Probabilistic Structured Event Modeling for Video Event Recognition under Perceptual Uncertainty

Nhan Thanh Ho[1⋆], Anh Dong Nguyen[2⋆], Thuan Duc Tran[3⋆,⋆⋆], Thang Van Nguyen[1], Bui Anh Thinh[1], Nguyen Bao Tran[4], and Long Nhat Ho[5]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology – VNU, Ho Chi Minh City, Vietnam
[2] School of Industrial Management, Ho Chi Minh City University of Technology – VNU, Ho Chi Minh City, Vietnam
[3] Department of Business Administration, Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam
[4] Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology – VNU, Ho Chi Minh City, Vietnam
[5] Faculty of Mechanical – Electrical and Computer Engineering, Van Lang University, Ho Chi Minh City, Vietnam

Abstract. Long-duration video understanding systems summarize continuous observations as structured event representations to enable semantic retrieval. However, these events are produced by imperfect perceptual pipelines (e.g., detectors, trackers, and action classifiers), meaning that every extracted attribute carries inherent uncertainty. Existing retrieval frameworks often ignore this uncertainty, relying on deterministic hard matching that fails when a single cue is misrecognized, or on embedding-based retrieval methods that do not explicitly preserve structured event attributes, particularly spatial information. This paper models structured events as probabilistic entities in which the correctness of each perceptual attribute is treated as a latent variable. A marginal-likelihood retrieval formulation integrates attribute-level uncertainty, allowing reliable attributes to compensate for uncertain ones instead of discarding an event after a single erroneous observation. For the spatial axis, detection covariance is propagated through the homography Jacobian to estimate uncertainty on the ground plane. Evaluated on an uncertainty-driven event retrieval benchmark using both the indoor CAVIAR and outdoor VIRAT datasets, the framework consistently outperforms deterministic hard matching under increasing attribute noise, reaching 3.3× higher Recall@5 (0.394 vs. 0.120) at the highest corruption level. Controlled ablations identify the per-event confidence as the operative signal precisely when perception is degraded, and delineate the boundary between robustness to attribute noise and failures caused by events missing from memory.

⋆ These authors contributed equally to this work.
⋆⋆ Corresponding author: thuantd@huit.edu.vn

Keywords: Video event retrieval · Probabilistic event modeling · Perceptual uncertainty · Confidence calibration · Structured event memory

1 Introduction

Long-duration surveillance and forensic video systems process continuous streams for event recognition, abnormal-event analysis, and semantic retrieval [2, 14, 23]. When retrieval-oriented systems organize observations into temporal, spatial, identity, and action fields, the resulting records enable interpretable compositional queries [14, 19]. This explicit representation is useful for long-video search, but its reliability is bounded by the perception modules that populate each field.

Existing structured retrieval pipelines typically treat observed event fields as deterministic symbols. In real video, detectors fail under occlusion, trackers switch identities [27], and action classifiers are uncertain; exact logical matching can therefore reject a relevant event after a single erroneous cue. Embedding-based retrieval methods such as CLIP [21] reduce dependence on explicit symbols, but they do not expose calibrated uncertainty for individual event axes and weaken the multi-axis structure needed for interpretable temporal, identity, action, and physical spatial reasoning. The missing capability is a retrieval model that preserves structured event attributes while accounting for their unequal reliability.

We propose a probabilistic structured event retrieval framework that represents each event attribute with calibrated reliability and ranks events by confidence-weighted evidence rather than hard acceptance or rejection. Attribute correctness is treated as latent, so unreliable observations contribute less while reliable temporal, location, actor, or action evidence can still support retrieval. As shown in Fig. 1, the system converts video observations into structured event memories with calibrated confidence scores and uses attribute-wise matching to let reliable axes dominate the final score. For spatial queries, the event memory carries ground-plane uncertainty instead of relying on a handcrafted location score, preserving physical interpretability while modeling localization ambiguity.

The contributions of this work are:

- A confidence-aware structured event memory that preserves explicit temporal, spatial, actor, and action fields while attaching calibrated reliability, including ground-plane spatial uncertainty.
- A probabilistic retrieval formulation that marginalizes latent attribute correctness and scores events by confidence-weighted evidence instead of deterministic hard matching, together with the linear surrogate it justifies.
- An uncertainty-driven structured retrieval benchmark and evaluation protocol for measuring robustness to perceptual degradation across event queries, separating corrupted attributes from events missing from memory.

