC of any ranking induced by $\{h_k\}$ is exactly $\frac{1}{2}$. Outside that family $h_k(x)$ can vary with x and k, but only through the model's belief about the shape of a normal future. The value of h is fixed by τ, which is calibrated on* observed *futures whose mean is displaced from the model side by $\mathrm{KL}(q \,\|\, p_N) + H[q] - H[p_N]$ for the true law q (Appendix D; the entropy term is not negligible and is often omitted). If S_{cal} is bounded above under p_N and τ exceeds that bound, $h = 0$; for a model with unbounded log-density—a Gaussian head, a flow with a Gaussian base—h is a strictly positive constant in population, and is zero only empirically, once τ exceeds the largest of finitely many draws.*

The constancy is the claim; the zero is a special case of it, and stating the zero alone would be weaker and, for a model with unbounded log-density, false. The two quantities in the corollary measure different things, and that is the whole difficulty: self-normalisation pins the model side by construction, while the threshold measures model–data divergence. The gap between them is not a tuning error to be fixed by moving τ. Moving τ down far enough to admit model draws makes it fire on normal observed futures at the same rate; moving it up sends h to zero. There is no setting of τ at which h_k tracks the realised future, because the realised future never enters it; what can vary is the model's own tail mass, a statement about p_N and not about q. Stated generally: *a pre-hoc hazard that thresholds a self-normalised score on model-drawn futures against a threshold calibrated on observed futures estimates the normal model's own tail mass, not the probability that the future is anomalous, at any horizon; for a location-scale model it is one constant.*

Numerical verification. We check the identities against artefact `prop2a_numeric_check.json`. Against a closed form, the estimated entropy matches the analytic value to 0.0081 nats. On the untrained spline-flow architecture—the sharper test, since Proposition 2 must hold for *any* density, so a deviation would be an implementation bug rather than a modelling failure—the model-side mean of S_{cal} is +0.0175 ($z = +1.87$). Sweeping the model–data divergence, the hazard is constant in x in every setting—the spline flow's conditional shape is not fixed across x, so this is what a threshold above the model-side support forces rather than what the corollary guarantees; it is numerically zero in 3 of 5 of them, once τ clears the model-side support, and 0.0104 at the smallest divergence. The constant, not the zero, is what is being confirmed: a Monte-Carlo hazard estimated from M draws cannot resolve a value below $1/M$, so a numerical zero is evidence of a small constant rather than of an exact one. Appendix E reports what happened when we ran this construction on real video, and why that run does not test the corollary either.

C.3 What escapes

The trap is a property of scores whose expectation under the sampling law is pinned. A likelihood *ratio* between two regimes is not: $S_{\mathrm{det}} = \log p_D - \log p_N$ has $\mathbb{E}_{y\sim p_N}[S_{\mathrm{det}}] = -\mathrm{KL}(p_N \| p_D) \leq 0$ and $\mathbb{E}_{y\sim p_D}[S_{\mathrm{det}}] = +\mathrm{KL}(p_D \| p_N) \geq 0$, both strictly non-zero whenever the two diets differ. We measure -0.619 and +0.653 on untrained flows, where the identities must hold for any pair of densities.

The escape is worth naming precisely, because it is narrower than it looks: p_D is a density of the *anomalous* regime, so building S_{det} requires anomalous data. The analytical message and the empirical one therefore agree. An unsupervised self-normalised pre-hoc hazard estimates the normal model's own tail mass rather than the future, by construction, and the supervised alternative is the one that has to be measured against floors—which is what the rest of this paper does.

The theory's scope. Corollary 1 is an identity about one construction—a pre-hoc hazard on a self-normalised score with model-drawn futures—and rules out only a hazard into which the realised future never enters. Its constancy is proved for location-scale conditionals, scalar or affine; the flow of Appendix E.2 is not one, and score-based or autoregressive predictors whose conditional shape varies with the state (Zhang et al., 2025; Cao et al., 2025) are not either, so for such models

the corollary bounds what the hazard can estimate—the model's own tail mass—and says nothing about whether that mass varies usefully with x; the chance result for our flow is empirical. A score whose expectation under the sampling law is not pinned escapes it; the ratio construction above is the natural candidate, though it needs a density of the anomalous regime, so the escape is supervised rather than free. Whether such a score carries usable anticipatory signal on real video is open: ours did not (Appendix E), and we report that as a negative result about our own attempt rather than as evidence against the possibility. Its support here is analytical, checked on synthetic densities.

D PROOFS

Throughout, x is a context, y a future window, and $p_N(\cdot \mid x)$ the density fit on normal video. All entropies are differential and assumed finite; where an expectation is taken the corresponding integral is assumed to exist.

PROPOSITION 1

Proof. Write $\hat{y} = \mathbb{E}[y \mid x]$ and $R(x) = \mathbb{E}[\|y - \hat{y}\| \mid x]$.

(i) R is a dispersion functional. Let T_c denote translation $y \mapsto y + c$. Under T_c the conditional mean maps to $\hat{y} + c$, so the deviation $y - \hat{y}$ is unchanged and R is translation-invariant. Under scaling $y \mapsto \lambda y$ with $\lambda > 0$ we have $\hat{y} \mapsto \lambda\hat{y}$ and, by absolute homogeneity of the norm, $R \mapsto \lambda R$. A functional of $p(\cdot \mid x)$ that is translation-invariant and positively homogeneous of degree one is a measure of dispersion; R depends on $p(\cdot \mid x)$ only through its centred law.

