U^2-REVISIT: PERSISTENT EVENT REASONING UNDER INTERMITTENT AERIAL REVISITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Route-based UAV inspection efficiently covers spatially distributed regions but observes each location only intermittently, with irregular gaps during which the region state remains unobserved. Verifying persistent events therefore requires reasoning across revisits while distinguishing confirmed absence from missing or unreliable evidence. A region-centered benchmark and a coarse-to-fine pipeline are introduced for this setting. Fixed semantic map regions provide spatial correspondence across revisits, and a lightweight proposal stage identifies candidate sequences for a learned multi-revisit verifier. U^2-Revisit combines visual, telemetry, map, and observation-quality information through quality-aware fusion. Separate spatial and instance memories represent region state and vehicle-related evidence. Under a seven-fold, site-ID-disjoint LOSO protocol, U^2-Revisit achieves the highest mean Event AP of 0.9594 ± 0.0082 across three seeds. Single Revisit and Unordered Set achieve 0.8840 ± 0.0380 and 0.9462 ± 0.0270, respectively. Repeated-observation aggregation accounts for most of the Event AP improvement over Single Revisit, and U^2-Revisit further improves Event AP beyond Unordered Set.

1 INTRODUCTION

Repeated UAV inspections provide an efficient way to monitor parking spaces, road segments, and other geographically anchored regions. Unlike continuous video, the same physical region is observed only during brief revisits separated by unobserved intervals. Verifying whether an event persists therefore requires evidence from multiple observations of the same location. This setting differs from continuous video tracking, which relies on densely sampled frames and temporal continuity. Irregular revisit intervals, imperfect image-to-map registration, limited spatial coverage, missed detections, and occlusion further complicate the task.

A lightweight proposal stage uses inexpensive cues such as vehicle detection frequency, motion stability within a revisit, and spatial consistency across revisits to identify candidate sequences. Learned verification then aggregates evidence across revisits for persistent-event reasoning.

A key challenge is to distinguish missing evidence from evidence of absence. A missing vehicle detection may result from incomplete coverage, occlusion, detector failure, or registration failure. Observations are therefore organized around fixed semantic map regions, with confirmed absence distinguished from missing or unreliable evidence. Reliability and modality-availability cues are retained and passed to the model rather than being resolved during preprocessing.

We introduce U^2-Revisit, a model for persistent event verification across intermittent UAV revisits. It combines visual, telemetry, and map information through quality-aware fusion, explicitly models the elapsed time between observations, and maintains separate spatial and instance memories for region state and vehicle-related evidence. These components allow the model to aggregate evidence across repeated observations with different acquisition conditions and revisit intervals.

The main contributions are:

- Persistent event verification under intermittent aerial revisits is formulated using fixed semantic map regions as stable units for reasoning across repeated observations. The benchmark distinguishes confirmed absence from missing or unreliable evidence.

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.

- An aerial revisit benchmark is constructed with visual, telemetry, map, and event annotations. Generalization to unseen sites is evaluated using a seven-fold leave-one-site-out (LOSO) protocol with disjoint site IDs.
- U^2-Revisit combines quality-aware multimodal fusion with separate spatial and instance memories. On the frozen LOSO benchmark, U^2-Revisit achieves a mean Event AP of 0.9594, compared with 0.8840 for Single Revisit and 0.9462 for Unordered Set.

2 Related Work

Driving and aerial benchmarks. Recent driving and aerial benchmarks cover multimodal perception, repeated traversal, vehicle tracking, and generalization across domains. MAVREC (Dutta et al., 2024) combines aerial and ground observations, VETRA (Hellekes et al., 2024) studies tracking from moving aerial platforms, and Open MARS (Li et al., 2024b) provides asynchronous data from repeated driving traversals. M3OT (Nie et al., 2025) provides a multimodal benchmark for tracking across multiple drones. DynUAV (Ye et al., 2026) evaluates tracking under complex motion and adverse observation conditions. MUSES (Brödermann et al., 2024) studies perception under multimodal uncertainty. Other datasets address georeferenced aerial traffic analysis (Fonod et al., 2025) and generalization in scene change detection (Kim & Kim, 2025). These benchmarks use frames, trajectories, traversals, or image pairs as their primary evaluation units. Our benchmark uses fixed semantic regions and models their occupancy across intermittent UAV revisits.

Tracking over long time spans and vehicle re-identification. Vehicle re-identification (ReID) and UAV tracking over long time spans mainly address identity preservation under appearance changes, viewpoint shifts, occlusion, and temporal gaps. Day–Night Cross-domain Vehicle Re-identification (Li et al., 2024a) addresses severe illumination changes, and the Aerial-Ground Cross-View Vehicle Re-ID benchmark (Shang et al., 2025) evaluates association across views. UCM-VeID V2 (Liu et al., 2025) studies vehicle ReID across UAV sensing modalities. Sea You Later (Yang et al., 2024) uses UAV metadata to recover identity after long temporal gaps, and UAST (Qin et al., 2026) studies active search after target loss. Our task separates region occupancy from vehicle identity because a region may remain occupied after the vehicle changes.

Irregular temporal modeling and censored events. Continuous-time models such as Rough Transformer (Moreno-Pino et al., 2024), ACSSM (Park et al., 2025), and Physiome-ODE (Klötergens et al., 2025) address observations sampled at irregular times, mainly through forecasting, interpolation, or latent dynamics modeling. Other work examines stable continuous time dynamics and evaluation under varying observation frequencies and missing values (Kuleshov et al., 2026; Spinnato & Landi, 2026). Change3D (Zhu et al., 2025) considers temporal change through video modeling. Our task focuses on persistent spatial states observed through sparse aerial revisits. States between revisits are unobserved, and observation reliability varies across visits. Exact event times are often unavailable, so we follow formulations for interval censoring (Shahin et al., 2024; García Meixide et al., 2024) and represent event boundaries as intervals between consecutive observations rather than single timestamps.

Registration between aerial images and maps. UAV localization and geospatial models combine reference maps with complementary sensing modalities to construct spatially consistent representations (Kim et al., 2024; Han et al., 2024). AirLock+ (Deng et al., 2026) studies registration between UAV and satellite imagery under substantial differences in viewpoint and domain. RobSense (Do et al., 2025) considers multimodal remote sensing representations with incomplete and temporally irregular inputs. DiffRegCD (Madani et al., 2026) jointly addresses registration and temporal change under viewpoint variation and misalignment. In our pipeline, registration between each aerial observation and the reference map provides the spatial correspondence needed to associate repeated observations with the same semantic region. Registration quality and modality availability are passed directly to the model so that unreliable observations can be masked or assigned lower weight.

Table 1: Comparison with representative driving and aerial benchmarks in terms of aerial acquisition, repeated revisits, irregular time gaps, fixed semantic regions, and event supervision.

Dataset	Aerial	Revisit	Irreg. gap	Fixed region	Event
MAVREC (Dutta et al., 2024)	✓	–	–	–	–
VETRA (Hellekes et al., 2024)	✓	–	–	–	–
Open MARS (Li et al., 2024b)	–	✓	✓	–	–
M3OT (Nie et al., 2025)	✓	–	–	–	–
DynUAV (Ye et al., 2026)	✓	–	–	–	–
Ours	✓	✓	✓	✓	✓

3 Intermittent Aerial Revisit Benchmark

Table 1 compares our benchmark with representative driving and aerial datasets in terms of aerial acquisition, repeated revisits, irregular time gaps, fixed semantic regions, and event supervision.

3.1 Data Collection and Flight Protocol

We collected UAV videos of road scenes at 11 sites using a DJI Mini 3 Pro. The data include 4K video, synchronized GPS/SRT telemetry, flight trajectories, georeferenced reference maps, fixed semantic regions, and vehicle annotations. In total, 87 videos were recorded.

