# WHEN ANOMALIES ARE TOO SMALL TO SEE: ORGANIZING EVIDENCE FOR VIDEO ANOMALY DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Video anomaly detection (VAD) aims to identify events that deviate from normal activity patterns specific to a scene. Given the open-set and scene-dependent nature of anomalies, recent work leverages semantic priors from pretrained vision-language models (VLMs) for detection without target-scene training. These methods have achieved strong performance in frame-level anomaly detection. However, they struggle with small-object anomalies in complex scenes, primarily because subtle anomaly cues from small objects are easily overwhelmed by the dominant normal background. This subproblem remains largely underexplored. To address this gap, we present SAVOR, a framework for detecting small-object anomalies from three complementary perspectives: temporal modeling, spatial refinement, and cross-frame association. Specifically, we propose Temporal Compression Mosaic (TCM) to adaptively select keyframes and highlight motion-salient objects. Then, Spatial Event Reasoning (SER) consolidates spatially sparse trajectory observations into persistent object events, enabling joint reasoning over complementary global and local anomaly evidence. Finally, to mitigate tracking errors, Anchor Prompt Retargeting (APR) realigns unreliable anchor prompts via mask feedback to guide bidirectional SAM2 propagation, preserving object identity and mask stability. The pipeline requires no in-domain training or parameter updates. Comprehensive experiments on ShanghaiTech and UBnormal demonstrate that our method outperforms existing zero-shot and training-free approaches at both pixel and object levels. Additional evaluations on SOAL, a newly constructed small-object anomaly benchmark curated from standard VAD datasets, confirm substantial improvements in spatial localization and temporal tracking.

## 1 INTRODUCTION

Video anomaly detection (VAD) aims to identify events that deviate from scene-specific activity patterns, with important applications in intelligent surveillance and public safety. Anomalous events are inherently open-set and scene-dependent: the definition of abnormality may vary substantially across scenes, while real-world anomalies cannot be exhaustively covered by finite training data.

Conventional VAD methods typically rely on normal samples or anomaly annotations from target scenes to learn scene-specific patterns, limiting their generalization to unseen scenes and anomalies. Recent advances in vision-language models (VLMs) and multimodal large language models (MLLMs) have driven zero-shot and training-free VAD, enabling anomaly detection with open-world knowledge without target-scene training or parameter updates, while achieving strong frame-level performance.

However, we observe that small-object anomalies in cluttered scenes remain a weakness of existing methods. As shown in Fig. 1, several representative methods exhibit reduced sensitivity to small or distant anomalies, leading to missed detections and inaccurate localization, despite achieving strong aggregate frame-level metrics. This issue is pronounced in VLM-based approaches relying heavily on global visual semantics (Zhou et al., 2024; Ahn et al., 2026), while also appearing across reconstruction- and prediction-based methods (Liu et al., 2021; Wang et al., 2022; Lee et al., 2026) and recent fine-grained approaches (Huang et al., 2025b).



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

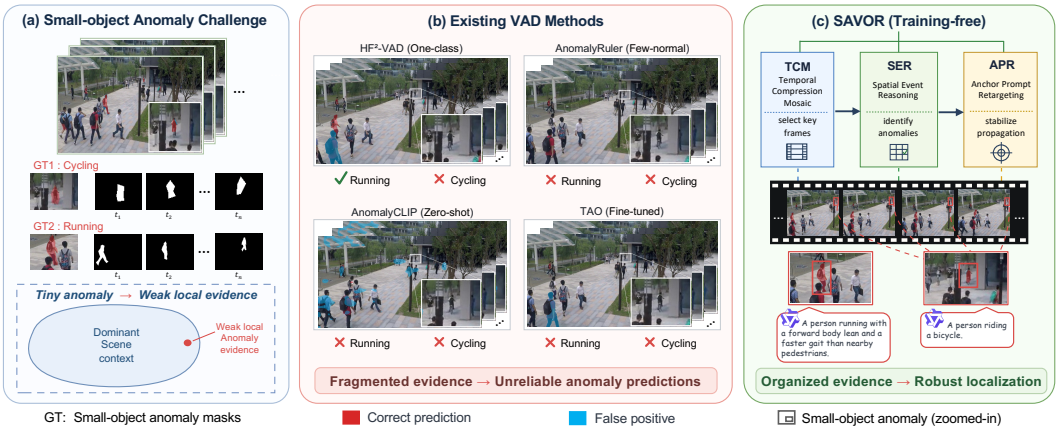



Figure 1: Qualitative visualization of anomaly localization results for ShanghaiTech test video 08_0058. HF$^2$-VAD (Liu et al., 2021) partially detects the salient nearby anomaly but misses the distant cyclist. AnomalyRuler (Yang et al., 2024) detects neither, while AnomalyCLIP (Zhou et al., 2024) and TAO (Huang et al., 2025b) generate inaccurate or fragmented predictions. Our training-free method accurately localizes anomalies with cleaner masks and generates anomaly descriptions

At the core of this challenge is the fragmentation of small-object anomaly evidence, which manifests along three interrelated dimensions: temporal observation, spatial representation, and cross-frame association. Temporally, anomaly cues concentrate in a few key moments, requiring effective observation allocation across long videos. Spatially, small-object evidence competes with cluttered backgrounds and dominant normal objects within full-frame observations. Across frames, scale variation, occlusion and imperfect initialization disrupt identity association and compound errors during propagation.

Existing studies have improved fine-grained VAD from several directions. Multi-scale and local modeling methods enhance detailed spatiotemporal representations (Zhang et al., 2024; Jiang, 2025; Wang et al., 2022; An et al., 2026), CLIP/VLM-based approaches improve region-level semantic perception through vision-language interaction (Wu et al., 2024b; Zhou et al., 2024; Zhu et al., 2026), and recent methods combine detection, segmentation, and tracking for pixel-level predictions (Huang et al., 2025b; Eltahir et al., 2026). While these approaches improve fine-grained anomaly understanding, they leave the fragmented evidence of small-object anomalies unresolved. Effectively integrating such evidence requires jointly considering temporal observations, spatial regions, and object continuity.

Building on this observation, we introduce **SAVOR** (Small-Anomaly Visual Evidence Organization and Reasoning), a training-free framework that uses motion consistency as the organizing bridge to link fragmented observations into persistent object-level events. It performs kinematic analysis to characterize motion dynamics and derive visual motion cues for VLMs, while global scene context and surrounding activity provide semantic references for anomaly assessment. Specifically, SAVOR comprises three components. Temporal Compression Mosaic (TCM) first selects motion-informative key frames from inter-frame dynamics and assembles them into a compact mosaic, reducing redundancy while preserving transient anomaly evidence. Spatial Event Reasoning (SER) then interprets candidate trajectories as persistent object events by aggregating fragmented cross-frame evidence, reasoning along complementary global and local pathways for anomaly assessment. To enhance propagation reliability, Anchor Prompt Retargeting (APR) uses mask feedback to correct unreliable anchors and stabilize bidirectional propagation, limiting error accumulation and maintaining consistent object identities across frames. Together, SAVOR transforms fragmented observations into object-level evidence for training-free anomaly detection and localization.

We conduct extensive experiments on multiple VAD benchmarks to evaluate SAVOR at pixel and object levels, and further examine cross-frame tracking and applicability across scenarios. Across these evaluations, SAVOR consistently outperforms representative zero-shot and training-free baselines. Notably, on our dedicated Small-Object Anomaly Localization (SOAL) benchmark derived

from ShanghaiTech and UBnormal, existing methods exhibit performance degradation, whereas SAVOR maintains stronger detection, localization, and cross-frame consistency.

Our main contributions are summarized as follows:

- We systematically investigate the underexplored problem of small-object anomalies in cluttered scenes, identifying three key challenges across temporal, spatial, and cross-frame dimensions.
- We propose SAVOR, a holistic training-free framework which organizes fragmented anomaly evidence into coherent object-level events for robust small-object anomaly detection and localization.
- We curate a small-object anomaly benchmark with frame-wise spatial and event-level category annotations. Extensive experiments demonstrate superior performance of SAVOR over existing training-free methods across diverse VAD benchmarks and anomaly scenarios.

## 2 Related Work

**Open-World and Vision-Language VAD** Conventional VAD methods employ weak supervision, frame reconstruction or prediction, and memory mechanisms for task-specific anomaly detection (Lv et al., 2023; Tang et al., 2020; Yang et al., 2023; Zhou et al., 2023). Open-set VAD extends detection to unseen anomalies through uncertainty-aware modeling (Zhu et al., 2022), while Open-Vocabulary VAD (Wu et al., 2024a) and Anomize (Li et al., 2025a) further incorporate pretrained semantic knowledge to detect and categorize unseen anomaly types. Recent work also leverages pretrained LLMs and VLMs for VAD. LAVAD converts VLM-generated frame descriptions into training-free LLM-based anomaly scores (Zanella et al., 2024), while Ex-VAD provides fine-grained visual-language explanations (Huang et al., 2025a). Holmes-VAU introduces anomaly-focused temporal sampling for long-term, multi-granular video understanding (Zhang et al., 2025a), and LAVIDA combines MLLM reasoning with pseudo-anomaly exposure and token compression for zero-shot VAD (Dai et al., 2026).

**Fine-Grained Anomaly Localization** VAD has been studied at multiple granularities, from frame-level anomaly detection to object- and pixel-level localization. Object-centric methods model localized video representations around detected instances (Ionescu et al., 2019; Micorek et al., 2024), while EVAL builds interpretable location-dependent scene models for spatial anomaly localization (Singh et al., 2023). Multi-Scale VAD learns coarse- and fine-grained spatio-temporal representations to improve sensitivity across anomaly scales, particularly for small-scale anomalies (Zhang et al., 2024). Recent work extends localization to dense masks: TAO uses anomaly boxes to prompt SAM2 for pixel-level tracking (Huang et al., 2025b), while GridVAD grounds open-set anomaly proposals and propagates their masks in a training-free pipeline (Eltahir et al., 2026).

**Video Object Segmentation and Propagation** Video object segmentation (VOS) propagates target masks across frames. XMem supports long-video segmentation through memory (Cheng & Schwing, 2022), while recent VOS methods strengthen target representation and propagation through object-level memory reading and spatial-semantic feature learning (Cheng et al., 2024; Li et al., 2025b). SAM 2 combines point, box, or mask prompts with streaming memory for video segmentation (Ravi et al., 2025), while SAM2Long adds training-free memory-tree search to reduce long-video error accumulation (Ding et al., 2025). In VAD, such propagation maintains localized anomalous targets across frames (Huang et al., 2025b; Eltahir et al., 2026).

Prior work has extended VAD toward semantic reasoning and spatial-temporal analysis, yet the fragmented evidence of small-object anomalies in cluttered scenes remains underexplored. Our training-free framework addresses this problem across temporal, spatial, and cross-frame dimensions.