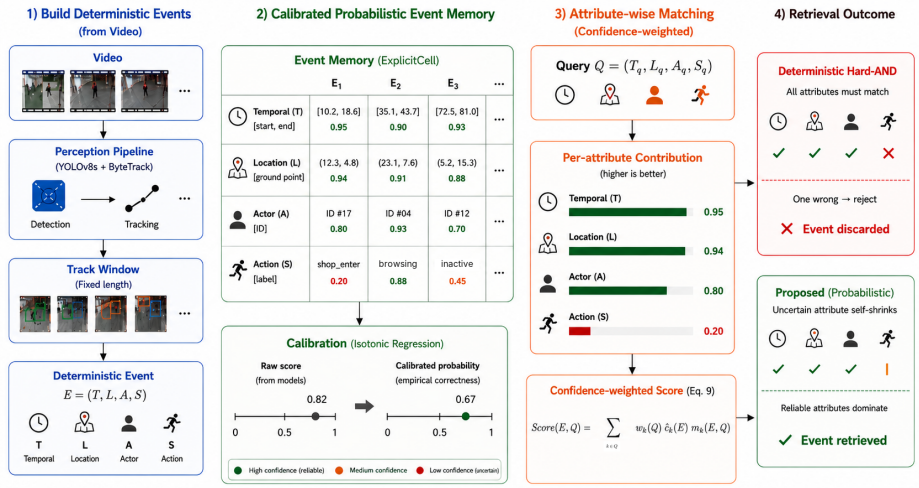


Fig. 1: Overview of the proposed confidence-aware event retrieval pipeline. Video observations are first converted into deterministic structured events $E = (T, L, A, S)$ through detection, tracking, and fixed-length track windows. The resulting temporal, location, actor, and action attributes are then stored in an ExplicitCell event memory together with calibrated confidence values obtained from isotonic regression. At query time, attribute-wise matching produces confidence-weighted contributions, allowing unreliable axes to self-shrink while reliable axes dominate the final score. Thus, unlike deterministic Hard-AND retrieval, which rejects an event after any single attribute mismatch, the probabilistic model can still retrieve the event when the aggregate calibrated evidence remains strong.

2 Related Work

2.1 Structured Event Modeling and Deterministic Retrieval

Surveillance video analysis processes continuous streams for event/activity recognition and abnormal-event detection [2, 23]. Retrieval-oriented systems further organize long videos into semantic, temporal, or spatial descriptors, enabling interpretable compositional queries and auditable long-video search [14, 19], and structured episodic memories such as STEM [3] recall events by conjunctive matching of stored attributes. This direction establishes the value of symbolic event memories, but retrieval commonly treats extracted fields as correct and applies deterministic predicates. It therefore lacks a mechanism to preserve explicit event structure while ranking candidates according to axis-wise perceptual reliability.

2.2 Recognition, Grounding, and Embedding-based Retrieval

Recognition backbones and temporal grounding models provide strong clip semantics and event boundaries [1,6,9,16,18,22,25,26]. Embedding-based retrieval, including CLIP-style models, further supports flexible visual matching in a continuous feature space [21]. These methods are effective for appearance and action

cues, but their representations do not expose calibrated reliability for temporal, location, actor, and action fields or encode physical ground-plane uncertainty for structured spatial queries.

2.3 Uncertainty-aware Perception and Probabilistic Retrieval

Uncertainty-aware detection, localization, noisy-label learning, and calibration estimate reliability within perception pipelines or map raw scores to better correctness probabilities [4, 10–12, 17, 24]. Probabilistic information retrieval similarly incorporates uncertainty into ranking functions or embedding spaces [5, 8]. These approaches provide useful reliability signals, but they typically stop at object-level decisions, score calibration, or embedding uncertainty rather than attaching calibrated perceptual reliability to explicit event attributes.

Overall, existing methods either (i) assume deterministic structured events, (ii) model uncertainty only at perception, or (iii) calibrate prediction confidence without structured event reasoning. These limitations motivate the framework developed in Section 3.

3 Method

The method follows directly from the limitations identified above. It preserves the explicit event axes valued by structured retrieval, adds the reliability information missing from deterministic event tuples, propagates spatial uncertainty into physical coordinates, and uses calibrated confidences to rank events without collapsing them into a single embedding.

3.1 Confidence-aware Structured Event Memory

The first design problem addresses the main limitation of deterministic structured retrieval: an event tuple is interpretable, but it treats every stored attribute as if it were correct. We therefore use a confidence-aware *ExplicitCell* : the event remains structured, while each axis carries the uncertainty needed by the retrieval model.

Figure 2 illustrates CIRCUIT, the multi-tier memory that stores ExplicitCells and indexes them along each axis, and the ExplicitCell representation used in this work.

As highlighted in Figure 2, raw perceptual outputs are grouped into unified *ExplicitCells* linked to parallel temporal, semantic, identity, and spatial indices. This organization attaches uncertainty to stored events while preserving the interpretable query interface.

With this structured-memory view, the traditional surveillance tuple $E = (t, l, a, s)$ is lifted into a probabilistic structured entity that preserves the same semantic axes but pairs each observed value with the reliability signal needed to reason about perceptual error:

$$E = \big\{(t, \hat{c}_t),\ (l, \Sigma_l),\ (a, \hat{c}_a),\ (s, \hat{c}_s)\big\}. \tag{1}$$