The UAV flew at approximately 48 m with the camera oriented near nadir and recorded video at 30 fps. Median horizontal speed varied from 4.4 to 7.5 m/s across sites. At each site, the UAV repeatedly followed a predefined closed loop or bidirectional route, producing repeated observations of the same areas at naturally varying time intervals.

The 11 sites serve different roles in the benchmark. Seven sites, R1, R2, R3, R4, R6, R7, and R8, form the frozen LOSO benchmark used for the main evaluation. R10 and R11 are reserved for additional external-site evaluation and are excluded from model training and selection. R5 is used only for diagnostic analysis because it was involved during method development. R9 was finalized after the primary benchmark protocol was frozen and is retained as a benchmark extension.

3.2 Spatial Organization and Revisit Construction

Each site is divided into fixed semantic regions, mainly parking and roadway polygons. An observation of a region during one UAV pass forms one revisit. Observations of the same region from different passes are ordered by time to form a revisit sequence.

Reference maps are constructed using photogrammetric reconstruction or 2D mosaics anchored by GPS, depending on the scene geometry. UAV telemetry provides an initial estimate of the frame location, and visual registration then aligns the frame with the reference map. The registered image footprint determines which regions are sufficiently observed. Multiple valid views of the same region within one pass are combined into a single observation.

Regions with insufficient visual support are marked as unavailable rather than empty. This preserves the distinction between missing evidence and confirmed absence. Details of image-to-map registration and the construction of region observations are provided in Appendices B.2 and B.3.

3.3 Observation and Event Annotations

Each region observation is assigned one of six evidence states:

$$\mathcal{C} = \left\{ \begin{array}{lll} \texttt{confirmed_empty}, & \texttt{vehicle_detected}, & \texttt{detection_missing}, \\ \texttt{region_not_observed}, & \texttt{occluded}, & \texttt{registration_failed} \end{array} \right\}.$$

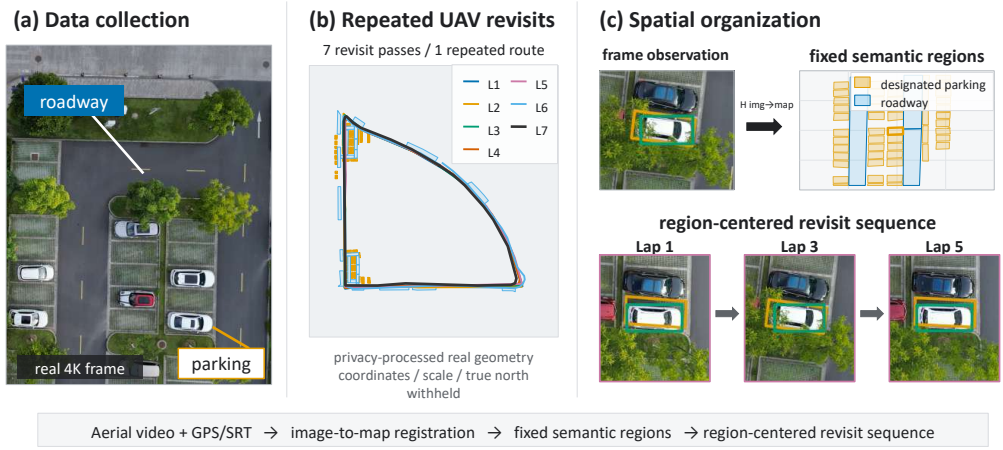


Figure 1: Data collection and spatial organization at one representative site. A real UAV frame from a near-nadir view, repeated flight passes, and registration to the reference map illustrate how observations are associated with fixed semantic regions to form revisit sequences.

The first two states provide direct evidence about region occupancy. The remaining states indicate that the observation is missing or unreliable. A missing vehicle detection is not treated as evidence that the region is empty.

Event annotation. A lightweight proposal stage first identifies region sequences that may contain persistent occupancy. The proposal rules use repeated vehicle observations, stationary evidence within a revisit, and spatial consistency across revisits. Unavailable observations provide no negative evidence and do not terminate a sequence by themselves.

For roadway regions, persistent occupancy is treated as a candidate abnormal parking event. For designated parking regions, it indicates continued parking occupancy. Proposed events are manually reviewed before labels are assigned. The frozen benchmark also includes additional region sequences with valid negative persistence labels. Details of proposal generation are provided in Appendix A.4, and the final benchmark population is summarized in Appendix A.5.

Identity across revisits and temporal boundaries. Track IDs are local to individual revisits and are not linked across revisits. Same vehicle and replacement labels are assigned only when identity across revisits can be manually verified. Exact transition times are usually unknown because each region is observed only intermittently. If a region is empty at time t_a and occupied at time t_b, the event start is recorded as $t_s \in (t_a, t_b]$. The event end is defined in the same way from an occupied observation followed by an empty observation.

3.4 Evaluation Protocol

Evaluation across sites. A seven fold LOSO protocol is used over R1, R2, R3, R4, R6, R7, and R8. In each fold, one complete site is held out for evaluation and the remaining six are used for training. No data from the held-out site are used for model fitting or checkpoint selection.

The frozen evaluation set contains 544 region sequences with valid persistence labels, including 325 positive and 219 negative examples. It contains both manually reviewed event proposals and additional sequences with valid negative persistence supervision. Predictions from the seven evaluation folds are pooled for the main evaluation. The composition of each site is reported in Appendix A.5.

Tasks and metrics. The primary task is to determine whether an evaluable region sequence contains a persistent event. Average precision (AP) is used as the primary metric, with AUROC reported as an additional threshold-independent metric. F1 and balanced accuracy are also reported at a threshold of 0.5. Occupancy, persistence at individual revisits, same-vehicle, and replacement labels

provide auxiliary supervision when valid annotations are available. All trainable methods use the same frozen labels and input features and are evaluated with three random seeds, $\{42, 2027, 3407\}$.

4 Problem Formulation

For a fixed semantic region r, intermittent UAV inspection produces an ordered sequence of revisits,

$$\mathcal{S}_r = \left\{ \mathbf{x}^v_{r,t}, \mathbf{x}^e_{r,t}, \mathbf{x}^m_r, \mathbf{q}_{r,t}, \mathbf{a}_{r,t}, \Delta t_{r,t} \right\}_{t=1}^{T_r}, \tag{1}$$

where $\mathbf{x}^v_{r,t}$, $\mathbf{x}^e_{r,t}$, $\mathbf{x}^m_r$, $\mathbf{q}_{r,t}$, $\mathbf{a}_{r,t}$, and $\Delta t_{r,t}$ respectively denote the visual observation, telemetry information, the fixed map region, observation quality, modality availability, and the elapsed time since the previous revisit.

Observation availability and supervision are handled separately. $m^{\text{upd}}_{r,t}$ is used to determine whether revisit t updates the recurrent state. A separate mask $m^k_{r,t} \in \{0, 1\}$ is used for each task, where

$$k \in \{\text{occ}, \text{pers}, \text{id}, \text{repl}\}.$$

These masks indicate whether a valid label is available for occupancy, persistence, vehicle identity, or replacement. An unreliable observation can therefore be excluded from the recurrent update while still providing supervision when a verified label is available.

The primary task is to determine whether a region sequence contains a persistent event. Given the complete revisit sequence, the model predicts

$$p^{\text{cand}}_r = P_\theta \left(y^{\text{cand}}_r = 1 \mid \mathcal{S}_r \right), \tag{2}$$

where $y^{\text{cand}}_r \in \{0, 1\}$ is the persistence label for the sequence.