## 3 Method

Given an untrimmed video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, our method produces object-specific spatiotemporal anomaly masks without target-scene training or parameter updates. We first link coherent moving regions into motion-supported candidate trajectories $\mathcal{C} = \{\mathcal{C}_i\}_{i=1}^{N}$, which provide a shared spatiotemporal index for evidence organization. **Temporal Compression Mosaic (TCM)** converts

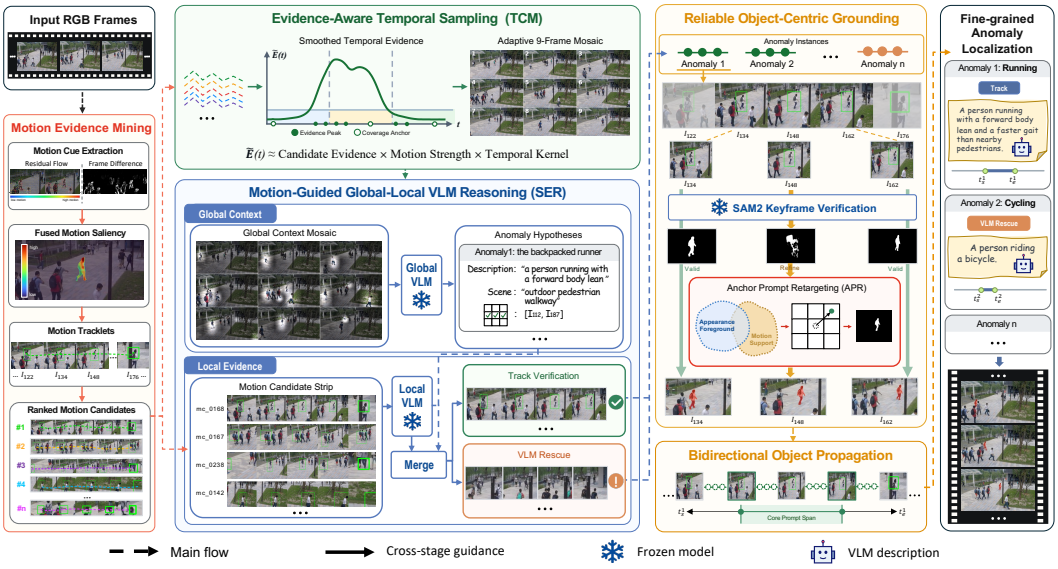



Figure 2: Overview of SAVOR. Motion-supported trajectories guide temporal evidence sampling. Global–local VLM reasoning then produces grounded anomaly instances. SAM2 feedback refines unreliable prompts before propagation produces fine-grained anomaly masks.

trajectory evidence into a compact global mosaic that balances informative moments and temporal coverage. **Spatial Event Reasoning (SER)** reorganizes the same trajectories into persistent local object-event views and combines global context with local evidence to recognize and ground anomalous instances. **Anchor Prompt Retargeting (APR)** then uses immediate SAM2 feedback to correct unreliable positive points before independent bidirectional propagation. Figure 2 summarizes this trajectory-guided global-local reasoning and segmentation pipeline.

### 3.1 Motion-Supported Trajectories

TCM and SER share category-free, motion-supported candidate trajectories that organize localized observations into temporally coherent object evidence. These trajectories provide a shared spatiotemporal representation for evidence organization, while anomaly semantics are deferred to subsequent reasoning. For each analyzed frame pair, we estimate dense optical flow and suppress dominant global translation using a robust componentwise median flow. Residual-flow and frame-difference cues identify localized motion components, which are linked across frames using predicted displacement together with spatial, geometric, and motion consistency. The resulting trajectory is

$$\mathcal{C}_i = \{(t_{i,k}, b_{i,k}, c_{i,k}, m_{i,k})\}_{k=1}^{K_i}, \tag{1}$$

where $t_{i,k}$ denotes the frame index, $b_{i,k}$ and $c_{i,k}$ denote the corresponding candidate box and center, $m_{i,k}$ denotes the motion magnitude, and $K_i$ is the number of observations in trajectory $\mathcal{C}_i$. Each trajectory is further characterized by motion- and track-quality cues that reflect temporal coherence and cross-frame correspondence. These cues quantify how consistently a candidate is supported over time and guide the organization of trajectory evidence.

Motion saliency alone is insufficient to infer abnormality, as salient motion may arise from normal activity while anomalous events may exhibit weak motion. We therefore decouple motion-based evidence organization from semantic anomaly reasoning. TCM uses trajectory evidence to select informative global observations, whereas SER uses identity-linked observations to construct persistent object events and infer anomaly semantics.

### 3.2 Temporal Compression Mosaic

Candidate trajectories provide temporally localized motion evidence, but constructing a compact global view requires preserving both salient events and broad temporal context. Uniform sampling

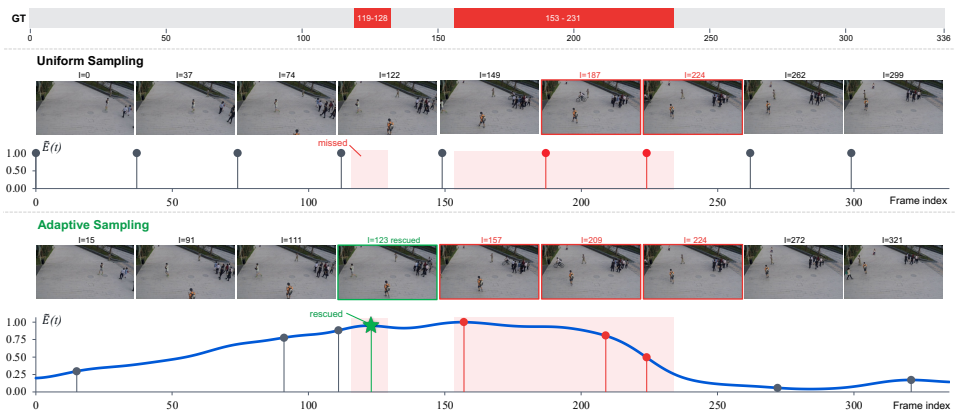



Figure 3: Uniform versus TCM adaptive temporal sampling with nine global observations. Uniform sampling misses the short anomaly interval at frames 119–128, while TCM selects frame 123 near its temporal-evidence peak and preserves broader temporal coverage.

may miss short-lived events, whereas evidence-only selection can overly concentrate its observations on the same event. TCM therefore allocates global observations by balancing event saliency and temporal coverage.

**Temporal Evidence Accumulation** For each trajectory $\mathcal{C}_i$, TCM aggregates fixed motion and trajectory-quality cues into an evidence weight, while robustly normalizing its keyframe motion magnitude over the video:

$$S_i = \mathrm{clip}_{[0,1]}\left(\sum_{\ell} w_\ell q_{i,\ell}\right), \qquad \widehat{m}_{i,k} = \mathrm{clip}_{[0,1]}\left(\frac{m_{i,k} - Q_{\mathrm{lo}}(\mathcal{M})}{Q_{\mathrm{hi}}(\mathcal{M}) - Q_{\mathrm{lo}}(\mathcal{M})}\right), \tag{2}$$

where $q_{i,\ell}$ denotes the $\ell$-th trajectory-level motion or quality cue, $w_\ell$ is its fixed coefficient, $\mathcal{M}$ contains the candidate-keyframe motion magnitudes, and $Q_{\mathrm{lo}}$ and $Q_{\mathrm{hi}}$ denote the corresponding lower and upper quantiles. The resulting $S_i$ captures trajectory-level evidence strength, while video-wide normalization preserves the relative motion saliency of $\widehat{m}_{i,k}$ across trajectories. TCM then accumulates these sparse trajectory observations over the full temporal axis:

$$E(t) = \sum_{i=1}^{N}\sum_{k=1}^{K_i} S_i \widehat{m}_{i,k} \exp\left(-\frac{|t - t_{i,k}|}{\tau}\right), \tag{3}$$

where $\tau > 0$ controls temporal decay. The accumulated evidence $E(t)$ measures the observation priority of frame $t$, assigning stronger evidence to frames near reliable motion-supported trajectories while accumulating support from nearby candidates. We apply local temporal smoothing as $\widetilde{E}(t) = |\mathcal{W}_t|^{-1}\sum_{u\in\mathcal{W}_t} E(u)$, which is subsequently balanced with temporal coverage for mosaic construction, as illustrated in Fig. 3.

**Evidence-Coverage Routing** To realize the evidence-coverage balance, TCM partitions the video into three macro temporal intervals. Within each interval $\mathcal{I}_r$, it selects two temporally separated high-evidence peaks and one coverage anchor. Let $e_r(t)$ denote the interval-normalized evidence and $d_r(t)$ the normalized distance to the nearest selected peak. Among the eligible frames $\mathcal{D}_r \subseteq \mathcal{I}_r$, the coverage anchor is selected as

$$a_r = \arg\max_{t\in\mathcal{D}_r} \left[\lambda e_r(t) + (1-\lambda) d_r(t)^\gamma\right]. \tag{4}$$

The evidence term favors stronger trajectory evidence, while the distance term promotes temporal separation from the selected peaks. Across the three intervals, TCM obtains nine chronologically ordered frames. The selected frames are then assembled into the mosaic with spatial emphasis on motion-supported candidate regions, locally brightening and sharpening them while retaining the surrounding scene as context. The resulting visual prompt directs the VLM toward motion-supported regions, while anomaly semantics are inferred from the full contextual view. Together, temporal routing and visual emphasis produce a compact global observation that preserves motion-relevant moments and scene context for SER.

### 3.3 SPATIAL EVENT REASONING

The TCM mosaic preserves scene context, whereas candidate trajectories preserve cross-frame object continuity. SER combines these complementary observations by reasoning globally over scene anomaly evidence and locally over persistent trajectory-centric object events. This separation allows anomaly semantics to be identified from visual context, while trajectory observations support object association and fine-grained local discovery.

**Persistent Object-Event Reasoning** A small candidate may recur across a video while occupying only a few pixels in each frame, leaving its visual evidence dispersed across the original scenes. SER uses the candidate trajectory as a temporal index to reorganize these sparse observations into a persistent object-event representation. For trajectory $\mathcal{C}_i$, we select representative observations distributed over the retained track. Each observation is cropped with surrounding context, enlarged to a common visual scale, and arranged in temporal order:

$$\mathcal{O}_i = \left[\psi(I_{t_{i,k}}, b_{i,k})\right]_{k \in \mathcal{K}_i^{\text{obs}}}, \tag{5}$$

where $\psi(\cdot)$ denotes a context-preserving crop with the tracked candidate marked. The resulting object-event view makes spatially sparse appearances persistently visible, allowing the VLM to inspect the same candidate and its temporal changes in a single representation.

Global reasoning analyzes the TCM mosaic to identify context-dependent anomaly hypotheses together with their supporting tile positions. Local reasoning independently examines each object event for anomalies revealed through persistent observation. A trajectory is associated with a global hypothesis only when its marked regions consistently correspond to the described abnormal target. If a locally identified anomaly is visible in the contextual crop but lies outside or only partly within the marked trajectory boxes, the trajectory is retained as spatiotemporal context and remains unassigned. SER then reconciles the global and local outputs at the instance level, merging duplicate evidence while preserving distinct discoveries. Each associated trajectory is verified in a focused view to ensure that its marked regions consistently follow the same abnormal instance before grounding.