(ii) Proportionality under a location-scale family. Suppose $y = \mu(x) + \sigma(x)\,\varepsilon$ with $\sigma(x) > 0$ scalar and $\varepsilon \sim p_0$ fixed, independent of x, with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}\|\varepsilon\| < \infty$. Then $\hat{y} = \mu(x)$, so $y - \hat{y} = \sigma(x)\varepsilon$ and

$$R(x) \;=\; \mathbb{E}[\|\sigma(x)\varepsilon\|] \;=\; \sigma(x)\,\mathbb{E}_{p_0}[\|\varepsilon\|] \;=\; c\,\sigma(x),$$

with $c = \mathbb{E}_{p_0}\|\varepsilon\|$ a constant not depending on x. □

The assumption that $\sigma(x)$ is a scalar matters: for a matrix scale $\Sigma(x)$ the residual is a norm-dependent functional of $\Sigma(x)$ rather than proportional to a single number. The qualitative conclusion—that R ranks contexts by conditional spread and not by whether an event occurs—does not depend on it.

PROPOSITION 2

Proof. By the definition of differential entropy, for each fixed x,

$$H[p_N(\cdot \mid x)] \;=\; -\int p_N(y \mid x) \log p_N(y \mid x)\,\mathrm{d}y \;=\; \mathbb{E}_{y\sim p_N(\cdot|x)}[-\log p_N(y \mid x)]\,.$$

Subtracting, $\mathbb{E}_{y\sim p_N(\cdot|x)}[S_{\mathrm{cal}}(x,y)] = H - H = 0$ for every x, with no assumption on p_N beyond finiteness of H. □

This is an identity rather than a theorem, and we do not claim otherwise; the definition of entropy is doing all the work. Its interest lies entirely in the corollary, and in the fact that the identity holds *regardless of how badly p_N is fit*—which is what makes the consequence unavoidable rather than a symptom of undertraining.

THE DATA SIDE, AND A CORRECTION

The threshold τ is calibrated on *observed* futures, drawn from the true law q rather than from p_N. The corresponding mean is not the model-side zero:

$$\begin{aligned}
\mathbb{E}_{y\sim q}[S_{\mathrm{cal}}(x,y)] &= \mathbb{E}_{y\sim q}[-\log p_N(y \mid x)] - H[p_N(\cdot \mid x)] \\
&= \Big(H[q] + \mathrm{KL}(q \,\|\, p_N)\Big) - H[p_N] \\
&= \mathrm{KL}(q \,\|\, p_N) \;+\; H[q] - H[p_N]\,, \qquad (3)
\end{aligned}$$

using $\mathbb{E}_q[-\log p_N] = H(q, p_N) = H[q] + \mathrm{KL}(q\|p_N)$ for the cross entropy. The identity is pointwise in the context: q and p_N abbreviate $q(\cdot \mid x)$ and $p_N(\cdot \mid x)$ for a fixed x, and a threshold calibrated over a whole split sees the average of equation 3 under the marginal law of x.

We record explicitly that equation 3 is *not* $\mathrm{KL}(q\|p_N)$ alone. The entropy difference $H[q] - H[p_N]$ does not vanish in general and is not small. For Gaussians with $d = 3$, $\sigma_q = 0.9$ and $\sigma_{p_N} = 1.7$ it contributes -1.91 nats against a divergence of 0.86, so the exact value is -1.04 and the KL-only form has the wrong sign. We check the identity against Monte Carlo in the same artefact, agreeing to 0.0001 nats. Only when $H[q] = H[p_N]$ does the data-side mean reduce to the divergence. The structural argument below does not depend on which of the two forms holds, since it is stated in terms of the model-side support and an upper quantile rather than in terms of means; but the identity is worth stating correctly.

COROLLARY 1

Proof. Fix x and k, write $p = p_N(\cdot \mid x, k)$ and $S = S_{\mathrm{cal}}(x, \cdot)$. The pre-hoc hazard draws its futures from p, so

$$h_k(x) \;=\; \mathbb{P}_{y\sim p}[-\log p(y) - H[p] > \tau]\,,$$

the probability that a draw from p is less typical than τ under p itself. This is a functional of p alone—of the law of $-\log p(Y)$ for $Y \sim p$, which is determined by the shape of p—and the realised future, the true law q and the anomaly label appear nowhere in it. That is the general claim. It is not, in general, constant in x or k: a horizon-conditioned model whose conditional changes shape with k, or a flow whose conditional changes shape with x, moves $h_k(x)$ through that shape; a Gaussian and a Laplace conditional of equal entropy, for instance, reject their own draws at different rates.