The model also predicts occupancy and persistence at individual revisits. Vehicle identity and replacement are predicted when verified identity labels are available. These predictions provide auxiliary supervision; persistence of the complete sequence remains the primary objective.

Exact transition times are usually unknown because the region is not observed between consecutive revisits. If a transition is bounded by observations at times t_a and t_b, its time is represented by

$$t_s \in (t_a, t_b].$$

These intervals are retained in the benchmark annotations rather than replaced by artificial point timestamps. Event boundary prediction is not optimized in the experiments reported in this work.

5 U^2-Revisit

5.1 Multimodal Revisit Representation

For each semantic region r and revisit t, U^2-Revisit combines visual, telemetry, and map information describing the same physical region. The visual observation describes its current appearance, telemetry records how and where the observation was acquired, and the map provides stable semantic information about the location. The three modalities are encoded independently as

$$\mathbf{e}^v_{r,t} = E_v(\mathbf{x}^v_{r,t}), \qquad \mathbf{e}^e_{r,t} = E_e(\mathbf{x}^e_{r,t}), \qquad \mathbf{e}^m_r = E_m(\mathbf{x}^m_r). \tag{3}$$

Here, E_v, E_e, and E_m are learnable encoders for the three modalities.

The visual input $\mathbf{x}^v_{r,t}$ is extracted from a crop around region r after alignment with the reference map. Its appearance feature is computed offline using a frozen DINOv2 backbone and then projected with E_v. The telemetry input $\mathbf{x}^e_{r,t}$ describes the UAV state relative to the region, including position, motion, temporal information, and indicators for available telemetry fields. The map input $\mathbf{x}^m_r$ describes the local semantic layout, including parking and roadway context. This map representation is shared by all revisits of the same region.

All three inputs refer to the same fixed semantic region. Visual appearance and acquisition conditions may change between revisits, but the spatial reference remains unchanged. The temporal model therefore processes repeated observations of the same physical location rather than unrelated image crops.

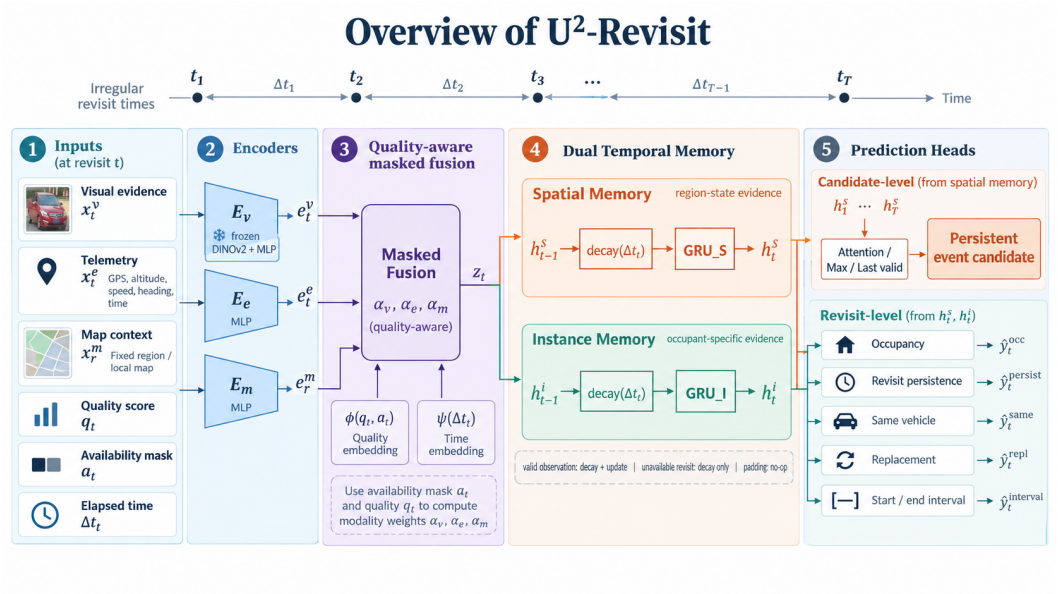


Figure 2: Overview of U^2-Revisit. Available modalities are fused using quality-aware masked weights. Spatial and instance memories decay with elapsed time and support region-state and vehicle-related predictions.

5.2 Quality-Aware Masked Fusion

At each revisit, the model needs to know whether a modality is available and whether the available evidence is reliable. These are different cases. A modality may be completely unavailable, or it may be present but unreliable because of poor visual quality, inaccurate image-to-map registration, low detection confidence, or inconsistent vehicle and region geometry.

For each revisit, a quality descriptor $\mathbf{q}_{r,t}$ and an availability mask $\mathbf{a}_{r,t} \in \{0,1\}^3$ are used for the visual, telemetry, and map modalities. Given the representations from Section 5.1, a gating network predicts one fusion logit for each modality,

$$\mathbf{g}_{r,t} = G([\mathbf{q}_{r,t}; \mathbf{a}_{r,t}]) \,. \tag{4}$$

For modality $k \in \{v, e, m\}$, unavailable inputs are removed before normalization:

$$\alpha_{r,t}^{k} = \frac{a_{r,t}^{k} \exp(g_{r,t}^{k})}{\sum_{j \in \{v,e,m\}} a_{r,t}^{j} \exp(g_{r,t}^{j})}, \qquad k \in \{v, e, m\}. \tag{5}$$

The fused representation is then

$$\mathbf{f}_{r,t} = \sum_{k \in \{v,e,m\}} \alpha_{r,t}^{k} \mathbf{e}_{r,t}^{k}. \tag{6}$$

An unavailable modality receives zero weight. The weights of the remaining modalities depend on the quality of the current observation.

The quality and availability information is also retained for temporal reasoning. The elapsed time since the previous revisit is encoded together with these reliability cues and the fused representation to form

$$\mathbf{z}_{r,t} = \Phi([\mathbf{f}_{r,t}; E_{\Delta}(\log(1 + \Delta t_{r,t})); E_q([\mathbf{q}_{r,t}; \mathbf{a}_{r,t}])]) \,. \tag{7}$$

The revisit token $\mathbf{z}_{r,t}$ therefore contains the fused observation, the elapsed time, and information about the reliability of that observation. It is used as input to the recurrent module in Section 5.3.

5.3 Dual Memory with Time Decay

Region occupancy and vehicle identity may change in different ways across revisits. A region can remain occupied after the vehicle changes. An unavailable observation also does not by itself indicate a change in either state. U^2-Revisit maintains two recurrent states: a spatial memory $h^S_{r,t}$ for region occupancy and persistence, and an instance memory $h^I_{r,t}$ for evidence about the observed vehicle.

Before processing revisit t, each memory decays according to the elapsed time:

$$\bar{h}^c_{r,t} = \exp\left(-\frac{\Delta t_{r,t}}{\tau_c}\right) h^c_{r,t-1}, \qquad c \in \{S, I\}, \tag{8}$$

where the retention time constants for the spatial and instance memories are set to $\tau_S = 120\,\mathrm{s}$ and $\tau_I = 90\,\mathrm{s}$, respectively.

For each non-padding revisit, the update mask $m^{\mathrm{upd}}_{r,t}$ determines whether the decayed state is updated with the current observation:

$$h^c_{r,t} = \begin{cases} \mathrm{GRU}_c\left(z_{r,t}, \bar{h}^c_{r,t}\right), & m^{\mathrm{upd}}_{r,t} = 1, \\ \bar{h}^c_{r,t}, & m^{\mathrm{upd}}_{r,t} = 0, \end{cases} \qquad c \in \{S, I\}. \tag{9}$$

The spatial and instance memories use separate GRU units. The spatial memory represents region occupancy and persistence. The instance memory retains information about the observed vehicle. An auxiliary head predicts vehicle replacement from these representations, but its prediction is not fed back into the recurrent transition.