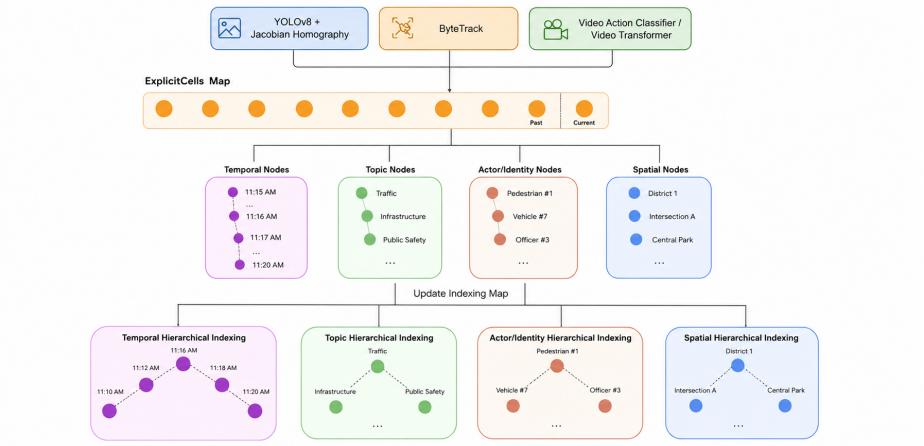


Fig. 2: CIRCUIT Memory Architecture and ExplicitCell Representation. The multi-tier topological framework decouples raw sensory components into unified event nodes (*ExplicitCells*), which project parallel indexing maps across semantic and physical dimensions.

Here, $\hat{c}_t$, $\hat{c}_a$, and $\hat{c}_s$ denote the calibrated probability of perceptual correctness for the temporal, identity, and action axes, while Σ_l is the spatial covariance matrix. Σ_l is not a scalar confidence: it induces the spatial likelihood of Section 3.2. This memory representation provides the interface between perception and probabilistic retrieval.

3.2 Geometric Uncertainty Propagation for Spatial Attributes

The location axis addresses a more specific gap: prior uncertainty-aware detectors estimate localization reliability in image coordinates, but spatial queries refer to ground-plane regions, and handcrafted spatial scores ignore that geometric transformation. We therefore propagate image-plane uncertainty through the camera homography so that location reliability reflects physical localization uncertainty.

Using local linear error propagation, a detection box has image covariance $\Sigma_{\text{image}} = \sigma_{\text{px}}(\text{conf})^2 I_2$, where $\sigma_{\text{px}}(\text{conf}) = \sigma_0 + \sigma_1(1-\text{conf})$ maps a low detection confidence to a looser foot-point localization; we use $\sigma_0 = 2\,\text{px}$ and $\sigma_1 = 10\,\text{px}$. It is propagated to the ground plane through the homography Jacobian J_H [13]:

$$\Sigma_{\text{ground}} = J_H \, \Sigma_{\text{image}} \, J_H^{\top}. \tag{2}$$

Spatial matching then becomes a probabilistic distance test through the Mahalanobis distance:

$$P(q_l \mid l) \propto \exp\left(-d_M(q_l, l)\right), \quad d_M = (x_q - x)^{\top} \Sigma_{\text{ground}}^{-1} (x_q - x). \tag{3}$$

The scalar location reliability used by the retrieval score is the probability that the true foot-point falls inside the queried ground-plane cell R, i.e. $\hat{c}_l = \int_R \mathcal{N}(x; x_l, \Sigma_{\text{ground}})\, dx$, which we evaluate with the axis-marginal approximation.

3.3 Attribute-level Probabilistic Observation Model

Given a confidence-aware memory, retrieval must address the hard-veto failure identified in prior structured systems: a single wrong attribute should not discard an otherwise relevant event. A fixed similarity score cannot distinguish true mismatch from low-confidence perception, and an embedding score does not preserve which axis caused the uncertainty. The proposed solution is to introduce a latent correctness variable $Z_k \in \{0, 1\}$, where $k \in \{t, l, a, s\}$ denotes the semantic axes (time, location, actor, action).

The intuition is that each observed attribute has two possible explanations: if perception is correct, the query should match the observed value; if perception is unreliable, the model should fall back to a background likelihood instead of vetoing the event. This yields the two-component mixture:

$$P(q_k \mid e_k) = \hat{c}_k \, s_k(q_k, e_k) + (1 - \hat{c}_k) \, b_k(q_k), \quad k \in \{t, l, a, s\}. \tag{4}$$

where:

- $\hat{c}_k$: the calibrated probability that perception is correct on axis k, i.e. $P(Z_k = 1)$.
- $s_k(q_k, e_k) = P(q_k \mid e_k, Z_k = 1)$: the similarity likelihood when perception is correct. For the categorical axes it is the indicator $\mathbf{1}[q_k = e_k]$; for the location axis (l) it is the multivariate normal distribution induced by the Mahalanobis distance:

$$P(q_l \mid l, Z_l = 1) = \frac{1}{2\pi\sqrt{|\Sigma_{\text{ground}}|}} \exp\left(-\frac{1}{2} d_M(q_l, l)\right). \tag{5}$$

- $b_k(q_k)$: the background noise likelihood, which we assume uniform over the observed value domain of axis k, i.e. $b_k = 1/|\mathcal{V}_k|$ with $\mathcal{V}_k$ the set of distinct perceived values on that axis, for simplicity and computational feasibility.

If perception is accurate ($\hat{c}_k \to 1$) the similarity component s_k determines the result; if it is uncertain ($\hat{c}_k \to 0$) the background b_k dominates, so uncertainty changes the strength of evidence rather than removing the event from consideration.