Constancy needs a fixed shape. Suppose $p_N(\cdot \mid x, k)$ is location-scale: $y = \mu(x, k) + \sigma(x, k)\,\varepsilon$ with $\varepsilon \sim p_0$ fixed. Writing d for the dimension, $-\log p_N(y \mid x, k) = -\log p_0(\varepsilon) + d \log \sigma(x, k)$ and $H[p_N(\cdot \mid x, k)] = H[p_0] + d \log \sigma(x, k)$, so the scale terms cancel in equation 1:

$$S_{\mathrm{cal}}(x, y) \;=\; -\log p_0(\varepsilon) \;-\; H[p_0].$$

The same cancellation holds for any affine conditional $y = \mu(x, k) + A(x, k)\,\varepsilon$ with $A(x, k)$ invertible: by the change-of-variables formula $-\log p_N(y \mid x, k) = -\log p_0(\varepsilon) + \log|\det A(x, k)|$ and $H[p_N(\cdot \mid x, k)] = H[p_0] + \log|\det A(x, k)|$, so the volume term—which is $\frac{1}{2}\log\det\Sigma(x, k)$ for a Gaussian head with covariance Σ—cancels identically, and a full-covariance parameterisation does not escape the corollary. The model-side law of S_{cal} is then the *same* for every x and k, and $h_k(x) = \mathbb{P}_{\varepsilon\sim p_0}[-\log p_0(\varepsilon) - H[p_0] > \tau] \equiv h$ is one constant for every context and every horizon. A ranking induced by a constant score has AUROC exactly $\frac{1}{2}$, and any survival or time-to-event quantity assembled from $\{h_k\}$—for instance $1 - \prod_k (1 - h_k)$—is a function of that constant.

Finally, the value of h. By Proposition 2 the model-side law of S is centred at zero with spread set by p_N's own log-density fluctuation, whereas τ is an upper quantile of S under q, whose location is displaced by equation 3. As the model misfits the data—in the sense that the displacement grows relative to the model-side spread— h falls; if S is bounded above under p_N and τ exceeds that bound, $h = 0$ exactly. □

Why the constancy, and not the zero, is the claim. An earlier form of this corollary asserted $h_k \equiv 0$ under the condition $\tau \geq \operatorname{ess\,sup} S$. That condition is unattainable for the models actually used: whenever p_N has unbounded log-density—a Gaussian head, or a normalising flow with a Gaussian base—the essential supremum is infinite and no finite τ satisfies it. In population h is then a strictly positive tail probability. It is nonetheless a *constant* tail probability, which is why the AUROC conclusion is unaffected and why the corollary is stated in the constant form. This also disciplines how the numerical check is read (Appendix C, artefact `prop2a_numeric_check.json`): sweeping the divergence, the hazard is constant in every setting, numerically zero once τ clears the sampled range and 0.0104 at the smallest divergence. A hazard estimated from M Monte-Carlo draws cannot resolve a value below $1/M$, so a measured zero is evidence of a small constant and not of an exact one.

PROPOSITION 3 (RATIO IDENTITIES, AND WHY THEY ESCAPE)

Proposition 3 (Ratio identities)**.** *Let $S_{\mathrm{det}}(x, y) = \log p_D(y \mid x) - \log p_N(y \mid x)$ for two conditional densities with common support. Then*

$$\mathbb{E}_{y\sim p_N}[S_{\mathrm{det}}] = -\mathrm{KL}(p_N \| p_D) \le 0, \qquad \mathbb{E}_{y\sim p_D}[S_{\mathrm{det}}] = +\mathrm{KL}(p_D \| p_N) \ge 0,$$

and both are strictly non-zero whenever $p_N \neq p_D$ on a set of positive measure.

Proof. $\mathbb{E}_{p_N}[\log p_D - \log p_N] = -\mathbb{E}_{p_N}[\log(p_N/p_D)] = -\mathrm{KL}(p_N \| p_D)$, and symmetrically for the second. Non-negativity of the Kullback–Leibler divergence gives the signs, and $\mathrm{KL}(p\|p') = 0$ if and only if $p = p'$ almost everywhere gives strictness. □

The contrast with Proposition 2 is the point: S_{cal} has its model-side mean pinned at zero by construction, whereas S_{det} has both means pinned *away* from zero whenever the two regimes differ at all. A hazard built on S_{det} is therefore not degenerate for the reason Corollary 1 gives. We verify all three identities on untrained flows, where they must hold for any pair of densities and a deviation would indicate an implementation error rather than a modelling one: the measured values are -0.619 and +0.653 (Appendix C).

REMARK: LINEAR RESIDUALISATION AS A FIRST-ORDER CONTRAST

Propositions 1 and 2 each identify a context-only nuisance term—a spread $c\,\sigma(x)$, an entropy $H[p_N(\cdot \mid x)]$—that a raw score carries. If such a nuisance enters additively, $S_{\mathrm{raw}}(x, y) = S^\star(x, y) + g(c(x))$ for an unknown smooth g of observable complexity proxies $c(x)$, then expanding g to first order about $\bar{c}$ gives $g(c) \approx g(\bar{c}) + \nabla g(\bar{c})^\top (c - \bar{c})$, and least-squares residualisation of S_{raw} on $[1, c]$ removes this term *provided the signal is uncorrelated with the proxies*: the population slope is $\mathrm{Var}(c)^{-1}\,\mathrm{Cov}(c, S_{\mathrm{raw}}) = \nabla g(\bar{c}) + \mathrm{Var}(c)^{-1}\,\mathrm{Cov}(c, S^\star)$ plus the Taylor remainder, so when $\mathrm{Cov}(c(x), S^\star(x, y)) \neq 0$—anomalies that co-occur with crowding, for instance—the projection absorbs part of the signal along with the nuisance. When the residualiser is fit on normal training video only (as in Appendix E.1) the premise holds by construction; when it is cross-fitted label-free across an evaluation split containing anomalies (as in Section 4.2) it is an assumption, and the adjustment is conservative in the sense that it can only remove correlated signal, which is why raw scores are reported beside adjusted ones. The remark covers the context-only nuisances of Propositions 1 and 2; the ratio S_{det} of Proposition 3 contrasts two densities of the realised y and is not approximated by a regression on context proxies. We state this as a remark rather than a proposition: it is a Taylor argument, it presumes the nuisance is additive and a function of the observed proxies alone, and neither presumption is verifiable from data.