Padding positions bypass both time decay and recurrent updates, leaving the two memory states unchanged.

5.4 Persistence Reasoning and Training Objectives

The final persistence decision uses the complete sequence of spatial memory states. This sequence is summarized using attention pooling, element wise max pooling, and the last valid state. For each revisit,

$$\gamma_{r,t} = w_a^\top h^S_{r,t} + b_a, \qquad \beta_{r,t} = \frac{m_{r,t}\exp(\gamma_{r,t})}{\sum_{j=1}^{T_r} m_{r,j}\exp(\gamma_{r,j})}. \tag{10}$$

Here, $m_{r,t}$ indicates whether position t is valid. The attention representation is

$$h^{\mathrm{att}}_r = \sum_{t=1}^{T_r} \beta_{r,t} h^S_{r,t}. \tag{11}$$

This representation is combined with the element-wise maximum state $h^{\max}_r$ and the last valid state h^{last}_r to predict persistence for the complete sequence:

$$p^{\mathrm{cand}}_r = \sigma\left(H_{\mathrm{cand}}\left[h^{\mathrm{att}}_r; h^{\max}_r; h^{\mathrm{last}}_r\right]\right). \tag{12}$$

The recurrent states are also used for auxiliary predictions. Occupancy and persistence at each revisit are predicted from the spatial memory,

$$\ell^{\mathrm{occ}}_{r,t} = H_{\mathrm{occ}}(h^S_{r,t}), \qquad \ell^{\mathrm{pers}}_{r,t} = H_{\mathrm{pers}}(h^S_{r,t}). \tag{13}$$

Verified identity and replacement outputs use both the spatial and instance memories,

$$\left[\ell^{\mathrm{id}}_{r,t}, \ell^{\mathrm{repl}}_{r,t}\right] = H_{\mathrm{inst}}\left([h^S_{r,t}; h^I_{r,t}]\right). \tag{14}$$

A heteroscedastic auxiliary objective is additionally used for occupancy and revisit-level persistence. From $[h^S_{r,t}; h^I_{r,t}]$, the model predicts task-specific log-variance and optimizes $\exp(-u)\mathrm{BCE} + \frac{1}{2}u$ over valid positions Appendix C.3.

The model is trained with the weighted objective

$$\mathcal{L} = 3.0\mathcal{L}_{\text{cand}} + 0.25\mathcal{L}_{\text{occ}} + 0.25\mathcal{L}_{\text{pers}} + 0.10\mathcal{L}_{\text{id}} + 0.10\mathcal{L}_{\text{repl}} + 0.05\mathcal{L}_{\text{uncert}}. \tag{15}$$

Each auxiliary term is computed only when the corresponding supervision mask is active. Persistence of the complete sequence remains the primary optimization target.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Evaluation protocol. All trainable methods follow the frozen seven-fold LOSO protocol defined in Section 3.4. Predictions from the seven held-out sites are pooled over 544 evaluable region sequences, including 325 positive and 219 negative examples. Event AP is the primary metric, with AUROC reported as an additional threshold-independent metric. All trainable methods are evaluated with the same three random seeds, $\{42, 2027, 3407\}$.

Training details. All learned methods use the same frozen input features and supervision and are trained for 20 epochs using AdamW with a learning rate of 10^{-4}, weight decay of 0.01, and batch size 32. The bfloat16 is used to mix precision without a learning rate schedule or early stopping. The hidden dimension is 192, and visual features are extracted offline using a frozen DINOv2 ViT-B/14 backbone. All LOSO results use the final checkpoint at epoch 20. Training-budget sensitivity is reported in Appendix D.1.

Baselines. U^2-Revisit is compared with several baselines that progressively incorporate repeated observation aggregation and temporal modeling. Single Revisit predicts from one observation only. Unordered Set aggregates repeated observations without using revisit order. GRU processes the ordered revisit sequence, and Time-aware GRU additionally uses the elapsed time between revisits. Single Memory uses one recurrent state with time decay for all prediction heads. U^2-Revisit uses the same input modalities and supervision but maintains separate spatial and instance memories for region state and vehicle-related evidence.

6.2 BENCHMARK CHARACTERISTICS

The frozen LOSO benchmark contains 544 region sequences with valid persistence labels, including 325 positive and 219 negative examples, distributed across seven sites. The class distribution varies substantially across sites. For example, R1 contains 204 sequences with a positive prevalence of 27.5%, whereas R2 contains 148 sequences with a positive prevalence of 89.2%. Several sites are much smaller, including R6 with 16 sequences and R8 with 15.

These differences reflect the natural event distribution rather than an artificially balanced benchmark. The original population at each site is retained, and predictions from all seven LOSO folds are pooled for the main evaluation. Event AP is used as the primary metric because it evaluates ranking performance under class imbalance. AUROC, F1, and balanced accuracy provide additional views of performance. Site-specific results are reported in Appendix D.4.

6.3 MAIN RESULTS AND ABLATION ANALYSIS

Table 2 reports the pooled results from the seven LOSO folds. Single Revisit achieves a mean Event AP of 0.8840, which increases to 0.9462 with Unordered Set and to 0.9594 ± 0.0082 with U^2-Revisit. U^2-Revisit achieves the highest mean Event AP among all evaluated methods, with absolute improvements of 0.0754 over Single Revisit and 0.0132 over Unordered Set.

The 0.0622 AP gain from Single Revisit to Unordered Set shows the benefit of aggregating evidence across repeated observations. U^2-Revisit, which additionally models revisit order, elapsed time, and separate spatial and instance states, further increases Event AP to 0.9594.

Table 3 reports the component ablations. Removing quality-aware fusion, time decay, and the separate instance memory reduces mean Event AP by 0.0112, 0.0073, and 0.0045, respectively. Removing time decay slightly increases mean AUROC from 0.9453 to 0.9468.

Table 2: Persistent-event detection under the seven-site LOSO protocol. AP and AUROC are reported as mean ± sample standard deviation over three random seeds. F1 and balanced accuracy are computed at a threshold of 0.5. Best results are shown in bold.

Method	Event AP	AUROC	F1	BAcc
Statistics	0.6047	0.4222	0.7696	0.5895
Single Revisit	0.8840 ± 0.0380	0.8857 ± 0.0266	0.8855	0.8373
Unordered Set	0.9462 ± 0.0270	$\mathbf{0.9465 \pm 0.0163}$	**0.9310**	**0.9059**
GRU	0.9506 ± 0.0131	0.9388 ± 0.0115	0.9136	0.8812
Time-aware GRU	0.9519 ± 0.0098	0.9399 ± 0.0082	0.9199	0.8879
Single Memory	0.9549 ± 0.0093	0.9427 ± 0.0047	0.9202	0.8891
U^2-Revisit	$\mathbf{0.9594 \pm 0.0082}$	0.9453 ± 0.0074	0.9269	0.9013

Table 3: Component ablations under the seven-site LOSO protocol. Results are mean ± sample standard deviation over three random seeds.

Variant	Event AP	AUROC
U^2-Revisit	$\mathbf{0.9594 \pm 0.0082}$	0.9453 ± 0.0074
w/o Quality-aware Fusion	0.9482 ± 0.0260	0.9436 ± 0.0109
w/o Time Decay	0.9521 ± 0.0196	$\mathbf{0.9468 \pm 0.0127}$
w/o Instance Memory	0.9549 ± 0.0093	0.9427 ± 0.0047

7 Conclusion

Persistent event verification under intermittent UAV revisits is studied using fixed semantic regions as spatial anchors. The benchmark preserves the distinction between confirmed absence and missing or unreliable evidence. U^2-Revisit combines quality-aware multimodal fusion, elapsed-time modeling, and separate spatial and instance memories to aggregate evidence across revisits.