**Evidence-Guided Instance Grounding** After semantic reconciliation and instance verification, SER grounds each recognized anomaly using its supporting spatial evidence. For an anomaly associated with a verified trajectory, the trajectory provides direct spatiotemporal evidence for grounding. Specifically, observations from a stable portion of the track are used to construct coarse spatial prompts, which are then interpolated over the corresponding temporal support. Semantic instances without an associated trajectory are grounded directly from the evidence supporting their anomaly semantics. The supporting tile positions of global hypotheses are mapped back to the corresponding source-video intervals to recover grounding observations. Local discoveries directly use their supporting frames and contextual crops as grounding observations. Conditioned on the anomaly description, a frozen VLM localizes the identified target in these observations with coarse bounding boxes and center points. Both grounding routes produce a unified anomaly-instance representation:

$$\mathcal{A}_j = \left(d_j, [t_j^s, t_j^e], \mathcal{P}_j\right), \qquad \mathcal{P}_j = \{(t, b_t, p_t) \mid t \in \mathcal{K}_j\}, \tag{6}$$

where $d_j$ denotes the anomaly description, $[t_j^s, t_j^e]$ its temporal support, and $\mathcal{P}_j$ the coarse spatial evidence over grounding frames $\mathcal{K}_j$, with bounding box $b_t$ and center point $p_t$ at frame $t$. This representation provides APR with the semantic identity, temporal support, and coarse spatial prompts required for subsequent refinement.

### 3.4 ANCHOR PROMPT RETARGETING

A coarse box center provides an initial anchor, but the induced SAM2 mask may not align with the intended object support. For each grounding keyframe with box $b$ and anchor $p_0$, APR queries the frozen SAM2 once to obtain an immediate mask $M_0$. We characterize the response by mask area, connected-component count $N_{\text{cc}}(M_0)$, and box containment ratio defined as

$$\rho(M_0, b) = \frac{|M_0 \cap b|}{\max(1, |M_0|)}. \tag{7}$$

These cues assess whether $M_0$ is nonempty, concentrated within $b$, and structurally coherent. APR retains $p_0$ when these structural checks are satisfied and $M_0$ adequately covers persistent target

evidence across neighboring keyframes. Otherwise, APR searches within the fixed SER box for an alternative anchor supported by complementary appearance, motion, and temporal evidence. Specifically, the candidate support region $R_t$ combines GrabCut (Rother et al., 2004) foreground, compatible trajectory motion when available, and foreground components persistent across neighboring keyframes.

$$\bar{p}_t = \arg\max_{x \in R_t} D_{R_t}(x), \qquad D_{R_t}(x) = \min_{y \notin R_t} \|x - y\|_2, \tag{8}$$

where $D_{R_t}(x)$ denotes the Euclidean distance from $x$ to the complement of $R_t$. This choice places $\bar{p}_t$ well inside the supported region and away from ambiguous object boundaries. APR evaluates at most one alternative anchor with one additional SAM2 query and adopts it only when the resulting mask satisfies the structural checks and remains consistent with the supported target evidence. APR retargets only the positive anchor, preserving the semantic identity, trajectory association, bounding box, and temporal support established by SER.

To keep mask propagation anchored to validated prompts across the full temporal span, APR propagates each anomaly instance independently in both temporal directions. For object $o$ with temporal support $[t_o^s, t_o^e]$, let $a_o^s$ and $a_o^e$ denote the first and last validated anchor frames within the stable prompt region, and let $\mathcal{P}_o$ denote the corresponding validated prompts. Denoting frozen SAM2 propagation by $\Phi(\cdot)$ and temporal reversal by $\mathcal{R}(\cdot)$, the two directional mask sequences are

$$\mathbf{M}^o_{\rightarrow} = \Phi\left(\mathcal{V}_{[a_o^s:t_o^e]}, \mathcal{P}_o\right), \qquad \mathbf{M}^o_{\leftarrow} = \mathcal{R}^{-1}\left(\Phi\left(\mathcal{R}\left(\mathcal{V}_{[t_o^s:a_o^e]}\right), \mathcal{R}(\mathcal{P}_o)\right)\right). \tag{9}$$

The backward branch thus reuses the same frozen propagation on the temporally reversed clip and maps the predictions back to the original timeline. The two directions are then fused at the object level as $M_t^o = M_{t,\rightarrow}^o \cup M_{t,\leftarrow}^o$ for $t \in [t_o^s, t_o^e]$. For videos containing multiple anomaly instances, the final frame-wise mask is $M_t = \bigcup_{o \in \mathcal{A}} M_t^o$, where $\mathcal{A}$ denotes the set of grounded anomaly instances.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets and Metrics** We evaluate standard VAD performance on UCSD Ped2 (Mahadevan et al., 2010) and ShanghaiTech (Luo et al., 2017). To specifically assess small-object anomaly localization, we curate SOAL, a 90-video benchmark from the ShanghaiTech and UBnormal (Acsintoae et al., 2022) test sets. Following prior VAD localization work, we report pixel-level AUROC, AP, AUPRO, and F1, together with region-level RBDC and track-level TBDC.

**Implementation** We use frozen Qwen3.6-Plus and SAM2.1 Hiera-B+ for VLM reasoning and mask propagation, respectively, without target-scene training or parameter updates. TCM constructs a $3 \times 3$ mosaic from nine frames, and the same VLM prompt templates are used across all datasets. All inference hyperparameters and decision thresholds are fixed across datasets. All experiments are conducted with a single NVIDIA RTX 3090 24GB GPU.

**Baselines** We compare SAVOR with representative methods spanning training-based, fine-tuned, zero-shot, and training-free settings: HF2VAD (Liu et al., 2021), DRAEM (Zavrtanik et al., 2021), DDAD (Mousakhan et al., 2023), SimpleNet (Liu et al., 2023), AnomalyRuler (Yang et al., 2024), TAO (Huang et al., 2025b), AdaCLIP (Cao et al., 2024), AnomalyCLIP (Zhou et al., 2024), AnyAnomaly (Ahn et al., 2026), and GridVAD (Eltahir et al., 2026). We report published results when directly comparable. Otherwise, we reproduce the methods using official code or released checkpoints under the same evaluation protocol.

## 4.2 Main Results

### 4.2.1 Standard Benchmark

Table 1 reports comparisons with representative VAD methods on ShanghaiTech and UCSD Ped2. On ShanghaiTech, SAVOR achieves the best results across all four pixel-level metrics, improving the best competing AP and AUPRO by 4.53 and 5.26 points, while remaining training-free. On Ped2, SAVOR achieves the highest AUROC and F1, surpassing the training-free GridVAD by 3.50 and 23.30 points. These results demonstrate strong localization across standard VAD benchmarks.

Table 1: Performance comparison on UCSD Ped2 and ShanghaiTech. Best and second-best are bolded and underlined, respectively. UBnormal results are reported in Appendix C.2.

| Method | Protocol | AUROC↑ | AP↑ | AUPRO↑ | F1↑ | RBDC↑ | TBDC↑ |
|---|---|---|---|---|---|---|---|
| *UCSD Ped2* | | | | | | | |
| DRAEM | Tr | 69.58 | 30.63 | 35.78 | 10.89 | 44.26 | 70.64 |
| DDAD | Tr | 55.87 | 5.61 | 15.12 | 2.67 | 18.01 | 13.29 |
| SimpleNet | Tr | 52.49 | 20.51 | 44.05 | 10.71 | 51.18 | 27.75 |
| AdaCLIP | FT | 53.06 | 4.97 | 50.66 | 11.19 | 12.30 | 15.50 |
| AnomalyCLIP | FT | 54.25 | 23.73 | 38.59 | 7.48 | 13.10 | 21.00 |
| AnomalyRuler† | PFT | 63.29 | 16.30 | 20.69 | 36.50 | 38.13 | 50.86 |
| TAO | PFT | 75.11 | **50.78** | **72.97** | 64.12 | **83.60** | **93.20** |
| AdaCLIP | ZS | 51.02 | 1.32 | 33.98 | 2.61 | 5.80 | 10.60 |
| AnomalyCLIP | ZS | 51.63 | 21.20 | 36.34 | 5.92 | 7.50 | 11.20 |
| GridVAD | TF | 77.59 | 38.53 | 66.82 | 42.09 | 38.96 | 37.70 |
| **Ours** | TF | **81.09** | 48.44 | 60.29 | **65.39** | 71.36 | 80.98 |
| *ShanghaiTech* | | | | | | | |
| HF2VAD† | Tr | 73.01 | 34.75 | 41.46 | 52.81 | 45.40 | 84.50 |
| AnomalyRuler† | PFT | 73.03 | 41.12 | 40.79 | 55.37 | 51.39 | 72.65 |
| TAO† | PFT | 67.63 | 22.08 | 33.46 | 34.07 | **62.10** | **85.40** |
| AdaCLIP† | ZS | 72.87 | 10.95 | 46.04 | 31.00 | 10.20 | 31.19 |
| AnomalyCLIP† | ZS | 72.71 | 5.07 | 37.19 | 18.68 | 5.14 | 19.61 |
| **Ours** | TF | **76.37** | **45.65** | **51.30** | **58.88** | 57.31 | 75.46 |

Tr = trained, FT = full fine-tuning, PFT = partial fine-tuning, ZS = zero-shot, TF = training-free.
† denotes reproduced results by us. Some baseline results are taken from TAO (Huang et al., 2025b).

Table 2: Small-object anomaly localization on SOAL and its ShanghaiTech subset. Object size is measured by normalized ground-truth box area $wh/(WH)$.

| Method | Protocol | Overall | | | | | | Object Size (Macro Pixel-AP) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | AP | AUPRO | F1 | RBDC | TBDC | ≤0.25% | 0.25–0.5% | 0.5–1% | 1–2% | >2% |
| *Full collection: 90 videos* | | | | | | | | | | | | |
| AdaCLIP | ZS | 81.31 | 16.30 | 57.29 | 24.10 | 18.48 | 43.50 | 2.54 | 4.26 | 5.70 | 7.37 | 7.30 |
| AnomalyCLIP | ZS | 76.81 | 4.53 | 47.56 | 12.59 | 8.37 | 17.86 | 2.29 | 3.30 | 2.29 | 1.98 | 4.03 |
| AnyAnomaly | ZS | 68.63 | 2.12 | 21.61 | 6.41 | 3.09 | 11.15 | 0.04 | 0.07 | 0.34 | 0.91 | 3.48 |
| **Ours** | TF | **82.94** | **55.25** | **64.37** | **69.92** | **56.08** | **71.17** | **20.74** | **25.24** | **28.91** | **29.90** | **31.74** |
| *ShanghaiTech subset: 43 videos* | | | | | | | | | | | | |
| HF2VAD | Tr | 76.92 | 17.34 | 51.09 | 35.25 | 2.19 | 6.43 | 0.19 | 0.72 | 1.73 | 3.52 | 7.12 |
| AnomalyRuler | PFT | 69.21 | 33.14 | 35.78 | 47.90 | **50.62** | **71.44** | 1.41 | 5.23 | 10.81 | 20.91 | 24.80 |
| AdaCLIP | ZS | 70.54 | 20.25 | 36.65 | 28.54 | 20.77 | 56.14 | 0.09 | 0.31 | 1.12 | 7.11 | 8.03 |
| AnomalyCLIP | ZS | 65.15 | 4.13 | 29.52 | 14.21 | 8.12 | 25.79 | 0.12 | 0.23 | 0.75 | 1.44 | 3.32 |
| AnyAnomaly | ZS | 73.31 | 6.40 | 30.04 | 12.14 | 7.19 | 18.97 | 0.01 | 0.05 | 0.24 | 1.24 | 4.06 |
| **Ours** | TF | **78.87** | **46.83** | **55.36** | **63.96** | 41.32 | 59.66 | **6.82** | **15.75** | **23.65** | **27.72** | **31.90** |

While SAVOR trails the best RBDC and TBDC results, prior works note limitations of the standard RBDC/TBDC protocol (Yang & Radke, 2025; Ramachandra & Jones, 2020), whose permissive matching allows limited coverage to count as valid detections. Consequently, over-extended responses may overstate RBDC/TBDC without exposing false discoveries. We therefore additionally analyze O-AP, O-Recall, and O-FDR to assess prediction quality (see Appendix C.1 for details). SAVOR retains over 80% O-Recall on both datasets while achieving the highest O-AP and lowest O-FDR, whereas the higher-coverage methods yield substantially more false discoveries. This indicates that SAVOR maintains broad object coverage with substantially fewer false discoveries, consistent with its instance verification and evidence-guided grounding.