E NEGATIVE RESULTS: OUR OWN ANTICIPATION ATTEMPTS

The floors in Section 3 were not built in the abstract. They came out of attempts to make an anticipation score work, and this appendix reports them rather than leaving them out. The first applies the video world model's own predictor to tokens from a full-window encode; the second replaces that point predictor with a full conditional density; the third asks whether pooling the tokens was to blame; the last runs the supervised alternative the theory points at. Pooled across videos, none of the unsupervised scores separates anomalous video from normal video; the first does per video, which is reported with equal weight; and the supervised arm clears a floor on one corpus only by measuring a different one. Nothing here is evidence that anticipation is impossible, and the runs that failed their own gates are marked as such and excluded from every conclusion.

E.1 THE WORLD MODEL'S OWN PREDICTOR, RUN ON FULL-WINDOW TOKENS

The first thing a reader from the world-model literature will ask is whether we ever ran V-JEPA 2's *predictor*, or only its encoder. The arms of Section 4.2 use frozen encoder features, and the density attempt below fits our own model on top of them, so on the evidence of the main text alone that objection would stand. We did run it, and this subsection reports how—including the respect in which the run falls short of the architecture's own contract. Artefacts: `e9_result.json`, `e9_refine_result.json`.

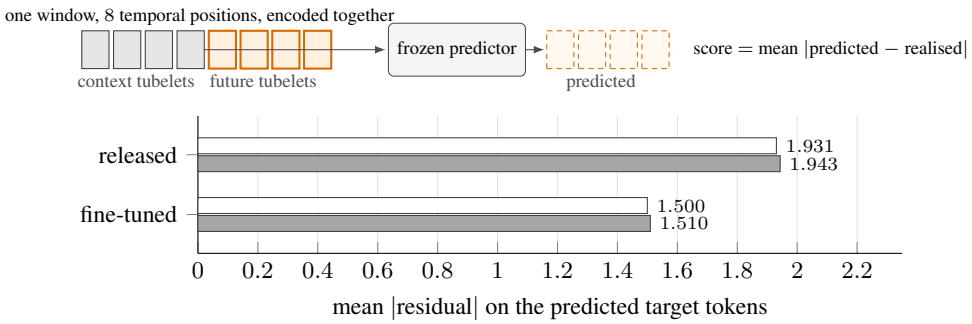

Figure 5: **Top:** how the predictor was called. Context tubelets are supplied as the context mask and the *future* tubelets of the same window are requested as the prediction target; the score is the mean absolute residual between the predicted and realised target tokens. The whole window was encoded together before the split, so the context tokens had already attended to the target frames—a weaker contract than pretraining's, where the mask precedes the encoder (Appendix E.1). **Bottom:** what fine-tuning on normal-only video changed; hollow bars are windows that are entirely normal, solid bars entirely anomalous. The residual fell a long way—0.431 on windows that are entirely normal and 0.433 on windows that are entirely anomalous—while the gap *between* normal and anomalous, the only quantity an anomaly score needs, stayed where it was. The model became a better predictor of everything, which is what future-token leakage predicts as readily as any property of normality. Pooled across videos both arms sit at chance (0.502 released, 0.488 fine-tuned); read per video the same residuals reach 73.4 (Appendix E.1), so the pooled chance verdict is a statement about aggregation under a leaky contract, not about the predictor.

The setup. The pretrained predictor is loaded from the released checkpoint and called with context tubelets supplied as the context mask and *future* tubelets requested as the target mask. The score for a window is the mean absolute residual between the predicted and the realised target tokens.