Under the seven-site LOSO protocol, U^2-Revisit achieves the highest mean Event AP of 0.9594 ± 0.0082 among all evaluated methods. Repeated observation aggregation provides the main improvement over single revisit reasoning, while ordered temporal modeling and separate spatial and instance memories further improve Event AP. Quality-aware fusion produces the largest performance drop when removed in the component ablations.

AI Use Statement

Generative AI tools were used to assist with research brainstorming, language translation, LaTeX preparation, and code generation and debugging for parts of the data-processing and experimental pipeline. All AI-assisted code was reviewed and tested by the authors, and all experimental results, mathematical formulations, citations, and scientific claims were independently checked against the underlying implementation and data. The authors take full responsibility for the final content of the paper and all artifacts produced with the assistance of generative AI.

Ethics Statement

The dataset contains aerial imagery of public road and parking environments together with UAV telemetry and geospatial information. To reduce privacy and security risks, released data will exclude information that could unnecessarily reveal sensitive geographic details or personal identity, and geographic metadata may be withheld or coarsened where appropriate. The benchmark is designed for research on aerial perception and persistent-event reasoning and does not provide ground-truth legal judgments or personal identity labels. Missing or ambiguous observations are explicitly represented rather than converted into unsupported labels.

REFERENCES

Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc van Gool. MUSES: The multi-sensor semantic perception dataset for driving under uncertainty. In *Proceedings of the European Conference on Computer Vision*, pp. 21–38, 2024.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

Zhiyun Deng, Austin Case, and Luis Sentis. AirLock+: Scaling UAV-to-satellite image registration for target geolocalization and geospatial augmented reality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3340–3349, 2026.

Minh Kha Do, Kang Han, Phu Lai, Khoa T. Phan, and Wei Xiang. RobSense: A robust multimodal foundation model for remote sensing with static, temporal, and incomplete data adaptability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7427–7436, 2025.

Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (MAVREC): Can multi-view improve aerial visual perception? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22678–22690, 2024.

Robert Fonod, Haechan Cho, Hwasoo Yeo, and Nikolas Geroliminis. Advanced computer vision for extracting georeferenced vehicle trajectories from drone imagery. *Transportation Research Part C: Emerging Technologies*, 178:105205, 2025.

Carlos García Meixide, Marcos Matabuena, Louis Abraham, and Michael R. Kosorok. Neural interval-censored survival regression with feature selection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 17(4):e11704, 2024.

Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27852–27862, 2024.

Jens Hellekes, Manuel Mühlhaus, Reza Bahmanyar, Seyed Majid Azimi, and Franz Kurz. VETRA: A dataset for vehicle tracking in aerial imagery — new challenges for multi-object tracking. In *Proceedings of the European Conference on Computer Vision*, pp. 52–70, 2024.

Jae-Woo Kim and Ue-Hwan Kim. Towards generalizable scene change detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24463–24473, 2025.

Youngjoo Kim, Seungho Back, Dongchan Song, and Byung-Yoon Lee. Aerial map-based navigation by ground object pattern matching. *Drones*, 8(8):375, 2024.

Christian Klötergens, Vijaya Krishna Yalavarthi, Randolf Scholz, Maximilian Stubbemann, Stefan Born, and Lars Schmidt-Thieme. Physiome-ODE: A benchmark for irregularly sampled multivariate time-series forecasting based on biological ODEs. In *International Conference on Learning Representations*, 2025.

Ilya Kuleshov, Evgenia Romanenkova, Vladislav Zhuzhel, Galina Boeva, Evgeni Vorsin, and Alexey Zaytsev. DeNOTS: Stable deep neural ODEs for time series. In *International Conference on Learning Representations*, 2026.

Hongchao Li, Jingong Chen, Aihua Zheng, Yong Wu, and Yonglong Luo. Day-night cross-domain vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12626–12635, 2024a.

Yiming Li, Zhiheng Li, Nuo Chen, Moonjun Gong, Zonglin Lyu, Zehong Wang, Peili Jiang, and Chen Feng. Multiagent multitraversal multimodal self-driving: Open MARS dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22041–22051, 2024b.

Xingyue Liu, Jiahao Qi, Chen Chen, KangCheng Bin, and Ping Zhong. UCM-VeID V2: A richer dataset and a pre-training method for UAV cross-modality vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22286–22295, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Seyedehanita Madani, Rama Chellappa, and Vishal M. Patel. DiffRegCD: Integrated registration and change detection with diffusion features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7503–7512, 2026.

Fernando Moreno-Pino, Álvaro Arroyo, Harrison Waldon, Xiaowen Dong, and Álvaro Cartea. Rough transformers: Lightweight and continuous time series modelling through signature patching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106264–106294, 2024.

Zhihao Nie, Luyi Xue, Zhenyu Fang, Jinchang Ren, Yufeng Wei, and Jiangbin Zheng. M3OT: A multi-drone multi-modality dataset for multi-object tracking. *Scientific Data*, 12:1927, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Byoungwoo Park, Hyungi Lee, and Juho Lee. Amortized control of continuous state space Feynman-Kac model for irregular time series. In *International Conference on Learning Representations*, 2025.

Liang Qin, Min Wang, Xingyu Lu, Aowen Qiu, Wengang Zhou, and Houqiang Li. UAST: Unified active search and tracking for arbitrary targets with UAVs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13464–13473, 2026.

Ahmed H. Shahin, An Zhao, Alexander C. Whitehead, Daniel C. Alexander, Joseph Jacob, and David Barber. CenTime: Event-conditional modelling of censoring in survival analysis. *Medical Image Analysis*, 91:103016, 2024.

Linzhi Shang, Chen Min, Juan Wang, Liang Xiao, Dawei Zhao, and Yiming Nie. Aerial-ground cross-view vehicle re-identification: A benchmark dataset and baseline. *Remote Sensing*, 17(15): 2653, 2025.

Francesco Spinnato and Cristiano Landi. PYRREGULAR: A unified framework for irregular time series, with classification benchmarks. In *International Conference on Learning Representations*, 2026.

Cheng-Yen Yang, Hsiang-Wei Huang, Zhongyu Jiang, Heng-Cheng Kuo, Jie Mei, Chung-I Huang, and Jenq-Neng Hwang. Sea you later: Metadata-guided long-term re-identification for UAV-based multi-object tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 805–812, 2024.

Jingtao Ye, Kexin Zhang, Xunchi Ma, Yuechan Li, Guangming Zhu, Peiyi Shen, Linhua Jiang, Xiangdong Zhang, and Liang Zhang. Breaking smooth-motion assumptions: A UAV benchmark for multi-object tracking in complex and adverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13594–13603, 2026.

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 *European Conference on Computer Vision*, pp. 1–21, 2022.

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3D: Revisiting change detection and captioning from a video modeling perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24011–24022, 2025.

A Dataset Construction and Benchmark Statistics

A.1 Acquisition Protocol

The dataset was collected using a DJI Mini 3 Pro during repeated near-nadir aerial inspection. Videos were recorded in 4K at approximately 30 fps together with synchronized GPS/SRT telemetry and flight trajectories. At each site, the UAV repeatedly followed a predefined loop or bidirectional route, producing observations of the same geographic areas at different times.