#### 4.2.2 Small-Object Anomaly Localization

SOAL contains 34,318 anomalous-object observations with a median normalized area of 0.38%, and 90% occupy no more than 2% of the frame, placing the evaluation predominantly in the small-object regime. To broaden the comparison to learning-based baselines, we also report results on the 43-video ShanghaiTech subset under the same localization protocol. Dataset construction and evaluation protocols are provided in Appendix A.

Table 2 shows that SAVOR achieves 55.25 AP on SOAL, surpassing the best baseline by 38.95 points. At the two smallest ranges, all baselines remain below 5% Macro Pixel-AP, while SAVOR reaches 20.74 and 25.24. It further achieves 56.08 RBDC and 71.17 TBDC, extending the gain

Table 3: Component ablation on ShanghaiTech. All variants use the same nine-frame global budget.

| Temporal sampling | TCM | SER | APR | Pixel-level | | | |
|---|---|---|---|---|---|---|---|
| | | | | AUROC ↑ | AP ↑ | AUPRO ↑ | F1 ↑ |
| Uniform 9 frames | - | - | - | 68.79 | 35.84 | 47.91 | 43.91 |
| Adaptive 9 frames | ✓ | - | - | 70.14 | 37.65 | 50.79 | 46.46 |
| Adaptive 9 frames | ✓ | ✓ | - | 76.23 | 43.67 | 51.07 | 57.61 |
| Adaptive 9 frames | ✓ | - | ✓ | 70.32 | 38.31 | 51.09 | 47.04 |
| Adaptive 9 frames | ✓ | ✓ | ✓ | **76.37** | **45.65** | **51.30** | **58.88** |

to region- and track-level localization. On the ShanghaiTech subset, SAVOR achieves the highest AP and AUPRO, with the best results across all five object-size ranges, highlighting its strength in fine-grained spatial localization. We further examine whether the localization gains extend across anomaly types. SAVOR achieves the highest Macro Pixel-AP in 15 of 19 anomaly types on SOAL and four of six on the ShanghaiTech subset (see Appendix A.3 for category-wise results). These results demonstrate that the localization gains span diverse anomaly semantics.

### 4.3 Ablation Studies

**Temporal Routing** Table 3 shows that TCM improves AP by 1.81 points under a nine-frame budget. Table 4 characterizes routing using two post-hoc statistics, the fraction of selected frames within anomaly intervals and the fraction of intervals reached by at least one selected frame. Evidence-focused routing raises the former from 44.44% to 57.53% but reduces coverage from 77.32% to 69.59%, whereas TCM maintains 57.63% while increasing coverage to 79.90%, supporting joint consideration of both factors. Further analysis appears in Appendix B.1.

Table 4: Temporal routing on ShanghaiTech. Ground-truth intervals are used only for analysis.

| Routing | Anomalous frames (%) | Interval coverage (%) |
|---|---|---|
| Uniform | 44.44 | 77.32 |
| Evidence-focused | 57.53 | 69.59 |
| **TCM routing** | **57.63** | **79.90** |

**Global-Local Evidence** Among components, SER yields the largest improvement, reaching 43.67 AP and 57.61 F1. Table 5 further evaluates the complementary roles of global scene cues and local tracks in SER. Full SER achieves 59.26% event recall and 45.07% target recall, outperforming the best global-only and local-only settings. Among targets visible in the global input frames, Full SER further improves target recall from 51.61% to 74.19%, showing that coupling global anomaly semantics with persistent local tracks improves trajectory-level grounding. Additional analysis is provided in Appendix B.2.

Table 5: Comparison of global-local anomaly grounding on ShanghaiTech.

| Evidence | Event Rec. | Target Rec. | FP/Video |
|---|---|---|---|
| Global (3×3) | 31.48 | 22.54 | 0.53 |
| Global (9 images) | 27.78 | 19.72 | 0.58 |
| Local Tracks | 22.22 | 15.49 | **0.00** |
| **Full SER** | **59.26** | **45.07** | 0.19 |

**Anchor Retargeting** APR improves AP with and without SER in Table 3, with the full model reaching 45.65 AP. Table 6 compares immediate SAM2 mask quality before and after anchor retargeting across all retargeting attempts and among those accepted by APR. Across all retargeting attempts, immediate-mask IoU/F1 improve by 7.77/8.25 points. For retargetings accepted by APR, the gains increase to 25.20/27.21 points, indicating that APR targets unreliable initial anchors.

Table 6: Immediate-mask quality before and after APR on ShanghaiTech.

| Subset | Metric | Original | Retargeted | Δ |
|---|---|---|---|---|
| All attempted | IoU | 36.77 | 44.54 | **+7.77** |
| | F1 | 48.23 | 56.47 | **+8.25** |
| APR-accepted | IoU | 25.49 | 50.68 | **+25.20** |
| | F1 | 34.19 | 61.41 | **+27.21** |

## 5 Conclusion

In this paper, we address the underexplored problem of small-object anomaly detection and propose SAVOR, a training-free framework that organizes fragmented anomaly evidence across temporal, spatial, and cross-frame dimensions. TCM selects informative views while preserving temporal coverage, SER organizes sparse observations as persistent object events with global scene context, and APR retargets unreliable anchors before mask propagation. We further construct SOAL to support scale- and category-stratified evaluation of small-object anomalies. Extensive experiments on SOAL and standard VAD benchmarks demonstrate that SAVOR substantially improves fine-grained anomaly localization without target-scene training or parameter updates.

## AI USE STATEMENT

Large language models (LLMs) were used to assist with manuscript preparation and literature search, including improving the clarity and readability of author-written text and identifying relevant prior work. They were not used for research ideation, methodology, experimental design, reported results, or scientific conclusions. Separately, pretrained vision-language models (VLMs) are used as frozen components of the proposed method, as described in the main paper. All LLM-assisted text was reviewed and edited by the authors, who take full responsibility for the work.

## REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20143–20153, 2022.

Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. AnyAnomaly: Zero-shot customizable video anomaly detection with LVLM. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3026–3035, 2026.

Dezhi An, Wenqiang Liu, Jun Lu, and Shengcai Zhang. Fine-grained video anomaly detection via adaptive feature refinement and semantic enrichment. *The Visual Computer*, 42, 2026. doi: 10.1007/s00371-025-04331-8.

Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, 2023. doi: 10.1016/j.cviu.2023.103656.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 813–824, 2021.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, pp. 640–658, 2022. doi: 10.1007/978-3-031-19815-1_37.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161, 2024.

Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: MLLM empowered zero-shot video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35648–35658, 2026.

Anja Delić, Matej Grcic, and Siniša Šegvić. Sequential keypoint density estimator: An overlooked baseline of skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11579–11589, 2025.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. SAM2Long: Enhancing SAM 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13614–13624, 2025. doi: 10.1109/ICCV51701.2025.01264.

Mohamed Eltahir, Ahmed O. Ibrahim, Obada Siralkhatim, Tabarak Abdallah, and Sondos Mohamed. GridVAD: Open-set video anomaly detection via spatial reasoning over stratified frame grids. *arXiv preprint arXiv:2603.25467*, 2026.

Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4505–4523, 2022. doi: 10.1109/TPAMI.2021.3074805.

Jihao Gu, Kun Li, He Wang, and Kaan Akşit. Text-guided fine-grained video anomaly understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1251–1261, 2026.

Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13545–13554, 2023.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-VAD: Explainable fine-grained video anomaly detection based on visual-language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 25750–25761, 2025a.

Yuzhi Huang, Chenxin Li, Haitao Zhang, Zixu Lin, Yunlong Lin, Hengyu Liu, Wuyang Li, Xinyu Liu, Jiechao Gao, Yue Huang, Xinghao Ding, and Yixuan Yuan. Track any anomalous object: A granular video anomaly detection pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8689–8699, 2025b.

Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7842–7851, 2019.

Yalong Jiang. Local patterns generalize better for novel anomalies. In *International Conference on Learning Representations (ICLR)*, 2025.

Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, Yoichi Sato, and Norimasa Kobori. WTS: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In *European Conference on Computer Vision (ECCV)*, 2024.

Kijung Lee, Youngwan Jo, Sunghyun Ahn, and Sanghyun Park. Unified video anomaly detection model for detecting different anomaly types. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 784–794, 2026. doi: 10.1109/WACV61042.2026.00083.

Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29203–29212, 2025a.

Xin Li, Deshui Miao, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Learning spatial-semantic features for robust video object segmentation. In *International Conference on Learning Representations (ICLR)*, 2025b.

Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, pp. 29713–29744, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.

Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13588–13597, 2021.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20402–20411, 2023.

Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8022–8031, 2023.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1975–1981, 2010.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. MULDE: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18868–18877, 2024.

Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2569–2578, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Nicolae-Catalin Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, and Mubarak Shah. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15984–15995, 2024.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004.

Ashish Singh, Michael J. Jones, and Erik G. Learned-Miller. EVAL: Explainable video anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18717–18726, 2023.

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

Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020.

Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2301–2312, 2022. doi: 10.1109/TNNLS.2021.3083152.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024b. doi: 10.1609/aaai.v38i6.28423.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision (ECCV)*, 2024.

Zhengye Yang and Richard J. Radke. Detecting contextual anomalies by discovering consistent spatial regions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 1083–1092, 2025.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14592–14601, 2023.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. PANDA: Towards generalist video anomaly detection via agentic AI engineer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, pp. 92745–92774, 2025.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536, 2024.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8330–8339, 2021.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853, 2025a.

Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17385–17394, 2024.

Ruixuan Zhang, Beichen Wang, Juexiao Zhang, Zilin Bian, Chen Feng, and Kaan Ozbay. When language and vision meet road safety: Leveraging multimodal large language models for video-based traffic accident analysis. *Accident Analysis & Prevention*, 219:108077, 2025b.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3769–3777, 2023. doi: 10.1609/aaai.v37i3.25489.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024.