What the run does not test. In pretraining the mask is applied *before* the encoder: the context encoder never sees the target tubelets, and the predictor must produce them from context alone. Here the frozen encoder was run once on the whole 32-frame window and the resulting 2048 tokens were split afterwards into the first four temporal positions (context) and the last four (target). Because the encoder attends across the entire window, every context token had already attended to the target frames before the predictor saw it; a mask applied at the predictor cannot remove information the encoder has already mixed in. The task the predictor faces is therefore auto-associative reconstruction under future-token leakage rather than forecasting, for normal and anomalous windows alike. The arm evaluates the predictor as a *readout* of features that already contain the target; it establishes neither an upper nor a lower bound on what a causal, masked-at-encoder forecaster would achieve, and the leakage itself is a sufficient mechanism for the residual to fall equally on both classes below. The corrected run—encode context tubelets with the target tubelets masked out at the encoder, predict, and compare against the target tokens of a full encode—requires re-encoding the corpus and is not reported here. We discovered the discrepancy in an external audit after the arm had been written up, and we report it rather than reword the arm around it. Two arms are scored identically: **(A)** the predictor exactly as released, and **(B)** the same predictor fine-tuned on normal-only training video of the evaluation corpus, minimising that same residual, with the encoder frozen throughout and fit/validation split by source video. This is on ShanghaiTech (Luo et al., 2017) rather than the two corpora of the main text, which is part of why it lives here; scoring covers 40,680 frames, with 111 frames falling outside every window and dropped rather than imputed. Scores are corrected for scene complexity by a nuisance model fit on the training split only—all normal, no labels—and applied unchanged to test, and intervals are bootstrapped over test sequences.

Pooled across videos, neither arm is distinguishable from chance. The released predictor reaches 0.502 corrected (0.596 raw), interval $[0.443, 0.560]$. Fine-tuning it on normality gives 0.488 (0.544 raw), interval $[0.429, 0.552]$. The difference between them is -0.015, interval $[-0.091, +0.067]$, $p = 0.66$: under this contract, fine-tuning a video world model's predictor on normal video does not make its surprise a better anomaly signal, and the point estimate moves the

wrong way. Both verdicts carry the readout caveat above, and both are pooled verdicts; the next paragraph is the same residuals read per video.

The chance verdict is a pooled verdict. Both numbers above pool every window in the corpus into one ranking, and Section 4.4 has shown that pooling across videos is where this representation loses its signal. Re-reading the *same saved residuals* per video says something different: the fine-tuned predictor reaches 73.4 per-video AUROC over 104 videos and the released one 70.4, against a position-in-video floor of 59.0—a paired +14.4 [+5.6, +23.8] for the fine-tuned arm. Pooled over the same per-window re-read (a window labelled anomalous when more than half its frames are; uncorrected), the same scores read 53.9; standardising them within each video lifts that to 69.0 with no refitting, an offline normalisation that sees the whole video. So the predictor's residual carries within-video anomaly information, and the collapse to chance is specific to unstandardised residuals pooled across heterogeneous cameras, which is what Proposition 1 predicts. We keep the arm in this appendix: the per-video metric was not pre-specified, the gap between the two arms remains within its interval, and the residuals being re-read are the same leaky full-window residuals. The sentence that survives is that aggregation changes the verdict, not that the predictor's surprise is empty.

Anticipation, though, stays empty. The obvious next move is to give that residual a lead time—anchor on windows that are entirely normal and ask whether the anomaly begins within α. It does not work, and the reason is instructive. A window is 32 raw frames scored on its second half, so an anchor admitted merely for being *mostly* normal can contain the onset in the very frames the residual is computed against; defined that way the arm appeared to clear its floor at three horizons. Restricting anchors to windows with no anomalous frame at all and measuring the lead from the window's end, 0 of 5 horizons from 0.32 s to 5.12 s clear the position floor; the best is +2.9 [-8.8, +14.7]. We report the leaky version's existence because it is the trap: at these horizons, anticipation and detection are separated by a single window, and an anchoring rule that is off by one turns the second into the first.

This null is weak and we will not dress it as strong. Restricting anchors to purely normal windows leaves as few as 1 positive anchor per video at the shortest horizon, so the design cannot resolve a margin between 11 and 15 points, depending on the horizon—an effect of the size we report elsewhere in this paper would pass through it undetected. What we can say is that this design is inconclusive on ShanghaiTech: the corpus does not carry enough pre-onset material to settle the question in either direction, and a null under family-wise control at this power is not evidence that precursors are absent. A corpus built for lead time, rather than cut around the event, would be needed to do better.

What the fine-tuning actually learned. The residual did drop substantially—the model became a much better predictor. It simply became better everywhere. Restricting to windows that are *entirely* normal (2,201) or *entirely* anomalous (1,561), so that no window mixes the two, the residual falls by 0.431 on the normal windows and 0.433 on the anomalous ones. The difference of drops is -0.002, interval $[-0.023, +0.018]$ over 5,000 resamples—centred on zero. Under this contract, training on normal-only video made the predictor better everywhere rather than selectively on normality. Because the context tokens had already attended to the target, an equal drop on both classes is also what leakage predicts, so we do not attribute the result to a mechanism; it is a description of what this run did, not of what a causal forecaster would do.

Why this belongs in front of the audit. Two negatives of different kinds point the same way. A latent-space predictor's own surprise, pooled across videos and read from a leaky full-window encode, is at chance, and fine-tuning it on normal video does not move that pooled number. Replacing the point predictor with a full conditional density—the obvious next move, since a point estimate cannot represent a distribution—does not help either, as the rest of this appendix reports. Only after both failed did we begin measuring the instrument instead of the model, which is the protocol of Section 3. We state the order because it is the honest one: the floors were learned, not derived. What none of this licenses is the stronger claim that a video world model's latents contain nothing about anomaly—these are two corpora, one architecture, and two score families, and Section 5 keeps the claim inside those bounds.