Data were collected at 11 physical sites, which serve different roles in the benchmark. The main benchmark uses seven sites: R1, R2, R3, R4, R6, R7, and R8. These sites cover different roadway layouts, designated parking areas, traffic conditions, and observation geometries. R10 and R11 are reserved for the additional evaluation reported in Appendix D.5 and are not used in the main benchmark. R5 is retained for diagnostic analysis because its registration pipeline was repaired and inspected repeatedly during method development. R9 completed processing and manual review after the main benchmark protocol had been frozen, so it is retained as an additional benchmark site rather than added retrospectively to the main evaluation.

A.2 Semantic Regions and Revisit Construction

Each site has a georeferenced reference map containing manually defined semantic regions, mainly roadway and designated parking polygons. These regions are fixed in map coordinates and remain consistent across repeated UAV passes.

For each revisit, UAV telemetry provides an initial geographic estimate, and visual image-to-map registration then refines the alignment. The registered image footprint determines which semantic regions are sufficiently observed. Multiple valid views of the same region within one pass are combined into a single observation. Observations associated with the same semantic region are ordered by acquisition time to form a revisit sequence for that region.

Figure 3 illustrates this construction process. Semantic regions are defined independently of temporary vehicle detections. The same map region provides a fixed spatial reference across revisits even when vehicle appearance, viewing geometry, or observation quality changes.

A.3 Observation Evidence

Each observation of a semantic region is assigned one of six evidence states:

$$\mathcal{C} = \left\{ \begin{array}{lll} \texttt{confirmed_empty}, & \texttt{vehicle_detected}, & \texttt{detection_missing}, \\ \texttt{region_not_observed}, & \texttt{occluded}, & \texttt{registration_failed} \end{array} \right\}.$$

The first two states provide direct evidence about region occupancy. The remaining four indicate that the observation is missing or unreliable. In particular, the absence of a reliable vehicle observation is not treated as evidence that the region is empty. This distinction prevents limited spatial coverage, registration failure, or occlusion from being converted into false negative occupancy evidence.

A.4 Persistent Event Annotation

Persistent event annotation follows the two-stage procedure described in the main paper. A lightweight proposal stage uses detector outputs and vehicle tracks aligned to the reference map. It identifies region sequences with repeated vehicle observations, stationary evidence within a revisit, and spatial consistency across revisits.

Across the seven main sites, this stage produces 362 event proposals. Among them, 353 have deterministic labels under the frozen protocol: 325 are accepted as persistent events and 28 are rejected. Proposals without deterministic labels are excluded from persistence evaluation.

The frozen benchmark also contains region sequences that were not produced by the proposal stage but have valid negative persistence labels from the annotation pipeline. These additional negatives extend the evaluation beyond the manually reviewed proposal set.

Table 4: Site-wise composition of the primary seven-site benchmark. "Evaluable" denotes region sequences with valid persistence labels under the frozen protocol.

Site	Sequences	Evaluable	Positive	Negative
R1	207	204	56	148
R2	194	148	132	16
R3	39	23	10	13
R4	124	91	69	22
R6	17	16	11	5
R7	50	47	35	12
R8	19	15	12	3
Total	**650**	**544**	**325**	**219**

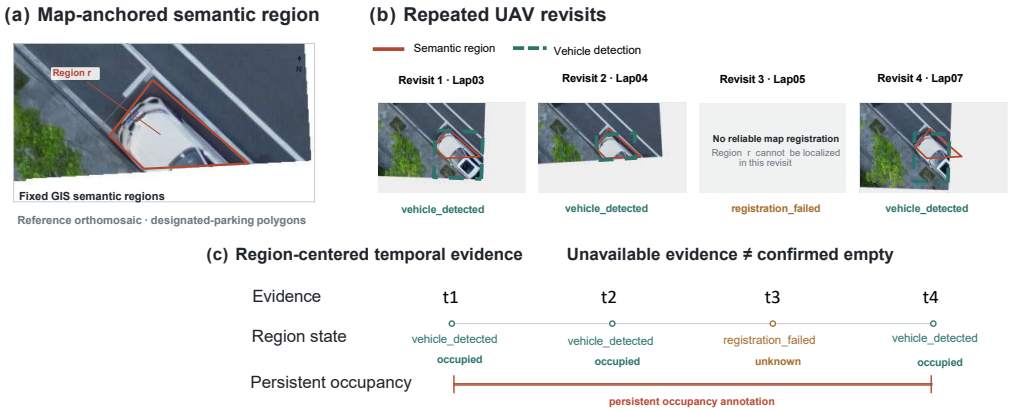


Figure 3: Illustration of the revisit benchmark. (a) A fixed semantic region is defined on the georeferenced reference map independently of vehicle detections. (b) The same physical region is observed across repeated UAV revisits, while some observations may be unavailable because of registration failure or other sensing limitations. (c) Evidence from multiple revisits is combined for persistence reasoning; unavailable evidence remains distinct from confirmed absence.

Vehicle identity is annotated separately from persistent occupancy. Track IDs are local to individual revisits and are not propagated across revisits. Same-vehicle and replacement labels are assigned only when identity can be manually verified. Event boundaries are recorded as time intervals between observations and are not optimized in the experiments reported in this paper.

A.5 Primary Benchmark Population

The seven main sites contain 650 revisit sequences in total. Of these, 544 have valid persistence labels for the complete sequence, including 325 positive and 219 negative examples. The 544 evaluable sequences consist of 353 proposals with deterministic labels and 191 additional negative sequences that were not produced by the proposal stage.

The remaining 106 sequences do not have valid persistence labels under the frozen protocol and are excluded from the primary evaluation. Table 4 reports the composition of each site.

B Preprocessing and Spatial Registration

B.1 Vehicle Detection and Within-Revisit Tracking

Vehicles are first detected using a YOLO detector. ByteTrack (Zhang et al., 2022) is then applied independently to each revisit to reduce fragmented detections and obtain vehicle tracks within that revisit. Track IDs are local to a single revisit and are not propagated to later observations.

The tracking stage therefore does not provide vehicle identity across revisits. Identity across repeated observations must be determined from the observations associated with the same semantic region.

B.2 Image-to-Map Registration

Each UAV observation is aligned with a georeferenced reference map. UAV telemetry provides an initial geographic estimate, and visual correspondences between the aerial image and the reference map are then used to refine the alignment. Robust geometric estimation rejects inconsistent correspondences and produces the transformation used to project image observations into map coordinates. If reliable registration cannot be obtained, the observation is marked as unavailable. Registration failure is not interpreted as evidence that the corresponding semantic region is empty.

B.3 Region-Centered Observation Construction

After image-to-map registration, each fixed semantic region is projected into the corresponding UAV observation. Vehicle detections and tracks that overlap the projected region are associated with that region, producing one observation for each region and revisit. Observations associated with the same semantic region are then ordered by acquisition time to form the revisit sequence used by the model.

Along with the visual, telemetry, and map inputs, the preprocessing pipeline records observation quality and modality availability for each revisit. These signals describe factors such as spatial coverage and registration reliability. They are passed to the fusion module so that unreliable observations can receive lower weight or be masked rather than being treated as equally reliable evidence.

C Implementation Details

C.1 Input Representations

Visual features are extracted offline using a frozen DINOv2 ViT-B/14 encoder, producing a 768-dimensional feature for each valid region observation. Telemetry and map inputs are represented by 32-dimensional and 128-dimensional vectors, respectively. Each revisit also retains a 5-dimensional quality descriptor together with modality availability indicators.

The elapsed time Δt between consecutive revisits is preserved as an explicit model input. Before fusion, the visual, telemetry, and map representations are projected into a common latent space.

C.2 Sequence Model

All trainable sequence models use a hidden dimension of 192. U^2-Revisit maintains separate spatial and instance memories with retention constants $\tau_S = 120\,\mathrm{s}$ and $\tau_I = 90\,\mathrm{s}$.