Yiyan Zhu, Menghao Zhang, Haifeng Sun, Pengfei Ren, Xianao Chu, Chenye Xu, Hong Tan, Jinghan Wang, Qi Qi, and Jingyu Wang. Alert-CLIP: Abnormality-aware latent-enhanced representation tuning of CLIP for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35545–35554, 2026.

Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards open set video anomaly detection. In *European Conference on Computer Vision (ECCV)*, pp. 395–412, 2022. doi: 10.1007/978-3-031-19830-4_23.

# A SOAL Dataset Details

## A.1 Construction and Annotation

Our small-object anomaly evaluation benchmark, SOAL, is constructed from the test splits of ShanghaiTech and UBnormal using common target-scale and scene-complexity screening criteria. It contains 90 complete videos, including 43 from ShanghaiTech and 47 base videos from UBnormal. SOAL preserves the full temporal extent of each selected video together with surrounding normal activity and scene clutter, providing a unified setting for evaluating small-object anomaly localization across object scales and anomaly semantics.

**Video Selection** The ShanghaiTech subset covers source scenes 02, 04, 05, 08, 09, and 12, while the UBnormal subset spans 19 test scenes. Each selected video contains at least 50 valid anomalous-object observations and 30 anomalous frames. For UBnormal scenarios with multiple visual variants, such as fog, smoke, or fire, the base video is used as the representative instance in SOAL.

**Spatial Annotation** SOAL harmonizes the spatial annotations of the two source benchmarks into a common scale-aware representation. For ShanghaiTech, frame-wise anomalous components are extracted from the official pixel-level masks, with each connected component represented by its bounding-box extent for scale characterization and the original masks retained for pixel-level evaluation. For UBnormal, the official anomaly tracks and their frame-wise spatial annotations provide object-level spatial support. All spatial observations are represented under the same normalized-area convention, enabling consistent scale characterization and scale-stratified localization evaluation across both source benchmarks.

**Semantic Annotation** SOAL further introduces a unified event-level semantic annotation layer across ShanghaiTech and UBnormal. The resulting annotation contains 140 semantic events organized into 19 unified anomaly types across all 90 videos. ShanghaiTech events are aligned with their temporal anomaly regions and frame-wise spatial components, while UBnormal events are associated with the corresponding official anomaly tracks. This event-level organization supports videos containing multiple anomaly types and allows each semantic event to be associated with multiple spatial tracks or components when necessary. The unified semantic layer enables the same class-agnostic localization predictions to be evaluated consistently across anomaly types. Together with the harmonized scale-aware representation, SOAL supports overall, scale-stratified, and category-stratified evaluation of small-object anomaly localization.

## A.2 Target Scale and Scene Complexity

We characterize target scale using the common scale-aware representation introduced above. For anomalous object $i$ in frame $t$, its normalized bounding-box area is defined as

$$a_{i,t}^{\mathrm{box}} = \frac{w_{i,t} h_{i,t}}{W_t H_t}, \tag{10}$$

where $(w_{i,t}, h_{i,t})$ denotes the spatial extent of the anomalous component and $(W_t, H_t)$ is the frame resolution. This normalization makes target scale comparable across videos and source datasets with different image resolutions. We use $a^{\mathrm{box}} \leq 2\%$ as the reference range for small-object observations. Across SOAL, the 34,318 anomalous-object observations have a median normalized area of 0.378%, with 90.06% occupying no more than 2% of the frame, demonstrating the strong concentration of SOAL in the small-object regime.

Scene complexity is further assessed along two complementary dimensions that can obscure small anomaly cues. Edge density measures static visual clutter by capturing structural and high-frequency content, while a frame-difference-based motion proxy measures dynamic scene activity. Together, these statistics characterize both visually cluttered and temporally active scenes, reflecting the diverse contexts in which small anomalous targets must be localized. Table 7 summarizes the composition, target-scale, and scene-complexity statistics of SOAL.

The two source subsets provide complementary target-scale distributions within SOAL. ShanghaiTech contributes 9,091 anomalous-object observations with a median normalized area of 1.10%,

Table 7: Composition, target-scale, and scene-complexity statistics of SOAL. Small-object denotes the percentage of anomalous-object observations with $a^{\text{box}} \leq 2\%$. Scene-complexity columns report equal-weight means of per-video statistics. Object obs. denotes anomalous-object observations.

| Source | Composition | | | | Target Scale | | Scene Complexity | |
|---|---|---|---|---|---|---|---|---|
| | Videos | Frames | Anomaly frames | Object obs. | Median area (%) | Small-object (%) | Edge density (%) | Motion proxy (%) |
| ShanghaiTech | 43 | 17,419 | 7,962 | 9,091 | 1.10 | 72.51 | 5.54 | 0.39 |
| UBnormal | 47 | 20,672 | 14,581 | 25,227 | 0.28 | 96.38 | 11.62 | 0.19 |
| **SOAL** | **90** | **38,091** | **22,543** | **34,318** | **0.38** | **90.06** | **8.71** | **0.29** |

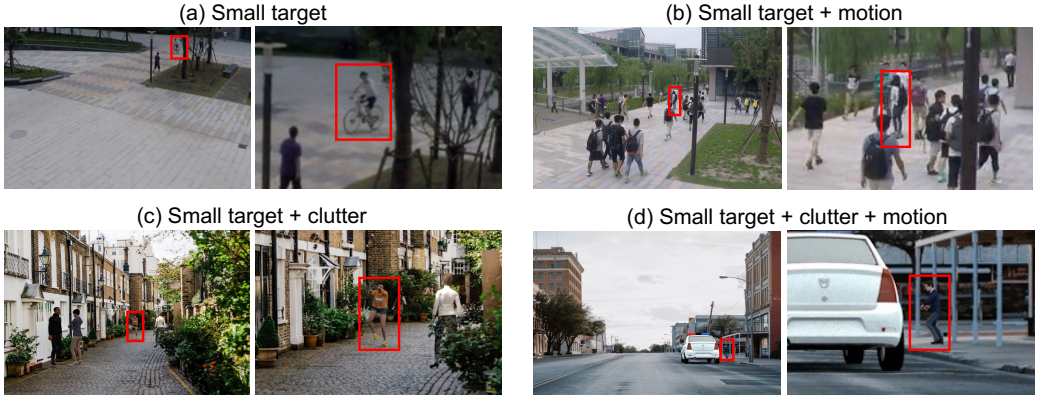



Figure 4: Representative challenge conditions in SOAL. Examples from ShanghaiTech in (a,b) and UBnormal in (c,d) illustrate small anomalous targets under complementary conditions, including strong motion, visual clutter, and their combination. Each example shows the full frame and an enlarged target crop, with red boxes indicating the ground-truth anomaly.

of which 72.51% occupy no more than 2% of the frame. UBnormal further concentrates the benchmark toward smaller targets, with a median area of 0.28% and 96.38% of its 25,227 observations falling within the same range. Together, the two sources broaden the coverage of small-object anomaly localization across diverse scenes and target scales. Scene 08 provides a representative example in which small target scale coincides with complex scene context. Its median anomalous-object area is only 0.63% of the frame, and 90.4% of its spatial observations fall below the 2% reference. Four of the five videos with the highest motion activity in SOAL originate from Scene 08, and all four also exhibit edge density above the ShanghaiTech median. This combination of small targets, visual clutter, and strong temporal activity illustrates the joint challenges captured by SOAL. Representative examples from ShanghaiTech and UBnormal are shown in Fig. 4.

### A.3 Scale and Category Evaluation

**Scale-Wise Evaluation** For scale-stratified evaluation, anomalous objects are grouped according to their normalized ground-truth bounding-box area $a^{\text{box}}$. We use five fixed ranges, $\leq 0.25\%$, $(0.25, 0.5]\%$, $(0.5, 1]\%$, $(1, 2]\%$, and $> 2\%$. When evaluating a scale range, pixels belonging to anomalous objects within the target scale are treated as positives, anomalous pixels from other scale ranges are ignored, and normal background pixels remain negative. A frame containing anomalous objects at multiple scales may therefore contribute to multiple strata. We report Macro Pixel-AP for scale-stratified localization, with results provided in Table 2 of the main paper.

**Category-Wise Evaluation** We further stratify the class-agnostic localization predictions by ground-truth semantic anomaly type to determine whether the aggregate localization gain extends across diverse anomaly types. For a given target type, its anomalous pixels are treated as positives, anomalous pixels belonging to other types are ignored, and normal background pixels remain negative. Category labels are used only for evaluation grouping, while the evaluated methods retain class-agnostic predictions. For frames in which multiple semantic events share an inseparable binary ground-truth region, the ambiguous region is excluded from category-specific spatial evaluation.

Table 8: Category-wise anomaly localization on SOAL. Panel (a) reports Macro Pixel-AP for all 19 semantic anomaly types in the full 90-video collection. Panel (b) reports the six types in the 43-video ShanghaiTech subset. $n$ denotes the number of supporting videos. The best result is marked in bold and the runner-up is underlined.

*(a) Full SOAL, 90 videos*

| Anomaly type | $n$ | AdaCLIP | AnomalyCLIP | AnyAnomaly | Ours |
|---|---|---|---|---|---|
| Running | 39 | 6.31 | 4.49 | 1.88 | **46.90** |
| Laying down | 14 | 21.74 | 18.06 | 0.59 | **49.88** |
| Bicycle | 11 | 13.48 | 1.47 | 4.48 | **41.85** |
| Jumping | 11 | 17.14 | 3.96 | 0.73 | **36.84** |
| Shuffling | 11 | 7.61 | 1.78 | 0.29 | **35.92** |
| Seizure | 7 | 16.69 | 15.40 | 0.41 | **44.16** |
| Skateboarding | 5 | 2.35 | 3.51 | 1.93 | **43.33** |
| Walking drunk | 5 | 8.93 | 9.42 | 0.33 | **46.06** |
| Stroller | 3 | 16.33 | 8.01 | 1.67 | **34.64** |
| Throwing | 3 | 11.45 | 4.67 | 0.93 | **33.38** |
| Car | 2 | 4.79 | 6.29 | 28.73 | **67.02** |
| Falling | 2 | **25.41** | 5.94 | 1.26 | 8.34 |
| Loitering | 2 | 18.60 | 11.27 | 0.73 | **23.53** |
| Pickup | 2 | **10.27** | 2.30 | 0.21 | 9.14 |
| Fighting | 1 | **26.60** | 8.01 | 13.94 | 1.27 |
| Jaywalking | 1 | 0.18 | 0.40 | 0.22 | **68.78** |
| Motorcycle | 1 | 15.38 | 2.18 | 2.89 | **52.55** |
| People–car accident | 1 | 5.81 | **25.44** | 0.55 | 20.53 |
| Unicycle riding | 1 | 2.45 | 3.62 | 0.31 | **47.96** |

*(b) ShanghaiTech subset, 43 videos*

| Anomaly type | $n$ | HF2VAD | AnomalyRuler | AdaCLIP | AnomalyCLIP | AnyAnomaly | Ours |
|---|---|---|---|---|---|---|---|
| Running | 13 | 4.94 | 10.82 | 4.70 | 3.40 | 2.03 | **39.59** |
| Bicycle | 11 | 11.10 | 26.26 | 13.48 | 1.47 | 4.48 | **41.85** |
| Jumping | 5 | 6.53 | **22.42** | 15.20 | 5.86 | 1.44 | 9.89 |
| Skateboarding | 5 | 8.59 | 34.39 | 2.35 | 3.51 | 1.93 | **43.33** |
| Stroller | 3 | 15.00 | **46.17** | 16.33 | 8.01 | 1.67 | 34.64 |
| Throwing | 3 | 8.66 | 3.44 | 11.45 | 4.67 | 0.93 | **33.38** |

SOAL contains 19 unified semantic anomaly types. Table 8 reports Macro Pixel-AP across all types, with $n$ denoting the number of supporting videos. On the full SOAL collection, SAVOR achieves the highest Macro Pixel-AP in 15 of the 19 anomaly types. The gains span categories with different semantic content and support sizes, indicating that the overall localization improvement extends beyond a small set of event types. On the ShanghaiTech subset, SAVOR achieves the highest Macro Pixel-AP in four of the six semantic types, further demonstrating consistent localization gains across anomaly categories.