3.4 Event-level Retrieval and Tractable Linear Approximation

The axis-level model is combined into one score for multi-attribute queries that rewards agreement on reliable axes without letting a single unreliable axis veto a relevant event. Assuming independent observations across axes, the marginal likelihood that the full event E matches a multi-axis query Q is the product over axes. Taking the logarithm yields the core retrieval formulation (maximum marginal likelihood):

$$\log P(Q \mid E) = \sum_k \log\left(\hat{c}_k \, s_k(q_k, e_k) + (1 - \hat{c}_k) \, b_k(q_k)\right). \tag{6}$$

This formulation mathematically explains graceful degradation: if perception is highly confident ($\hat{c}_k = 1$) but the query value does not match ($s_k = 0$), the log function applies a large penalty ($-\infty$), similar to Hard-AND. Conversely, if perception is uncertain ($\hat{c}_k \to 0$), the event is only penalized by a finite constant $\log b_k(q_k)$, preventing it from being removed from the candidate set.

The log-likelihood gives the desired probabilistic behavior, but direct inference requires estimating $b_k(q_k)$ for every attribute class and evaluating a non-linear mixture at retrieval time. It also justifies a practical linear rule. Consider the single-axis component function $f(\hat{c}_k) = \log\big(\hat{c}_k s_k + (1 - \hat{c}_k) b_k\big) = \log\big(b_k + \hat{c}_k(s_k - b_k)\big)$, where $s_k = s_k(q_k, e_k)$ and $b_k = b_k(q_k)$. Performing a first-order Taylor expansion of $f(\hat{c}_k)$ around the absolute background-noise point ($\hat{c}_k = 0$) yields:

$$f(\hat{c}_k) \approx \log b_k(q_k) + \hat{c}_k \, \frac{s_k(q_k, e_k) - b_k(q_k)}{b_k(q_k)}. \tag{7}$$

The expansion separates a query-dependent background term from an event-dependent reliability term; summed over axes, $\sum_k \log b_k(q_k)$ is constant across candidates and can be dropped without changing the ranking. By introducing query-side axis weights $w_k(Q)$ and explicitly defining the normalized similarity metric as:

$$m_k(E, Q) = \frac{s_k(q_k, e_k) - b_k(q_k)}{b_k(q_k)}, \tag{8}$$

the strict probabilistic inference reduces to an efficient linear scoring function:

$$\text{Score}(E, Q) = \sum_{k \in Q} w_k(Q) \, \hat{c}_k(E) \, m_k(E, Q), \tag{9}$$

where $\hat{c}_k(E)$ represents the empirically calibrated perceptual correctness probability of axis k. A noisy axis self-shrinks its contribution while reliable axes preserve the rank of the correct event. Here $w_k(Q)$ expresses how much the *query* trusts its own cue on axis k, as opposed to $\hat{c}_k(E)$, which expresses how much the *event* trusts its stored value; unless stated otherwise we set $w_k(Q) = 1$ for every queried axis, and Section 5.4 is the only experiment in which $w_k(Q) < 1$. We report Eq. (6) as the proposed model and Eq. (9) as its tractable surrogate.

3.5 Confidence Calibration

The formulation is probabilistic only if the reliability values imported from perception have the same meaning across axes. Raw perceptual confidences are often miscalibrated [10], so inserting them directly into the mixture would make axis weights incomparable; the final design step is therefore to calibrate them before retrieval.

Isotonic regression [24] is used to map a raw confidence score r_k into a calibrated probability $\hat{c}_k$. It is fitted by two-fold cross-fitting, so no event is calibrated with a model that saw its own label. This calibration synchronizes heterogeneous

perceptual signals into a shared probabilistic reference frame, making it possible to embed them into the latent-variable model and assign meaningful weights during retrieval.

4 Uncertainty-Driven Event Retrieval Benchmark

4.1 Benchmark Motivation

Conventional structured event retrieval evaluation typically assumes fixed, clean event records and reports aggregate retrieval accuracy. This setting is insufficient for uncertainty-aware retrieval: a failure may arise from an incorrectly recorded attribute, an event missing from memory, or an ambiguous user query, yet these cases require different behavior from the retrieval model. We therefore ask whether calibrated confidence can help retrieval remain reliable as perceptual evidence degrades, and construct a controlled-corruption benchmark in which ground-truth labels define relevance only, while ranking is performed solely over noisy perceptual attributes.

4.2 Evaluation Protocol and Degradation Mechanisms

The benchmark controls perceptual difficulty with severity $s \in \{0, 1, 2, 3\}$ and axis-corruption probability $p(s) \in \{0, 0.15, 0.35, 0.60\}$. It separates three uncertainty regimes:

- Attribute Noise (main mode): one perceptual axis is replaced by a random incorrect value, and its latent confidence $\hat{c}_k$ is scaled by $\gamma = 0.2$, simulating misclassification and ID swaps.
- Event Omission (miss mode): target events are removed from memory, simulating severe occlusion or tracker loss; an omitted relevant event counts as a miss for every method.
- Query Corruption (query-noise mode): a queried cue is replaced by a wrong value and the query reports a lower weight $w_k(Q)$ on that cue, reducing the corrupted axis in the final score.

A query asks for events sharing a seed event's values on a set of axes; ties are broken at random (three draws).

4.3 Performance Metrics and Scientific Scope

Retrieval performance is measured by Recall@5, Recall@10, and mean reciprocal rank (MRR); calibration quality by expected calibration error (ECE, ten bins) together with the AUROC of confidence against correctness, since ECE alone cannot distinguish a well-scaled constant from a per-event discriminator. Robustness curves summarize how each method degrades across the severity spectrum. The benchmark is a controlled overlay on off-the-shelf perceptual models rather than a new physical dataset.