For each non-padding revisit, both memories first decay according to the elapsed time. If the recurrent update mask is active, the decayed states are then updated using the current revisit representation. If the update mask is inactive, only time decay is applied. Padding positions leave both memory states unchanged.

The spatial memory represents region occupancy and persistence, while the instance memory retains evidence about the observed vehicle. Replacement predictions are not fed back into the recurrent transition, and no replacement-conditioned reset or gating mechanism is used.

The final persistence prediction combines attention pooling, element-wise max pooling, and the last valid spatial memory state. Occupancy and persistence at individual revisits are predicted from the spatial memory. Same-vehicle and replacement predictions use the spatial and instance memories jointly when valid identity labels are available.

C.3 OPTIMIZATION

All learned methods are trained for 20 epochs using AdamW with a learning rate of 1×10^{-4}, weight decay of 0.01, and batch size 32. Training uses bfloat16 mixed precision where supported, without a learning rate schedule or early stopping. All formal comparisons use the final checkpoint at epoch 20 (`last.pt`); `best.pt` is not used for evaluation or checkpoint selection.

All methods are evaluated with the same three random seeds, $\{42, 2027, 3407\}$. The training objective is

$$\mathcal{L} = 3.0\mathcal{L}_{\text{cand}} + 0.25\mathcal{L}_{\text{occ}} + 0.25\mathcal{L}_{\text{pers}} + 0.10\mathcal{L}_{\text{id}} + 0.10\mathcal{L}_{\text{repl}} + 0.05\mathcal{L}_{\text{uncert}}.$$

The uncertainty term is a heteroscedastic auxiliary objective for occupancy and revisit-level persistence. From the joint state $[h^S_{r,t}; h^I_{r,t}]$, the model predicts task-specific log-variance $u^k_{r,t}$ for $k \in \{\text{occ}, \text{pers}\}$:

$$\left[u^{\text{occ}}_{r,t}, u^{\text{pers}}_{r,t}\right] = \text{clip}\left(H_{\text{uncert}}\left([h^S_{r,t}; h^I_{r,t}]\right), -10, 10\right).$$

For each task k, let $\ell^k_{r,t}$ denote the corresponding classification logit and $y^k_{r,t} \in \{0, 1\}$ its binary target. The binary cross-entropy loss at each revisit is

$$b^k_{r,t} = -y^k_{r,t} \log \sigma(\ell^k_{r,t}) - (1 - y^k_{r,t}) \log\left(1 - \sigma(\ell^k_{r,t})\right),$$

where $\sigma(\cdot)$ denotes the sigmoid function. The heteroscedastic loss for that prediction is then

$$\ell^{\text{het}}_{r,t,k} = \exp(-u^k_{r,t})\, b^k_{r,t} + \frac{1}{2} u^k_{r,t}.$$

Let $m^k_{r,t} \in \{0, 1\}$ denote the validity mask for task k, including the sequence-validity mask that excludes padding positions. The masked loss for each task is

$$\mathcal{L}^{\text{het}}_k = \frac{\sum_{r,t} m^k_{r,t}\, \ell^{\text{het}}_{r,t,k}}{\max\left(1, \sum_{r,t} m^k_{r,t}\right)}.$$

The final uncertainty objective averages the occupancy and persistence terms:

$$\mathcal{L}_{\text{uncert}} = \frac{1}{2}\left(\mathcal{L}^{\text{het}}_{\text{occ}} + \mathcal{L}^{\text{het}}_{\text{pers}}\right).$$

No independent uncertainty target is introduced. The uncertainty objective reuses the binary occupancy and revisit-level persistence targets together with their task-specific validity masks. Missing or unreliable observations without valid task supervision are therefore excluded from this loss rather than treated as uncertainty-positive examples.

Persistence of the complete sequence remains the primary optimization target, and event boundary intervals are retained in the benchmark annotations but are not optimized in the experiments reported in this paper.

C.4 FORMAL EVALUATION PROTOCOL

The primary experiments use a seven-fold LOSO protocol over R1, R2, R3, R4, R6, R7, and R8. In each fold, one site is used for evaluation and the remaining six are used for training. Site IDs are disjoint between training and evaluation.

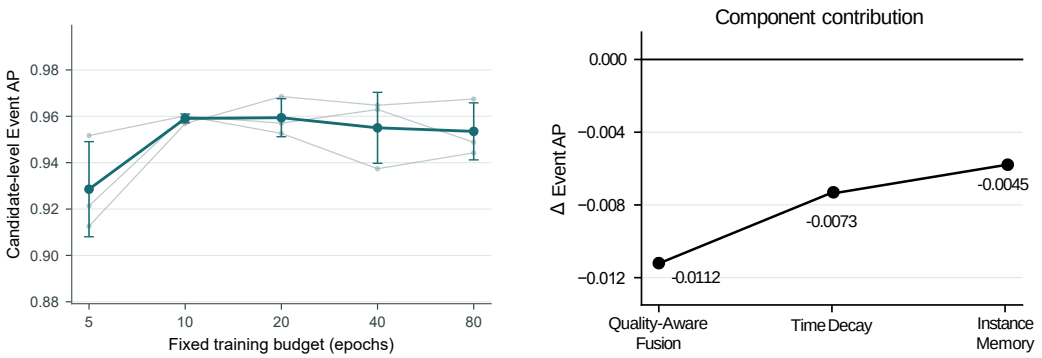
Predictions from the seven evaluation folds are pooled over 544 region sequences with valid persistence labels, including 325 positive and 219 negative examples. Data from the evaluation site are not used for checkpoint selection.

R10 and R11 are evaluated separately as additional external sites. They are not included in the primary LOSO benchmark and are not used for model training, checkpoint selection, or hyperparameter tuning.

Table 5: Key implementation and optimization settings of U^2-Revisit.

Architecture		**Training**	
Setting	**Value**	**Setting**	**Value**
Visual backbone	DINOv2 ViT-B/14	Optimizer	AdamW
Visual feature dim.	768	Learning rate	1×10^{-4}
Telemetry dim.	32	Weight decay	0.01
Map feature dim.	128	Batch size	32
Quality dim.	5	Training budget	20 epochs
Hidden dimension	192	Evaluation checkpoint	`last.pt`
Spatial retention	$\tau_S = 120$ s	Mixed precision	bfloat16
Instance retention	$\tau_I = 90$ s	LR scheduler	None
Memory structure	Dual	Early stopping	None
Candidate readout	Attention + max + last	Random seeds	42 / 2027 / 3407

Loss weights: $\lambda_{\text{cand}} = 3.00$, $\lambda_{\text{occ}} = 0.25$, $\lambda_{\text{pers}} = 0.25$, $\lambda_{\text{id}} = 0.10$, $\lambda_{\text{repl}} = 0.10$
$\lambda_{\text{uncert}} = 0.05$; interval-boundary loss: not active.



(a) Training budget Sensitivity

(b) Effect of component removal

Figure 4: Additional analysis of training stability and model components. (a) Mean Event AP under different fixed training budgets. (b) Change in mean Event AP after removing individual components from U^2-Revisit.

D Additional Experimental Results

D.1 Training Budget Sensitivity

Training budgets of 5, 10, 20, 40, and 80 epochs are evaluated using the same seven site LOSO protocol and three random seeds. Each run is evaluated using its final checkpoint. A fixed training budget of 20 epochs is therefore used for all main comparisons.

D.2 Results Across Random Seeds

Table 6 reports the results for each random seed. For every seed, predictions from the seven LOSO folds are pooled over the same 544 evaluable region sequences. All methods use the same labels, input features, 20-epoch training budget, and final checkpoint evaluation rule.