E.2 A CONDITIONAL DENSITY ON THE SAME LATENTS

Artefact: `nwpu_campus_e12.json`.

The attempt. Because a point estimate cannot represent a distribution, the natural response to Appendix E.1 is to score futures under a density rather than a residual. We fit a conditional density $p_N(y \mid x)$ on *mean-pooled* frozen encoder features of the corpus's normal training video—a rational-quadratic spline flow over a PCA basis, with x a context window and y a future window—and scored test video under the anticipation protocol of Cao et al. (2023): the score at time t uses only observations up to t, and the label covers only frames after t. Six horizons were swept, from 0.5 to 3.0 seconds.

It does not work. Across all 6 horizons the frame-level AUC sits between 57.9 and 58.5, against published values of 64.0–65.8 for the same protocol. It beats none of them, and more importantly the $99\,\%$ interval spans chance at every horizon—the widest runs $[48.0, 67.8]$—leaving 0 of the 6 horizons distinguishable from a coin flip. We report the point estimates only because omitting them while keeping the intervals would flatter the result.

The hazard does not rise toward the event either. If a score were anticipating, its per-video trajectory would trend upward as the event approaches. Over 118 videos, 66 have a positive median slope—a sign test gives $p = 0.23$, which is what one expects from noise.

The rollout hazard, and why it is a diagnostic rather than a test. The same artefact contains the pre-hoc construction of Corollary 1 run on this corpus: τ a conformal quantile of the score on observed normal futures, futures drawn from p_N, and the hazard the fraction of draws exceeding τ. It comes out constant across every context—$\tau = 379.9$ nats, with a spread across contexts of 0.0—which is consistent with Corollary 1, though for a flow the corollary does not require it.

We do not present it as confirmation, because it cannot be. In that run the entropy term of equation 1 is estimated from the *same* model draws that are then scored, so the sample mean of S_{cal} is zero per context by construction rather than by Proposition 2; with the score centred at zero and τ several hundred nats above it, a hazard of zero is forced. A zero spread across contexts also makes every pair a tie, so the AUROC that follows is $\frac{1}{2}$ by definition and not by measurement. A test of the corollary on real video needs an entropy estimate independent of the scored draws, a τ placed inside the model-side support, and a draw count large enough to resolve a small constant. We have not run it. The corollary's support in this paper is analytical, plus the synthetic checks of Appendix C.

E.3 WAS IT THE POOLING?

The density above consumes a slice state that is a mean over the 256 spatial tokens, and pooling before predicting is a choice that deserves to be named rather than buried: a reader may reasonably suspect that it, not the representation, is what fails. The objection is tested on both axes. With *unpooled* tokens and a point predictor, Appendix E.1 is that experiment—under the full-window caveat recorded there—and, pooled, reaches chance. With unpooled tokens and a density—a raw-token autoregressive conditional and a PaDiM-style per-position marginal, both pooled over patches only at the very end—the scores reach chance for detection, and this subsection scores the same four under the *anticipation* protocol above. Artefact: `nwpu_campus_x4_unpooled_anticipation.json`, seed 437017364.

The protocol makes this cheap: the horizon lives in the label, not in the score. An anchor is scored from observations up to t and the label asks whether any frame in the next α seconds is anomalous, so the same per-token score serves both readings and only G_α changes. All arms are scored on the 93,506 anchors common to the pooled and unpooled sets, recomputed per α because G_α drops anchors whose horizon runs past the end of the video.

Unpooling did not help in this run. Across 4 unpooled scores and 6 horizons the AUC runs from 49.5 to 53.2, the best being PaDiM-marginal, mean-pooled at $\alpha = 3.0\,\mathrm{s}$. 0 of the 24 arm-horizons have a $99\,\%$ interval excluding chance, 0 reach the published figure, and on a paired test over identical anchors 0 beat the pooled score—which itself runs 57.9–58.5 and so sits *above* every

unpooled arm. We do not conclude from this that spatial unpooling removes signal: the run failed its own gates (below), so the degradation could as easily be threshold drift or optimisation instability in a high-dimensional token space as a property of unpooling. The one descriptive observation that survives is that the best unpooled arm is the one with *no dynamics at all*—a per-position marginal—consistent with Appendix F.

The comparator reproduces, which is what makes the rest readable. The pooled arm here is rebuilt from scratch: a separate script, its own anchor selection, its own G_α. It lands on the numbers of the previous subsection at all 6 horizons, with a maximum difference of 0.00 AUC and identical anchor counts. Had it not, the unpooled figures beside it would mean nothing. The pooled comparator nevertheless inherits that run's gate status below, and the paired differences inherit it with them.

Gate status, stated rather than omitted. This artefact does not meet the standard the rest of the paper holds to, which is why nothing in the main text depends on it. It records `ran_under_w_override: true`; its own pre-registered acceptance gate is recorded as failed, as are the two calibration gates—the realised false-alarm rate exceeds its nominal budget by a factor of two to four, meaning exchangeability broke between the calibration and evaluation normals. The artefact does record its resolved seed (437017364), so the unpooled arms are reproducible as stored; the pooled comparator it inherits from Appendix E.2 records none, so the paired differences against it are not. We report it because a negative result about our own method is what motivated the protocol, and because the alternative—describing the floors as though they had been derived rather than learned the hard way—would misrepresent how this work happened.