Method	Action	Location	Actor	L+A	S+L+A
Oracle (ground-truth attributes)	.635	.847	.802	.920	.931
Proposed (exact, Eq. (6))	.184	.640	.554	.679	.675
Confidence heuristic	.185	.661	.558	.669	.644
Equal-weight ($\hat{c}_k \equiv 1$)	.286	.624	.559	.652	.643
Hard-AND (STEM-style)	.172	.659	.542	.610	.558
CLIP-NN (appearance)	.262	.310	.310	.363	.364
# queries	65	149	164	219	213

Table 1: Recall@5 on clean CAVIAR (s_0), averaged over the query sets counted in the last row. Ground truth defines relevance only; every method ranks perceived attributes. CLIP-NN is an appearance-embedding nearest-neighbor baseline over the same events and queries.

5 Experiments & Results

Implementation Details. Experiments use CAVIAR [7] for indoor retrieval and VIRAT Ground [20] for outdoor transfer, producing 1,428 and 5,262 ExplicitCells, respectively. Person detections are obtained with YOLOv8s [15] (person class, score threshold 0.10, input resolution 960 px) and linked by ByteTrack [27]. Perception runs on an NVIDIA T4 GPU; all remaining stages run on an AMD Ryzen 7 7735HS CPU. Tracks are split into non-overlapping 75-frame windows (3 s at 25 fps), discarding windows with fewer than 20 detections; each retained track-window pair forms $E = (T, L, A, S)$. Confidence estimates follow Section 3.2 and Section 3.5: temporal completeness, homography-based location uncertainty, actor track reliability, and out-of-fold probabilities of a gradient-boosting action classifier over trajectory features. Location and action confidences are calibrated with cross-fitted isotonic regression [24]. Queries are formed over the action, location and actor axes. Attribute corruption uses $s \in \{0, 1, 2, 3\}$, $p(s) \in \{0, 0.15, 0.35, 0.60\}$, and $\gamma = 0.2$, averaged over five random seeds.[1] The CLIP baseline uses ViT-B/32 [21] on cropped detections and cosine similarity.

5.1 Clean-data Retrieval and CLIP Baseline (CAVIAR)

We first ask how the structured probabilistic memory behaves when no synthetic corruption is applied. Table 1 compares Recall@5 on clean CAVIAR data (noise level $s = 0$) across single-axis and multi-axis queries.

Table 1 and Fig. 5a highlight a complementary behavior between appearance embeddings and structured event memory. CLIP-NN remains stronger on isolated action queries, where visual appearance captures fine behavioral cues, whereas the proposed model is best on multi-axis queries and roughly doubles CLIP-NN on location-aware ones, suggesting that explicit physical structure matters when retrieval depends on true spatial locations. The Oracle gap is an

[1] All results were regenerated from one event snapshot after fixing ECE binning, in-sample calibration and index-based tie-breaking; numbers therefore differ slightly from the review version, and every conclusion is unchanged.

Method	s_0 (0.00)	s_1 (0.15)	s_2 (0.35)	s_3 (0.60)
Oracle (GT attributes)	.931 / 1.000	.931 / 1.000	.931 / 1.000	.931 / 1.000
Proposed (exact, Eq. (6))	.675 / .716	.637 / .679	.559 / .622	.394 / .462
Taylor surrogate (Eq. (9))	.687 / .711	.631 / .661	.558 / .613	.396 / .474
Confidence heuristic	.645 / .698	.522 / .597	.429 / .516	.274 / .343
Conf. heuristic, uncalibrated	.616 / .634	.510 / .559	.425 / .487	.245 / .336
Equal-weight ($\hat{c}_k \equiv 1$)	.640 / .701	.512 / .576	.396 / .484	.223 / .282
Hard-AND (STEM-style)	.558 / .652	.382 / .517	.249 / .412	.120 / .202

Table 2: Retrieval under attribute noise (CAVIAR, S+L+A queries, Recall@5 / MRR; five corruption seeds, three tie-break draws). Severity s corrupts each queried attribute with probability $p(s)$ and scales that axis' confidence by $\gamma = 0.2$; all rows share one snapshot and one calibration fit.

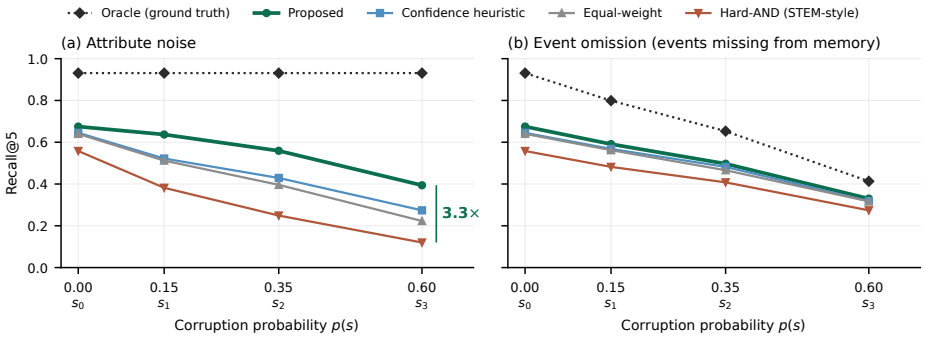


Fig. 3: Degradation curves (CAVIAR, S+L+A, Recall@5) under (a) attribute noise and (b) the boundary case in which events are removed from memory.

end-to-end perception gap: Oracle uses ground-truth labels, whereas every other method ranks perceived events.