U^2-Revisit achieves the highest Event AP for two of the three seeds and the highest mean Event AP overall.

Table 6: Per-seed persistent-event detection results under the seven-site LOSO protocol. Predictions from all seven held-out folds are pooled before computing each metric.

	Event AP			AUROC		
Method	**42**	**2027**	**3407**	**42**	**2027**	**3407**
Single Revisit	0.8748	0.9258	0.8516	0.8870	0.9116	0.8585
Unordered Set	0.9208	**0.9745**	0.9433	0.9402	**0.9650**	0.9342
GRU	0.9432	0.9657	0.9428	0.9363	0.9512	0.9287
Time-aware GRU	0.9467	0.9633	0.9458	0.9385	0.9486	0.9325
Single Memory	0.9456	0.9641	0.9551	**0.9427**	0.9474	0.9379
U^2-Revisit	**0.9526**	0.9685	**0.9570**	0.9419	0.9538	**0.9402**

Table 7: Per-seed ablation results under the seven-site LOSO protocol. Predictions from all seven held-out folds are pooled before computing each metric.

	Event AP			AUROC		
Variant	**42**	**2027**	**3407**	**42**	**2027**	**3407**
U^2-Revisit	0.9526	0.9685	0.9570	0.9419	0.9538	0.9402
w/o Quality-aware Fusion	0.9193	0.9696	0.9557	0.9401	0.9559	0.9349
w/o Time Decay	0.9317	0.9707	0.9538	0.9428	0.9610	0.9366
w/o Instance Memory	0.9456	0.9641	0.9551	0.9427	0.9474	0.9379

D.3 Component Ablations

Table 7 reports the ablation results for each random seed, and Figure 4(b) summarizes the corresponding changes in mean Event AP.

Removing quality-aware fusion reduces mean Event AP from 0.9594 to 0.9482, the largest decrease among the tested components. Removing time decay reduces mean Event AP to 0.9521, while removing the separate instance memory gives 0.9549. The corresponding AP drops are 0.0112, 0.0073, and 0.0045, respectively, with quality-aware fusion showing the largest effect on the primary metric.

D.4 Generalization Across Sites

The pooled LOSO evaluation summarizes performance across unseen sites but can hide differences between individual environments. Event AP is reported separately for each evaluation site in Table 8. The number of evaluable sequences and the class balance vary substantially across sites, from 15 sequences in R8 to 204 in R1, so these results should be interpreted together with the corresponding sample counts.

Performance varies across environments. U^2-Revisit achieves the highest mean Event AP on R3, R4, and R7, and ties with the other methods using repeated observations on R8. Unordered Set performs best on R1 and R2, while the recurrent baselines perform better on R6. R3 remains the most challenging site, whereas performance on R7 and R8 is close to saturation.

For this reason, pooled Event AP over all seven LOSO folds remains the primary benchmark metric. Results for individual sites are used to show variation across environments rather than to define the main comparison.

D.5 Additional Evaluation on External Sites

Models trained on the full seven site development population, R1, R2, R3, R4, R6, R7, and R8, are additionally evaluated on two external sites, R10 and R11. Neither site is included in the primary LOSO benchmark. They are not used for checkpoint selection, hyperparameter tuning, or threshold tuning.

Table 8: Site-wise Event AP under the seven-site LOSO protocol. Results are mean over three random seeds, and positive/negative counts are shown in parentheses. Bold denotes the highest unrounded mean within each site.

Site	Single Revisit	Unordered Set	GRU	Time-aware GRU	Single Memory	U^2-Revisit
R1 (56/148)	0.743	**0.961**	0.953	0.951	0.941	0.925
R2 (132/16)	0.988	**0.992**	0.989	0.988	0.988	0.990
R3 (10/13)	0.518	0.468	0.460	0.492	0.530	**0.534**
R4 (69/22)	0.900	0.972	0.973	0.973	0.975	**0.976**
R6 (11/5)	0.936	0.958	**0.967**	**0.967**	0.964	0.955
R7 (35/12)	0.975	0.998	0.998	0.998	0.998	**0.998**
R8 (12/3)	0.994	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 9: Additional external-site evaluation on R10 and R11. Results are mean $\pm$ sample standard deviation over three random seeds. R10 contains 131 candidates (122/9), and R11 contains 80 candidates (68/12).

Method	R10		R11		R10+R11	
	AP	AUROC	AP	AUROC	AP	AUROC
Single Revisit	0.970±0.000	0.733±0.012	0.888±0.014	0.599±0.040	0.944±0.005	0.663±0.016
Unordered Set	0.985±0.003	0.838±0.019	0.922±0.012	0.734±0.006	0.963±0.007	0.789±0.010
GRU	0.979±0.003	0.785±0.024	0.936±0.008	0.760±0.014	0.964±0.006	0.780±0.026
Time-aware GRU	0.984±0.001	0.824±0.003	0.938±0.010	0.757±0.024	0.969±0.002	0.797±0.008
Single Memory	0.982±0.003	0.815±0.018	0.945±0.002	0.779±0.006	0.972±0.002	0.812±0.010
U^2-Revisit	0.983±0.004	0.825±0.021	0.945±0.006	0.780±0.017	0.973±0.001	0.817±0.010

Table 9 reports the results. On the pooled R10+R11 evaluation, U^2-Revisit achieves a mean Event AP of 0.973 ± 0.001 and a mean AUROC of 0.817 ± 0.010. Differences on the individual sites are smaller and vary across methods. Unordered Set obtains the highest Event AP on R10, while U^2-Revisit and Single Memory obtain the same mean Event AP on R11.

R10 and R11 are strongly class imbalanced, with 122/9 and 68/12 positive/negative examples, respectively. These results are reported separately from the primary LOSO benchmark and serve as additional evidence of transfer to unseen sites.

E Qualitative Analysis

Representative revisit sequences. Figure 5 shows representative sequences from fixed semantic regions observed over multiple UAV revisits. Reliable occupancy evidence may be interrupted by missing or unreliable observations. U^2-Revisit combines the available evidence across revisits without treating an unavailable observation as evidence that the region is empty. This allows persistent occupancy to be inferred from incomplete observation sequences.

Challenging observation conditions. Figure 6 shows several conditions that make persistent event verification difficult, including registration failure, partial spatial coverage, state transitions during unobserved intervals, and long gaps between revisits. These examples illustrate why unavailable observations must be distinguished from confirmed absence. Observation quality, modality availability, and elapsed time are therefore retained as inputs to the model.

F Reproducibility Statement

An anonymized implementation of the benchmark and model pipeline will be released. The release will include candidate generation, preprocessing, LOSO split manifests, training and evaluation configurations, and scripts for reproducing the main results.



Figure 5: Representative revisit sequences for fixed semantic regions. Each row follows a fixed semantic region across repeated revisits. (a) Persistent occupancy remains supported despite an unavailable observation caused by registration failure. (b) A region transitions from confirmed empty to occupied between consecutive revisits. (c) Persistent occupancy is observed across strongly irregular revisit intervals. Semantic regions are fixed in map coordinates, while vehicle detections provide transient observation evidence.



Figure 6: Representative challenging observation conditions under intermittent aerial inspection. (a) Registration failure prevents reliable localization of a semantic region in the current revisit. (b) A vehicle near the image boundary can still provide valid occupancy evidence when sufficient spatial support remains. (c) A state transition occurs during an unobserved interval, so its exact time is unknown. (d) Persistent occupancy spans a long and irregular interval between valid revisits.

The benchmark release will also include region observations with privacy protection, semantic region annotations, observation evidence labels, telemetry features, and local map representations needed to reproduce the reported experiments. Sensitive absolute geographic information will be removed or converted to local coordinate systems.