E.4 The supervised alternative, tested

Appendix C argues that what escapes Corollary 1 is a likelihood ratio between regimes, so the informative alternative is *supervised*. That is an assertion about an experiment we had not run. We ran it, on both corpora of the main text, and it is reported here because one corpus says nothing and the other says something the paper needs.

The task and its two gates. An anchor is scored before the onset and labelled by whether the anomaly begins within α; we report $\alpha = 1.5\,\text{s}$. An anchor's time is the moment its features can exist: the encoder attends across its whole 32-frame window, so none of a window's eight slices is available before the window's last frame, and every lead below is measured from there (140 positive windows at the headline horizon on NWPU-Campus, 186 on UCF-Crime). Positives are anchors in anomalous videos; negatives are anchors in *normal* videos, resampled so their position-in-video distribution matches the positives'. Folds are grouped by video, out-of-fold scores are residualised on complexity, intervals are video-clustered, and Holm is applied over the horizon family—the protocol of Section 3 throughout. Two properties are measured *before* any probe is fit. The *position gate* asks how well anchor position alone predicts the label inside an anomalous video; the *identity gate* is the first-window diagnostic of Section 4.1. NWPU-Campus scores 70.3 on the position gate and 49.3 on the identity gate; UCF-Crime scores 65.3 and 89.9. The learned arm throughout is the pooled 128-d V-JEPA 2 state (a state + delta variant is also run and never wins), with the better of a linear and an MLP head reported, MLP at the headline horizon on both corpora. Artefacts: `nwpu_campus_x7_anticipation.json` (seed 744605746) and `ucf_crime_x7_anticipation.json` (seed 2069931602).

NWPU-Campus: nothing that survives correction. With 123 anomalous and 118 normal videos, the best learned arm reaches 66.6 against a position floor of 49.8: a paired +16.8, interval [+0.5, +33.0], and 0 of 4 horizons clear the floor after Holm. The point estimate sits above position at every horizon, but the intervals include zero at three of the four and the fourth does not survive the family-wise level; with 140 positive windows the design is not powered to settle a margin of this size, and we do not read it as one. Trained to separate near-onset from far-from-onset anchors *within the same video*, it reaches 59.4 against 77.6—worse than position. Supervision does not produce an anticipation result that survives multiplicity here, and the result is inconclusive rather than negative: the intervals are wide enough to contain a margin of the size reported elsewhere in this paper. That

this corpus's anomalies, largely rule violations rather than physical events, lack visible precursors is one reading; low power at this clip count is another, and we cannot separate them.

UCF-Crime: the trap, and it is the paper's own thesis. The same arm reaches 89.7, clearing the position floor of 49.9 by +39.8, interval [+27.3, +53.0], with all 4 horizons surviving Holm. Read on its own that is a strong anticipation result—stronger than anything else in this paper—and it is not anticipation at all. Trained to tell near-onset from far-from-onset anchors *inside the same anomalous video*, where provenance is held constant and only imminence can separate, the same features reach 63.3 against a position floor of 70.1: -6.8, interval [-16.6, +2.2], 0 of 4 horizons. The model cannot locate the event inside a video it has already recognised. What the matched design measured was the identity floor—89.9 on this corpus—read back to us through a supervised head.

Where the score actually sits. The clearest way to see what the UCF arm learned is to ask where in an anomalous video it puts its score. Take the same matched-trained model and score every window. On footage more than twenty seconds *before* the onset—where, by construction, nothing is happening—it separates anomalous from normal videos at 89.9. Asked instead to tell that same far-from-onset footage from the moments immediately before the event, inside the same videos, it reaches 55.9. Almost all of the 89.7 is available before anything occurs. The arm separates *corpora*, not moments, and no amount of lead time is being predicted.

NWPU-Campus inverts every part of this. Its far-from-onset footage is indistinguishable from its normal videos—49.2, chance—and its within-video imminence separation, 62.1, is slightly *higher* than UCF's. The corpus with the far worse headline number is the one whose signal, such as it is, is about the moment rather than the source. A matched-position AUROC ranks these two corpora in exactly the wrong order.

Why we report it. This is the sharpest instance of the paper's argument we have, and it arrived after the protocol was fixed rather than being designed to illustrate it. A design can match the confound's distribution exactly, clear its floor by +39.8 with corrected intervals under multiplicity control, and still be measuring provenance. Matching a floor is not the same as removing it, and the control that separates the two—holding the video constant and varying only the distance to the event—is not part of the protocol we proposed in Section 3. It should be. This is precisely the "floor we did not think of" that Section 5 says we would want back, and we found it in our own work.