5.2 Attribute-noise Degradation

We next test whether the retrieval rule remains robust when event attributes are corrupted. Table 2 summarizes S+L+A retrieval under attribute noise, and Fig. 3a visualizes the degradation curves as corruption severity increases. The probabilistic model retains a clear advantage at the hardest level (s_3): 0.394 Recall@5 versus 0.120 for Hard-AND (3.3×), showing that soft marginalization turns unreliable axes into weaker evidence instead of brittle vetoes. The confidence heuristic $\sum_k \hat{c}_k s_k$, which uses the same confidences without probabilistic marginalization, reaches 0.274, so marginalization adds 0.120 beyond confidence weighting alone; the Taylor surrogate of Eq. (9) stays within 0.002 of exact inference. An empirical instead of uniform b_k leaves the trend unchanged (0.677 to 0.407 across s_0–s_3).

5.3 Confidence Calibration on Real Data

Because the scoring rule treats confidences as probabilities, we next verify whether raw perceptual scores are sufficiently calibrated. Fig. 4 visualizes the

Axis	Acc.	$\mathrm{ECE}_{\mathrm{raw}}$	$\mathrm{ECE}_{\mathrm{cal}}$	AUROC
Action	0.862	0.113	0.018	0.811
Location	0.920	0.552	0.000	0.507

Table 3: ECE before/after cross-fitted isotonic regression and AUROC (n = 1,218 labeled events per axis).

Severity	q_{flat}	q_{conf}	Gain
s_0 (0.00)	0.675	0.675	+0.000
s_1 (0.15)	0.521	0.577	+0.056
s_2 (0.35)	0.402	0.493	+0.091
s_3 (0.60)	0.282	0.376	+0.094

Table 4: Retrieval under query corruption (S+L+A, Recall@5): flat vs. confidence-aware query.

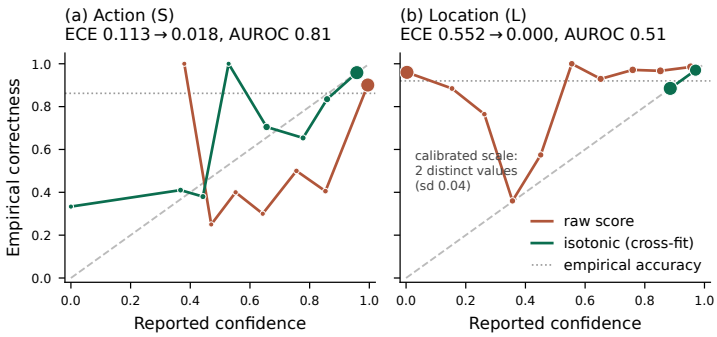


Fig. 4: Reliability diagram: calibration turns raw scores into probabilities for marginal-likelihood inference (marker area $\propto$ bin count; dashed: perfect calibration).

reliability diagram before and after isotonic regression, while Table 3 reports ECE for the two axes with directly measurable correctness: action (S) and location (L).

ECE drops from 0.552 to 0.000 for location and from 0.113 to 0.018 for action. This supports using calibrated $\hat{c}_k$ as $P(\text{axis } k \text{ correct})$ in the marginal-likelihood score rather than as an uncalibrated heuristic weight. The AUROC shows that the two axes differ: the action confidence ranks correct above incorrect predictions (0.811), whereas the location confidence (0.507) acts as a well-scaled axis-level reliability.

5.4 Retrieval under Query Corruption

The previous experiments corrupt stored event attributes; we now ask whether confidence also helps when uncertainty appears on the query side. Table 4 compares a flat query q_{flat}, which weights all axes equally, with a confidence-aware query q_{conf}, which down-weights less reliable axes. Query-side confidence provides an additional protection mechanism: as severity increases, q_{conf} increasingly outperforms q_{flat}, reaching a +0.094 Recall@5 gain at s_3. On clean data the two queries coincide, indicating that confidence changes the ranking primarily when the query contains unreliable evidence.

Method	s_0 (0.00)	s_1 (0.15)	s_2 (0.35)	s_3 (0.60)
Oracle (GT attributes)	.701 / 1.000	.701 / 1.000	.701 / 1.000	.701 / 1.000
Proposed (exact, Eq. (6))	.508 / .796	.439 / .772	.339 / .681	.221 / .518
Taylor surrogate (Eq. (9))	.414 / .719	.360 / .652	.285 / .577	.200 / .476
Confidence heuristic	.488 / .801	.385 / .739	.262 / .630	.153 / .466
Equal-weight ($\hat{c}_k \equiv 1$)	.492 / .804	.373 / .700	.237 / .557	.098 / .360
Hard-AND (STEM-style)	.473 / .775	.358 / .701	.220 / .570	.085 / .362

Table 5: Cross-domain transfer to VIRAT Ground (5 scenes, 5,262 ExplicitCells, L+A queries, Recall@5 / MRR). The isotonic maps fitted on CAVIAR are applied unchanged; refitting them on VIRAT changes Recall@5 by at most .023.