# B Additional Ablations

## B.1 TCM Temporal Routing

**Uniform Sampling** Under the fixed nine-frame observation budget, uniform sampling spreads observations evenly over time without accounting for where informative anomaly evidence occurs. We therefore examine the anomaly evidence retained by uniform sampling at both the dataset and video levels. The anomaly-ratio columns in Table 9 show similar aggregate anomaly ratios for all frames and the nine sampled observations, with 44.72% versus 44.44% on ShanghaiTech and 83.01% versus 81.48% on Ped2. However, this aggregate similarity masks substantial per-video sparsity. The per-video distribution in Table 9 shows that 68 of the 107 ShanghaiTech videos contain at most four anomaly-bearing observations among the nine sampled frames, compared with only one of the 12 Ped2 videos. Uniform sampling can therefore leave sparse anomaly evidence in individual videos despite comparable dataset-level ratios, motivating adaptive temporal routing under the same observation budget.

**TCM Routing** Adaptive temporal routing must account for both where informative evidence occurs and how observations are distributed over time. Table 10 extends the routing analysis in Table 4 with two additional routing principles under the same nine-frame observation budget. These routing principles emphasize different aspects of the design. Uniform routing distributes observations without evidence awareness, evidence-focused routing prioritizes informative moments, segmented

Table 9: Anomaly evidence under uniform nine-frame sampling. $k$ is the number of anomaly-bearing observations among the nine sampled frames, and entries report the corresponding percentage of videos.

| Dataset | Videos | Anomaly ratio (%) | | Proportion (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | All frames | Uniform-9 | $k=0$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k\geq 5$ |
| ShanghaiTech | 107 | 44.72 | 44.44 | 0 | 7.48 | 22.43 | 15.89 | 17.76 | 36.45 |
| Ped2 | 12 | 83.01 | 81.48 | 0 | 0 | 0 | 0 | 8.33 | 91.67 |

Table 10: Temporal routing analysis on ShanghaiTech under a fixed nine-frame observation budget.

| Routing emphasis | Anomalous frames (%) | Interval coverage (%) |
|---|---|---|
| Uniform | 44.44 | 77.32 |
| Evidence-focused | 57.53 | 69.59 |
| Segmented coverage | 55.24 | 78.35 |
| **TCM routing** | 57.63 | 79.90 |
| Dynamic allocation | 64.59 | 74.74 |

coverage broadens temporal allocation, and dynamic allocation allows more flexible concentration of observations. TCM combines evidence-aware selection with distance-aware coverage, directly addressing both requirements within a single routing rule.

Anomaly-bearing observations measure the fraction of selected frames within annotated anomaly intervals, while interval coverage measures the fraction of the 194 anomaly intervals reached by at least one frame. Ground-truth intervals are used only to compute these post-hoc statistics and are not used by the routing procedure. The comparison reveals the complementary roles of informative evidence and temporal coverage. A stronger concentration of anomaly-bearing observations does not necessarily yield broader coverage, as the highest anomaly-bearing ratio is accompanied by lower interval coverage than TCM routing. TCM maintains anomaly concentration while achieving the broadest interval coverage, supporting the joint consideration of both factors.

### B.2 SER Analysis

We analyze SER on 43 ShanghaiTech test videos in SOAL with fine-grained annotations covering 54 anomalous events and 71 persistent anomalous targets. The annotations provide event-level correspondence and frame-level target localization, enabling evaluation at both the event and trajectory levels. All analyses use the same frozen trajectory candidate set and evaluation protocol.

**Semantic Guidance** Table 11 isolates the contribution of global anomaly semantics to trajectory grounding. Frame-wise Voting independently evaluates the five observations of each candidate and retains a trajectory by majority voting over the five VLM predictions. The two SER variants instead use the same persistent trajectory evidence in a single VLM request and share identical association settings. The w/o Semantic Guidance variant preserves the hypothesis structure while replacing each global anomaly label and description with the fixed texts *unspecified anomalous event* and *An unspecified anomalous event involving a scene entity*.

Specific anomaly semantics increase event recall from 24.07% to 59.26% and target recall from 18.31% to 45.07%. On candidate-supported cases, Full SER reaches 82.05% event recall and 74.42% target recall, compared with 33.33% and 30.23% without semantic guidance. Full SER further outperforms the frame-wise reference while using one VLM request instead of five under this protocol. These results show that specific global anomaly semantics guide the association of anomalous events with persistent local trajectories.

**Global-Local Evidence** We further analyze the global-local comparison in Table 5. The two global settings use the same nine TCM-selected scene frames, presented either as a single $3\times3$ mosaic or as nine separate images. The local setting uses only persistent trajectory evidence without

Table 11: Semantic guidance for trajectory grounding on ShanghaiTech. Candidate-supported denotes targets covered by the frozen trajectory candidate set and their corresponding events.

| Setting | VLM Calls | Overall | | Candidate-supported | | FP/Video |
|---|---|---|---|---|---|---|
| | | Event Rec. | Target Rec. | Event Rec. | Target Rec. | |
| Frame-wise Voting | 5 | 42.59 | 32.39 | 58.97 | 53.49 | 0.21 |
| w/o Semantic Guidance | 1 | 24.07 | 18.31 | 33.33 | 30.23 | **0.05** |
| **Full SER** | 1 | **59.26** | **45.07** | **82.05** | **74.42** | 0.19 |

global anomaly semantics, while Full SER associates global anomaly hypotheses with the persistent local trajectories.

Conditioning on global visibility, we evaluate the 31 targets that appear in at least one TCM-selected scene frame. Within this subset, Full SER reaches 77.42% event recall and 74.19% target recall, compared with 54.84% and 51.61% for the $3\times3$ global input. The gain persists even when the target is already observable in the global frames, supporting the role of persistent local evidence in associating global anomaly semantics with the responsible trajectory. Presenting the same nine global frames as separate image inputs yields 27.78% event recall and 19.72% target recall, compared with 31.48% and 22.54% for the $3\times3$ mosaic. Simply separating the global frames therefore does not reproduce the grounding gain obtained by Full SER. Full SER selects nearly the same number of trajectories as the nine-image global setting, 37 versus 38, while achieving substantially higher recall and reducing false positives from 0.58 to 0.19 per video. This indicates that more accurate event-to-trajectory association drives the gain, while the number of retained candidates remains nearly unchanged.

The local-only setting is highly selective, yielding no false-positive trajectories while recovering 22.22% of events and 15.49% of targets. Full SER expands event and target coverage to 59.26% and 45.07% while maintaining 0.19 false positives per video. Together, these results highlight the complementary roles of global anomaly semantics and persistent local evidence in trajectory-level grounding.

# C Additional Results

## C.1 Object-Level Results

We extend the object-level evaluation in Table 1 by reporting O-AP, O-Recall, and O-FDR. RBDC and TBDC follow the region- and track-based protocol introduced by Ramachandra & Jones (2020), where a ground-truth region is detected at IoU $\geq 0.1$ and a track is detected when at least 10% of its annotated regions are detected. For O-AP, O-Recall, and O-FDR, predictions are matched one-to-one to ground-truth objects within the same video and frame at IoU $\geq 0.1$, with unmatched and duplicate predictions counted as false positives. Using predictions pooled over all videos and frames, O-Recall is computed as $\mathrm{TP}/N_{\mathrm{GT}}$ to measure anomalous-object coverage and O-FDR as $\mathrm{FP}/N_{\mathrm{pred}}$ to measure false discoveries, while O-AP summarizes the pooled object-level precision-recall performance. Table 12 reports the full comparison on UCSD Ped2 and ShanghaiTech.

As shown in Table 12, SAVOR achieves the highest O-AP and the lowest O-FDR on both UCSD Ped2 and ShanghaiTech while maintaining over 80% O-Recall. On UCSD Ped2, SAVOR improves O-AP from 62.71 for TAO to 67.54 and lowers O-FDR from 55.16 to 17.99 while achieving 82.33 O-Recall. On ShanghaiTech, SAVOR improves the best competing O-AP from 38.55 to 40.95 and lowers O-FDR from 33.41 to 24.43 while achieving 81.79 O-Recall. These results demonstrate that SAVOR combines broad anomalous-object coverage with fewer false object-level detections.

## C.2 UBnormal Results

Alongside UCSD Ped2 and ShanghaiTech, we evaluate SAVOR on UBnormal as a standard VAD benchmark. Following the benchmark protocol, we report results on both the test and validation splits, using the test set for the primary comparison and the validation set as complementary evidence.

Table 12: Object detection results on UCSD Ped2 and ShanghaiTech.

| Method | Protocol | O-AP ↑ | O-Recall ↑ | O-FDR ↓ |
|---|---|---|---|---|
| *UCSD Ped2* | | | | |
| AnomalyRuler | PFT | 24.71 | 56.33 | 56.12 |
| TAO | PFT | 62.71 | **91.95** | 55.16 |
| **Ours** | TF | **67.54** | 82.33 | **17.99** |
| *ShanghaiTech* | | | | |
| HF2VAD | Tr | 29.31 | **94.91** | 89.43 |
| AnomalyRuler | PFT | 38.55 | 57.89 | 33.41 |
| TAO | PFT | 37.97 | 86.23 | 65.67 |
| AdaCLIP | ZS | 4.27 | 56.11 | 94.67 |
| AnomalyCLIP | ZS | 1.33 | 35.10 | 98.12 |
| **Ours** | TF | **40.95** | 81.79 | **24.43** |

Tr = trained, PFT = partial fine-tuning, ZS = zero-shot, and TF = training-free.