A timing correction, disclosed. An earlier version of this experiment timed each anchor at the first frame of its own four-frame slice rather than at the end of its encoder window—up to twenty-eight frames before the features could exist, so an anchor could be labelled "before the onset" while its window had already read the onset. An external audit caught it after the arm had been written up. The correction changes the positive set in both directions: it drops slices whose window straddled the onset, and admits the earlier slices of windows that end within α, which the slice-start rule had placed beyond the horizon. Under the slice-start rule the UCF-Crime matched arm read 86.5, its within-video arm 63.5 and its imminent-versus-far separation 58.4; NWPU-Campus's matched arm read 54.3. Every number above is from the corrected run; the earlier artefacts are released beside them with the suffix `slicestart`. The reading did not depend on the leak—identity is available at every point in a clip, which is the finding. The numbers a leak would flatter are the within-video and imminence ones: on UCF-Crime they fell, to 63.3 and 55.9; on NWPU-Campus the within-video arm fell to 59.4 while the matched point estimate rose to 66.6, inside intervals wide enough that neither movement is a result. The unsupervised anticipation arms of Appendices E.2 and E.3 time their anchors at the end of the scored slice rather than at the end of its encoder window, which carries the same defect one level down: a scored slice at temporal position j of its window is available $(7 - j) \times 4$ raw frames later than its own end, up to 28 frames or 1.12 s at 25 fps, so those arms carried up to that much lookahead relative to their nominal anchor and their conformal thresholds were calibrated under it. They were at chance with it, and we have not re-timed them; their negatives are reported with that caveat.

F TOKEN READ-OUTS: MEAN, MAX AND LEARNED ATTENTION

The 6 token read-outs of Section 4.2, run inside the *primary* fold partition (seed 1750412224) so that they may be read against Table 2 and paired against the same control on identical rows. Comparisons run in their own partition cannot be quoted beside Table 2; these can.

Two checks make the table readable. First, the object-count control re-run inside this harness reproduces the primary partition's 0.686 exactly, which establishes that the row selection and the fold assignment match. That check alone is not sufficient, because the counting arm never touches the token array: so, second, the row alignment of the token array is verified directly. The token arm of Section 4.2 and the array used here are built from the same encoder output by two independent paths—one reads the raw per-window tensors and reduces the frame mean to 64 dimensions, the other projects to 128 and writes a single array—so if the rows correspond, one is nearly a linear function of the other. Regressing the first on the second gives $R^2 = 0.992$, against 0.045 on randomly permuted rows.

Pooling	Head	AUROC	99 % CI
mean	linear	0.425†	[0.340, 0.514]
mean	mlp	0.521†	[0.432, 0.611]
max	linear	0.418†	[0.328, 0.521]
max	mlp	0.500†	[0.436, 0.567]
attn	linear	0.429†	[0.334, 0.526]
attn	mlp	0.533†	[0.452, 0.616]
object counts (9-d)	mlp	0.686	[0.611, 0.742]

† marks an interval that includes chance. The paired test the text quotes is below: each MLP pooling against the count control under its MLP head, on identical rows, Holm within this family of 3.

Arm − counts (MLP)	Δ	99 % CI	p_{Holm}	
tokens, max pooling, mlp	-0.186	[-0.277, -0.081]	0.0007	loses
tokens, mean pooling, mlp	-0.164	[-0.268, -0.047]	0.0007	loses
tokens, attn pooling, mlp	-0.152	[-0.254, -0.040]	0.0020	loses

These 3 poolings are a separate, later family from the 8 pre-registered comparisons of Section 4.2 and are Holm-corrected within themselves; folding them into that family would inflate the count of arms reported as losing to counting. Two further caveats belong with this table. The token arms are standardised as a single mean and variance over all 256 patch positions, so that no position can win the max pooling by having a larger scale; the object-count control is standardised per column, as everywhere else in the paper. These are different normalisations, and the paired test therefore compares the arms as their constructions define them rather than isolating the pooling operator. Second, the spread across the three poolings at a fixed read-out is 0.034 AUROC, while the spread across all 6 arms is 0.115: the choice of read-out head moves this arm considerably more than the choice of pooling does, and a claim that pooling is immaterial should not be read as a claim that the read-out is. The linear poolings, which Table 1 suggests would fare better against a linear count arm, were not paired here; the fixed-head families of Table 3 cover the frame-global token arm under both heads. Unlike the arms of Figure 3, these were run in the primary partition only, so no seed sweep accompanies them.

G NORMAL-ONLY ESTIMATORS ON THE OBJECT COUNTS

The counting baseline of Section 4.4 under 5 different normal-only density estimators, all on the same 9-d count features, the same 305 normal-only training videos and the same frame-level evaluation over 242 test videos. The Gaussian row is Section 4.4's own arm, re-run inside this harness: it reproduces 54.1, which is what makes the other four readable. Artefact: `nwpu_campus_x2b_estimators.json`, seed 20260906. The same five estimators applied to the representation are in Table 7.

Normal-only estimator on the 9-d counts	Micro-AUC	Macro-AUC	Beats any published?
per-scene kNN	60.6	58.8	no
per-scene isolation forest	60.4	59.6	no
per-scene histogram	55.8	55.9	no
per-scene GMM	55.1	58.5	no
per-scene Gaussian (X2 harness check)	54.1	57.7	no

H PER-VIDEO AGENT COVERAGE

The lowest-coverage test videos on NWPU-Campus. Coverage is the fraction of sampled frames in which the detector finds at least one agent. Videos below the pre-registered floor are marked; they handicap the counting control rather than favour it (Section 5).

Video	Agent coverage	
D121_01	0.28	below floor
D121_02	0.38	below floor
D122_06	0.53	below floor
D047_05	0.78	
D048_01	0.85	