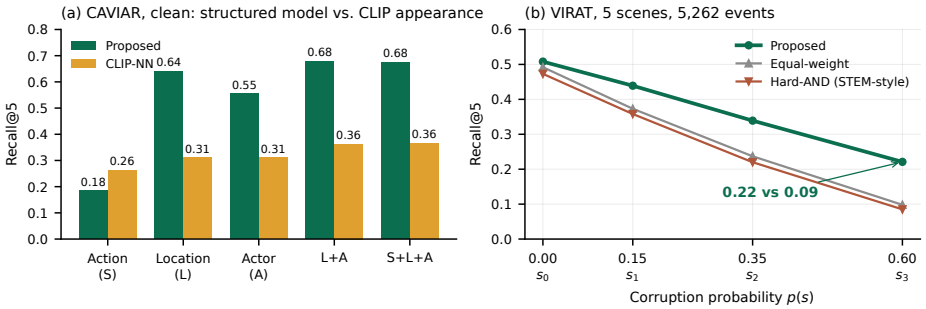


Fig. 5: CLIP + VIRAT: comparison between a black-box embedding baseline and the structured model (a), and cross-domain transfer to VIRAT (b).

5.5 Cross-domain Generalization on VIRAT

We then test whether the same retrieval rule transfers beyond the indoor CAVIAR setting. Fig. 5b and Table 5 evaluate L+A retrieval on outdoor VIRAT scenes without domain-specific recalibration. The trend mirrors the indoor results: the probabilistic model remains consistently above Hard-AND as noise increases and reaches 0.221 Recall@5 versus 0.085 at s_3, suggesting that calibrated structured retrieval transfers across scene type; refitting the isotonic maps on VIRAT moves Recall@5 by at most 0.023. The Taylor surrogate is less reliable here (0.414 vs. 0.508 on clean data) because a small b_A amplifies $(s_k - b_k)/b_k$.

5.6 Real Resolution Degradation

Finally, we distinguish attribute uncertainty from event loss. Unlike synthetic attribute noise, resolution degradation can remove detections entirely, preventing some events from entering memory; Table 6 evaluates this boundary case on CAVIAR. Because the pipeline emits a different number of events at each resolution, scores are comparable only within a row, where the ordering of methods is preserved. The controlled omission experiment (Fig. 3b) isolates recall loss with the candidate pool fixed: the Oracle ceiling itself falls from 0.931 to 0.413 and all methods converge (0.330 versus 0.317 for Equal-weight at s_3). The framework is designed to reason over uncertain observed attributes, not to recover events removed from memory by missed detections (recall loss); resolution degradation

Resolution	Events	Proposed	Conf. heuristic	Equal-weight	Hard-AND
1/1 (original)	1,428	.675 / .716	.644 / .696	.643 / .708	.558 / .652
Down 1/2	1,721	.690 / .743	.678 / .731	.676 / .725	.618 / .691
Down 1/4	739	.749 / .756	.744 / .735	.724 / .719	.567 / .672

Table 6: Performance under real resolution degradation (CAVIAR, S+L+A, Recall@5 / MRR). The pipeline emits a different number of events at each level, so only the ordering of methods within a row is comparable.

Variant	Replacement for $\hat{c}_k(E)$	R@5 s_0	R@5 s_3	MRR s_3
Oracle	Ground truth (upper bound)	.931	.931	1.000
Proposed	Calibrated per-event $\hat{c}_k(E)$	.675	.394	.462
Uncalibrated	Raw perceptual score	.633	.367	.441
Per-axis mean	Axis average, no per-event information	.673	.265	.335
Shuffled	Shuffle $\hat{c}_k$ across events	.654	.227	.301
Random	Uniform $\mathcal{U}[0,1]$	.634	.240	.309
Inverted	$1 - \hat{c}_k$	.590	.139	.188
Equal-weight	$\hat{c}_k = 1$ (deterministic)	.640	.223	.282
Hard-AND	Hard matching (STEM-style)	.558	.120	.202

Table 7: Controlled latent-variable ablation (CAVIAR, S+L+A): $\hat{c}_k(E)$ is replaced while the exact scoring rule of Eq. (6) and the queries are held fixed.

therefore marks the transition from noisy observations, where probabilistic retrieval helps, to missing observations, where no retrieval rule can rank absent events.

6 Ablation Studies

The ablation study asks whether the gains come from meaningful, event-specific uncertainty estimates rather than from arbitrary soft weighting, keeping the inference function of Eq. (6) fixed.

6.1 Latent-variable Controls

Table 7 shows that confidence becomes informative precisely when perception is unreliable. The per-axis mean, which keeps each axis' average reliability but removes all per-event information, matches the full model on clean data (0.673 versus 0.675) yet loses 0.129 under heavy noise. Likewise, shuffled confidence costs 0.021 at s_0 but 0.167 at s_3, and inverted confidence falls to 0.139, barely above Hard-AND. Thus, the benefit is not merely soft weighting but depends on preserving event-specific confidence. A separate γ sweep at s_3 gives the same conclusion: for $\gamma \in \{0.1, 0.2, 0.4, 0.6, 1.0\}$, the proposed model remains above Hard-AND with R@5 ratios of $3.51\times$, $3.28\times$, $2.84\times$, $2.58\times$, $1.55\times$, so robustness persists even without confidence–corruption coupling ($\gamma = 1$).

Perturbed axis	**Randomized**, s_0	**Decrease**	**Per-axis mean**, s_3	**Decrease**
None (full model)	.675	—	.394	—
Location $\hat{c}_L$	.644	−0.031	.328	−0.066
Action $\hat{c}_S$	.655	−0.020	.377	−0.017
Actor $\hat{c}_A$	.659	−0.016	.341	−0.053

Table 8: Per-axis attribution (CAVIAR, S+L+A, Recall@5): one axis' calibrated confidence is randomized on clean data or replaced by its per-axis mean under heavy noise.

6.2 Per-axis Attribution

Table 8 identifies location confidence as the strongest single contributor: randomizing $\hat{c}_L$ causes a 0.031 Recall@5 decrease on clean data and replacing it by its mean costs 0.066 under heavy noise, larger than the action or actor axes. This supports the role of homography-based spatial uncertainty as a useful reliability signal rather than a cosmetic score attached to the location attribute.

6.3 Calibration as a Reliability Mechanism

Finally, we test whether the calibrated probability scale itself is necessary. Replacing calibrated by uncalibrated confidences lowers Recall@5 at every severity: by 0.029 at both s_0 and s_3 for the confidence heuristic (Table 2) and by 0.027 at s_3 for the exact rule (Table 7). Calibration thus provides a consistent reliability scale, while the robustness itself comes from marginalization and event-specific confidence.

7 Limitations

Our framework does not recover events absent from memory: missed detections remove events, and track fragmentation or identity switches split them, which our experiments capture only through the actor-axis confidence. Eq. (6) assumes conditional independence, whereas real perceptual errors are often correlated across axes. Confidence-distribution shift between domains is not modeled. Finally, CAVIAR ships no camera calibration, so the homography is built from four floor points per camera and σ_{px} is a heuristic rather than a measured localization error.

8 Conclusion

This work presents a probabilistic structured event memory for retrieval under perceptual uncertainty. By storing calibrated axis-wise confidence and marginalizing latent attribute correctness, the framework degrades gracefully under noisy observations and improves over deterministic Hard-AND matching by 3.3× Recall@5 under high corruption. The experiments show that event-specific confidence, calibration, and spatial uncertainty provide useful reliability signals, while resolution degradation marks the boundary between noisy attributes and missing events.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: International Conference on Machine Learning (ICML) (2021)
2. Bhaumik, H., Bhattacharyya, S., Chakrabarti, S.: Event and activity recognition in video surveillance for cyber-physical systems. In: Emergence of Cyber Physical System and IoT in Smart Automation and Robotics, pp. 51–68. Springer (2021)
3. Chang, P.H., Tan, A.H.: Encoding and recall of spatio-temporal episodic memory in real time. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 1490–1496 (2017). https://doi.org/10.24963/ijcai.2017/206
4. Choi, J., Chun, D., Kim, H., Lee, H.J.: Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
5. Cohen, D., Mitra, B., Lesota, O., Rekabsaz, N., Eickhoff, C.: Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In: ACM SIGIR Conference on Research and Development in Information Retrieval (2021)
6. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
7. Fisher, R.B.: The CAVIAR video datasets. EC-funded CAVIAR project, IST 2001 37540 (2004)
8. Fuhr, N.: Probabilistic models in information retrieval. The Computer Journal **35**(3), 243–255 (1992)
9. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: Temporal activity localization via language query. In: IEEE International Conference on Computer Vision (ICCV) (2017)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML) (2017)
11. Hall, D., Dayoub, F., Skinner, J., Zhang, H., Miller, D., Corke, P., Carneiro, G., Angelova, A., Sünderhauf, N.: Probabilistic object detection: Definition and evaluation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)
12. Han, H., Zheng, Q., Luo, M., Miao, K., Tian, F., Chen, Y.: Noise-tolerant learning for audio-visual action recognition. IEEE Transactions on Multimedia **26**, 7761–7774 (2024)
13. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2nd edn. (2003)
14. Hu, W., Xie, D., Fu, Z., Zeng, W., Maybank, S.: Semantic-based surveillance video retrieval. IEEE Transactions on Image Processing **16**(4), 1168–1181 (2007)
15. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (YOLOv8). Software release, https://github.com/ultralytics/ultralytics (2023)
16. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
17. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
18. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video Swin Transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

19. Lyu, Z., Zhang, Y.: A novel temporal moment retrieval model for apron surveillance video. Computers & Electrical Engineering **107**, 108616 (2023)
20. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J.K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3153–3160 (2011)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
22. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
23. Wan, S., Xu, X., Wang, T., Gu, Z.: An intelligent video analysis method for abnormal event detection in intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems **22**(7), 4487–4495 (2021)
24. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 694–699 (2002)
25. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
26. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2D temporal adjacent networks for moment localization with natural language. In: AAAI Conference on Artificial Intelligence (2020)
27. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision (ECCV) (2022)