Table 13: Results on the UBnormal test set. "–" denotes an unavailable result. ‡ denotes UBnormal results reported in PANDA (Yang et al., 2025).

| Method | Setting | Micro-AUC ↑ | RBDC ↑ | TBDC ↑ |
|---|---|---|---|---|
| *Training / adaptation-based* | | | | |
| BAF-AT (Georgescu et al., 2022) | Normal only | 59.3 | 21.91 | 53.44 |
| BAF-AT | Normal + abnormal | 61.3 | 25.43 | 56.27 |
| TimeSformer (Bertasius et al., 2021) | Fine-tuned (SR=1/32) | 68.5 | 0.04 | 0.05 |
| TimeSformer | Fine-tuned (SR=1/8) | 64.1 | 0.04 | 0.05 |
| TimeSformer | Fine-tuned (SR=1/4) | 61.9 | 0.04 | 0.06 |
| SSMTL++v1 (Barbalau et al., 2023) | Self-supervised | 62.1 | 25.63 | 63.53 |
| SSMTL++v2 | Self-supervised | 56.0 | 20.33 | 57.76 |
| STG-NF (Hirschorn & Avidan, 2023) | Unsupervised | 71.8 | 31.70 | 62.30 |
| STG-NF | Supervised | **79.2** | 43.60 | **75.60** |
| AED-MAE (Ristea et al., 2024) | Unsupervised | 58.5 | 23.58 | 50.36 |
| SeeKer (Delić et al., 2025) | Normal only | 77.9 | 42.00 | 58.50 |
| AnomalyRuler (Yang et al., 2024) | Few-normal-shot | 71.9 | – | – |
| *Training-free / zero-shot* | | | | |
| ZS CLIP‡ (Radford et al., 2021) | Zero-shot | 46.20 | – | – |
| LLaVA-1.5‡ (Liu et al., 2024) | Training-free | 53.71 | – | – |
| LAVAD‡ (Zanella et al., 2024) | Training-free | 64.23 | – | – |
| PANDA (Yang et al., 2025) | Training-free (Offline) | 75.78 | – | – |
| PANDA | Training-free (Online) | 72.41 | – | – |
| URF (Lin et al., 2025) | Zero-shot | 69.02 | – | – |
| AnyAnomaly (Ahn et al., 2026) | Zero-shot | 74.50 | – | – |
| **Ours** | Training-free | 75.68 | **65.47** | 74.99 |

Table 13 reports the full test-set comparison across frame-level anomaly detection and localization. Without target-scene training or parameter updates, SAVOR obtains 75.68 Micro-AUC, closely matching PANDA offline at 75.78 while outperforming all other compared training-free and zero-shot baselines. For localization, SAVOR achieves the highest RBDC of 65.47 and the second-highest TBDC of 74.99 among methods with reported localization results, with its TBDC only 0.61 points below supervised STG-NF. These results demonstrate that SAVOR generalizes its training-free detection and localization capabilities to UBnormal, with particularly strong localization performance.

We further report results on the UBnormal validation set as complementary evidence. Table 14 reports Micro-AUC and Macro-AUC for frame-level anomaly detection together with RBDC and TBDC for localization. The validation comparison is presented separately because it uses a different data split and includes a distinct set of available baselines. SAVOR achieves the highest RBDC of 72.59 and TBDC of 78.35 on the UBnormal validation set, exceeding the strongest competing

Table 14: Results on the UBnormal validation set. "–" denotes an unavailable result. † denotes our reproduction using the official implementation.

| Method | Setting | Micro-AUC | Macro-AUC | RBDC | TBDC |
|---|---|---|---|---|---|
| *Training / adaptation-based* | | | | | |
| BAF-AT (Georgescu et al., 2022) | Normal only | 58.5 | 94.4 | 18.58 | 48.21 |
| BAF-AT | Normal + abnormal | 68.2 | **95.3** | 28.65 | 58.10 |
| TimeSformer (Bertasius et al., 2021) | Fine-tuned (SR=1/32) | 86.1 | 89.2 | 0.01 | 0.02 |
| TimeSformer | Fine-tuned (SR=1/8) | 83.4 | 90.6 | 0.01 | 0.02 |
| TimeSformer | Fine-tuned (SR=1/4) | 78.5 | 89.2 | 0.01 | 0.02 |
| T-VAU/AHD (Gu et al., 2026) | One-shot | 94.5 | 85.2 | 64.30 | 74.40 |
| T-VAU/AHD | Fine-tuned | **94.8** | 87.8 | 67.80 | 76.70 |
| AnomalyRuler† (Yang et al., 2024) | Few-normal-shot | 72.3 | – | 47.95 | 64.98 |
| *Training-free* | | | | | |
| **Ours** | Training-free | 70.7 | 86.7 | **72.59** | **78.35** |

Table 15: Controlled evidence presentation on 20 UCF-Crime test videos. Entries denote recognized events out of 20.

| Setting | Evidence | | Recognition | |
|---|---|---|---|---|
| | Temporal | View | Event-specific | Behavior-level |
| Uniform Global | Uniform | Full | 8/20 | 10/20 |
| Oracle Global | Oracle interval | Full | 7/20 | 13/20 |
| Oracle Local | Oracle interval | Local | 8/20 | 13/20 |
| Mixed Global | Uniform + oracle | Full | 7/20 | 13/20 |
| **Mixed Global-Local** | Uniform + oracle | Full + local | **10/20** | **16/20** |

results by 4.79 and 1.65 points, respectively. At the frame level, SAVOR obtains 70.7 Micro-AUC and 86.7 Macro-AUC without target-scene training or parameter updates. Together with the test-set results, these findings demonstrate that SAVOR maintains strong localization performance across both UBnormal splits.

### C.3 Cross-Dataset Analyses

**UCF-Crime Context Preservation** To examine whether context-preserving evidence transfers beyond the primary benchmarks, we conduct a controlled analysis on UCF-Crime (Sultani et al., 2018), a real-world surveillance dataset with diverse anomaly categories. We use 20 anomalous test videos spanning eight categories and annotate the target entity at nine fixed timestamps within the official anomaly interval, yielding 180 sparse bounding-box annotations. All five conditions are evaluated on this same set with the VLM, prompt, decoding configuration, nine-observation budget, and one model call per video fixed, while varying only temporal sampling and spatial evidence presentation.

Table 15 summarizes the five controlled conditions. The first three compare uniform global observations, oracle temporal concentration, and isolated local views. The final two use a mixed temporal composition of six uniformly sampled observations and three observations from the annotated anomaly interval, differing only in spatial presentation. Generated descriptions are evaluated under a blind protocol with randomized responses and hidden condition identities. Event-specific recognition requires the core event semantics and principal entities, while behavior-level recognition also accepts a correct coarser description of the abnormal behavior.

Oracle temporal concentration primarily benefits behavior-level recognition, while isolated local views provide only marginal additional gains. In contrast, under the shared mixed temporal composition, adding enlarged local evidence improves event-specific recognition from 7/20 to 10/20 and behavior-level recognition from 13/20 to 16/20, supporting the joint use of scene context and object-focused evidence.

Table 16: Traffic-event grounding results on WTS. † denotes our reproduction using Qwen3.6-Plus.

| Method | VLM | Successes / Evaluated | Success Rate (%) |
|---|---|---|---|
| *Published SeeUnsafe results* | | | |
| SeeUnsafe (Zhang et al., 2025b) | GPT-4o mini | 35/136 | 25.74 |
| SeeUnsafe | GPT-4o | 70/136 | 51.47 |
| *Shared-backbone comparison* | | | |
| SeeUnsafe† | Qwen3.6-Plus | 48/136 | 35.29 |
| **Ours** | Qwen3.6-Plus | **73/136** | **53.68** |

**WTS Traffic-Event Grounding** Beyond the standard VAD benchmarks, we evaluate SAVOR on WTS (Kong et al., 2024) to examine whether its event-centric reasoning and object grounding transfer to complex traffic interactions. We follow the visual-grounding protocol introduced by SeeUnsafe (Zhang et al., 2025b), which evaluates whether the road users involved in collision and near-miss events are correctly identified.

The SeeUnsafe grounding protocol considers 145 critical-event videos, including 106 collision and 39 near-miss cases, and retains the 136 videos containing at least three objects to avoid trivial grounding cases. A video is counted as successfully grounded only when both involved road users are correctly identified. Since most existing WTS evaluations focus on captioning or VQA, we use SeeUnsafe as the directly comparable object-grounding baseline. Since SeeUnsafe originally reports results with GPT-4o family models, we additionally reproduce it with Qwen3.6-Plus and compare both methods under the same VLM backbone.

As shown in Table 16, under the shared Qwen3.6-Plus backbone, SAVOR successfully grounds the event participants in 73 of 136 videos, yielding 25 additional cases over the reproduced SeeUnsafe baseline. These results show that SAVOR transfers its event-centric reasoning and object grounding capabilities to complex traffic interactions while maintaining favorable grounding performance.

# D Experimental Details

## D.1 Implementation

**Models and Inference** We use Qwen3.6-Plus as the frozen VLM and SAM2.1 Hiera-B+ (Ravi et al., 2025) with the official pretrained weights for mask propagation, without target-dataset training or parameter updates. The same VLM prompt templates are used across all datasets, with the complete prompts provided in Appendix D.2. We use a temperature/top-$p$ of 0.1/0.5 for global and local reasoning and 0/0.2 for evidence-guided grounding. All experiments are conducted with a single NVIDIA RTX 3090 24GB GPU.

**Motion and Temporal Routing** We analyze frame pairs using a temporal stride of 2 and compute dense Farnebäck optical flow at a working width of at most 640 pixels while preserving the original aspect ratio. For each analyzed pair, we remove dominant global translation as $\mathbf{R}_t(\mathbf{x}) = \mathbf{F}_t(\mathbf{x}) - \widetilde{\mathbf{f}}_t$, where $\widetilde{\mathbf{f}}_t$ denotes a robust componentwise median estimate of the flow field. We combine residual-flow and frame-difference saliency to identify localized moving regions. These regions are linked across analyzed frame pairs using predicted displacement together with spatial, geometric, and motion consistency.

Following Eq. 2, we set $Q_{\text{lo}}$ and $Q_{\text{hi}}$ to the 5th and 95th percentiles of the candidate-keyframe motion magnitudes in each video. We set $\tau = 10$ in Eq. 3 and use a 15-frame moving-average window to obtain $\widetilde{E}(t)$. TCM uses a fixed budget of nine global observations and partitions the video timeline into three macro temporal intervals. For Eq. 4, we set $\lambda = 0.45$ and $\gamma = 0.50$. The selected frames are ordered chronologically and arranged into a $3 \times 3$ mosaic.

**TCM Visual Emphasis** TCM locally enhances motion-supported regions through brightness and sharpness adjustment while preserving the surrounding scene context. The resulting mosaic is provided to the global VLM with instructions to treat the emphasized regions as attention cues rather

than anomaly labels and to determine anomaly semantics from visual evidence and the surrounding scene context.

**SER Event Construction** For each retained trajectory, SER samples five temporally distributed observations to construct its object-event view. Each observation is extracted using a square context crop whose side length is $2.8\times$ the longer side of the corresponding trajectory box. We retain at most eight candidate trajectories for each video across all datasets.

**APR Retargeting and Propagation** SAM2 grounding prompts are generated at a keyframe stride of 10 frames, and propagation uses a prompt stride of 1. An immediate mask is considered structurally valid when its box containment satisfies $\rho(M, b) \geq 0.90$ and it contains at most 10 connected components. For retargeting, candidate locations are quantized to a fixed $3 \times 3$ grid inside the SER box, with the original center excluded from the alternatives. APR evaluates at most one alternative anchor with one additional SAM2 query and retains the original anchor if the alternative fails the structural or target-consistency checks. After anchor validation, each anomaly instance is propagated independently in both temporal directions with frozen SAM2, and the two directional mask sequences are combined by union. Propagation stops when the mask remains below 200 foreground pixels for five consecutive frames after the target has first appeared.

**Baseline Reproduction** We reproduce baseline results when the results required for our evaluation protocol are unavailable from the original publications, using the official implementations and released models when available. All other baseline results are taken from the corresponding publications unless stated otherwise.

### D.2 VLM Prompts and Example

We use the same VLM prompt templates across all datasets. To illustrate how these prompts interact with the inference stages, we trace ShanghaiTech test video 08_0058 as a running example.

**Global Reasoning Prompt** Our global prompt partially follows the structured prompting practice of GridVAD (Eltahir et al., 2026), while we substantially redesign the task definition, visual evidence cues, and output protocol for SAVOR. The prompt identifies anomalies from scene-relative deviations and dominant behavior without specifying scene-dependent anomaly categories.

***Global Input***

Motion-emphasized $3 \times 3$ TCM mosaic of ShanghaiTech video 08_0058. The tiles are arranged in temporal order from left to right and top to bottom. The emphasized regions provide visual attention cues and carry no anomaly labels.



***Global Prompt***

You are an advanced Video Anomaly Detection assistant. You are analyzing a video clip where a $3 \times 3$ grid of frames (Temporal mosaic) is stitched into a single image. The frames are in temporal order and

indexed from 1 to 9, from left to right and top to bottom. Brightened areas are only visual hints, not anomaly labels. Your task is to identify possible anomalous events based on visual content.

*Definition of task.* Consider the following cues:

1. An anomaly is an event, action, motion pattern, object, or interaction that deviates from the surrounding scene or dominant behavior.
2. Spatial & Contextual Mismatch. An object or person appears, moves, or behaves in an unusual area for the scene.
3. Kinematic Contrast. An object or person shows unusual speed, direction, trajectory, or motion compared with nearby entities.
4. Social & Safety Violations. Actions that could cause danger, disruption, or injury to others. Do not infer danger without visible evidence.
5. Motion and Pose Cues. Pay attention to body posture across frames, such as larger stride changes and faster posture transitions than normal behavior.

*Output Format:*
Return only a strictly valid JSON object.
If multiple people show the same anomalous behavior, output one JSON item for each person. Do not include analysis, reasoning, or explanations outside the JSON.
The JSON structure must be:

```
{
  "anomalies": [
    {
      "description": "concise noun phrase describing the abnormal
    object or event, emphasizing visible appearance when possible.
    No locations.",
      "instance_hint": "appearance, object type or color, motion
    timing, or tile location. Use temporal cues to distinguish
    similar anomalies.",
      "scene": "short noun phrase describing video location",
      "evidence_frame_indices": [1, 2]
    }
  ]
}
```

If no visually supported anomaly is found, return:

```
{
  "anomalies": []
}
```

***Global Outputs***

```
{
  "anomalies": [
    {
      "description": "a person running with a forward body lean and
    a faster gait than nearby pedestrians",
      "instance_hint": "the backpacked runner moving rapidly through
     tiles 4 to 6",
      "scene": "outdoor pedestrian walkway",
      "evidence_frame_indices": [4, 5, 6]
    }
  ]
}
```

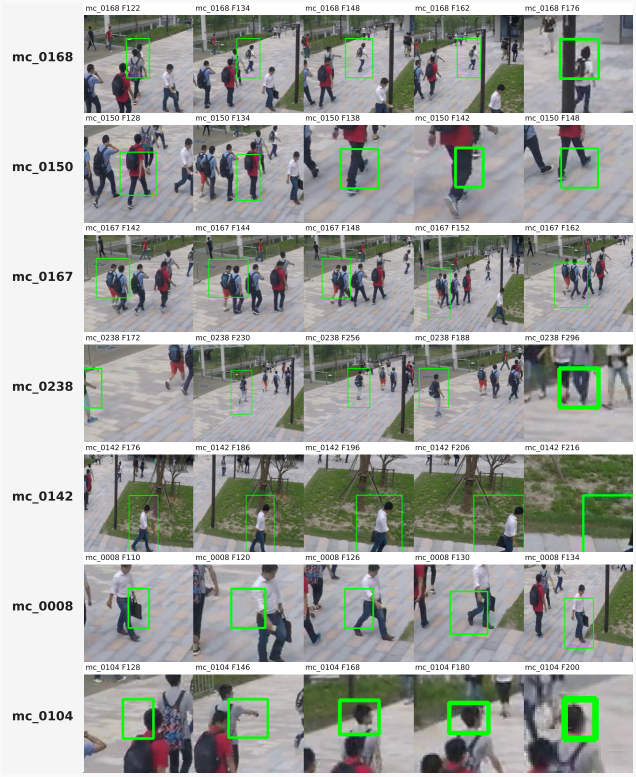
**Local Object-Event Prompt** SER performs local object-event reasoning over trajectory-centric candidate views using two complementary passes. The example below illustrates how candidate trajectory `mc_0168` in video 08_0058 is matched to the anomaly identified by global reasoning.

### *Local Input*

*Global anomaly hypotheses:*
If available, the preceding {`global anomaly descriptions`} are used as textual input. For the running example:

```
{
  "video_id": "08_0058",
  "anomalies": [
    {
      ...
    }
  ]
}
```

*Candidate trajectories:*
Each row corresponds to one retained candidate trajectory and presents temporally distributed observations of the same candidate. Green boxes indicate the object followed by the trajectory, while the surrounding context remains visible for local anomaly reasoning.




### *Local Prompt*

You are given candidate tracks over time, with green boxes marking the tracked objects. Inspect the marked objects and their surrounding context.

*Trajectory association rule:*

1. Select a trajectory only if its green boxes consistently cover the abnormal target.
2. A nearby, background, partially overlapping, reflection-only, or shadow-only target is not a valid association.
3. If the abnormal target is visible but not reliably covered, use the row only as context, keep the trajectory unassigned, and trigger semantic rescue.

*Pass A. Global-conditioned association:*

For each global anomaly hypothesis, associate valid candidate trajectories according to the rule above. The hypotheses and candidate trajectory IDs are provided at inference time.
*Output Format:*

```
{
  "anomalies": [
    {
      "description":"{global_description}",
      "instance_hint":"{global_instance_hint}",
      "scene":"{global_scene}",
      "evidence_frame_indices": [{global_evidence_frame_indices}],
      "selected_track_id": "candidate track id",
      "selected_track_valid": true,
      "box_target_relation": "inside | uncertain",
      "boxed_target_description": "object enclosed by the green
    boxes",
      "target_enclosed_by_green_box": true,
      "rescue_required": false
    }
  ]
}
```

*Pass B. Independent local discovery:*
Independently inspect all candidate views, including regions outside the green boxes, for additional visually supported anomalies. Use global hypotheses only for duplicate checking, apply the same association rule above, and return an empty list if no additional anomaly is found.
*Output Format:*

```
{
  "anomalies": [
    {
      "description": "concise abnormal object or event",
      "instance_hint": "short visual or temporal cue",
      "scene": "short scene description",
      "evidence_frame_numbers": [0, 1],
      "context_track_id": "candidate track id or empty string",
      "box_target_relation": "outside | partial | uncertain",
      "boxed_target_description": "object enclosed by the green
    boxes",
      "target_enclosed_by_green_box": false,
      "rescue_required": true
    }
  ]
}
```

***Local Outputs***

For video 08_0058, `mc_0168` matches the runner, while the bicyclist is found outside the green trajectory boxes and sent to semantic rescue. Here, `rescue_crops` denotes clean original-frame crops guided by the contextual trajectory and used as inputs for semantic rescue.

```
{
  "anomalies": [
    {
      "description": "a person running with a forward body lean and
    a faster gait than nearby pedestrians",
      "instance_hint": "the backpacked runner moving rapidly through
     tiles 4 to 6",
      "scene": "outdoor pedestrian walkway",
      "evidence_frame_indices": [4, 5, 6],
      "selected_track_id": "mc_0168",
      "selected_track_valid": true,
      "box_target_relation": "inside",
```

```
      "boxed_target_description": "the green boxes follow the
    running pedestrian",
      "target_enclosed_by_green_box": true,
      "rescue_required": false
    },
    {
      "description": "a person riding a bicycle",
      "instance_hint":"The cyclist appears in the context of
    candidate mc_0168 but remains outside its green box.",
      "scene":"outdoor pedestrian walkway",
      "evidence_frame_numbers": [0, 37, 74, 112, 149],
      "context_track_id": "mc_0168",
      "box_target_relation": "outside",
      "boxed_target_description": "The green boxes enclose the
    running pedestrian while the bicyclist remains visible outside
    them",
      "target_enclosed_by_green_box": false,
      "rescue_required": true,
      "rescue_crops": [{ ... }]
    }
  ]
}
```


**Semantic Rescue Prompt** When no candidate trajectory can be reliably associated with an anomaly, SER re-localizes the target from its supporting visual evidence. In the example, the bicyclist lies outside the boxes of `mc_0168`, so evidence-guided grounding uses the supporting observations to recover coarse spatial grounding.

***Rescue Input***

*Unassigned anomaly:*


```
{
  "description": "a person riding a bicycle",
  "instance_hint": "The cyclist appears in the context of candidate
    mc_0168 but remains outside its green box.",
  "scene":"outdoor pedestrian walkway",
  "evidence_frame_numbers": [0, 37, 74, 112, 149],
  "context_track_id": "mc_0168",
  "rescue_required": true
}
```


*Supporting observations:*
Supporting frames are recovered without the trajectory boxes to re-localize the anomalous target.



***Rescue Prompt***

*Task:* The image supports the target anomaly {`anomaly description`}. Locate the anomalous target itself, which may appear anywhere in the image.

*Localization rules:*

1. The bounding box should fully contain the visible anomalous target.
2. If a person is using or carrying an object, include both as one target.
3. Do not localize background regions, shadows, object parts, or unrelated nearby people.
4. Estimate the target center as `[center_x, center_y]` and its bounding box as `[x_min, y_min, x_max, y_max]`.

5. Use normalized image coordinates from 0 to 1000, with `[0, 0]` at the top-left and `[1000, 1000]` at the bottom-right.

*Output Format:*

```
{
  "object": "{anomaly description}",
  "center": [center_x, center_y],
  "bbox": [x_min, y_min, x_max, y_max]
}
```

***Rescue Outputs***

For the bicyclist in frame 074, the VLM returns the localization below. The other supporting observations are processed similarly.

```
{
  "object": "a person riding a bicycle",
  "center": [715, 485],
  "bbox": [680, 410, 750, 560]
}
```

The resulting coarse localization is shown below:

