ANOMMERGE: SPATIALLY BOUNDED TOKEN MERGING FOR VIDEO ANOMALY UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly understanding (VAU) with multimodal large language models (MLLMs) aims to recognize, temporally localize, and explain abnormal events. Under limited visual token budgets, existing VAU systems often trade temporal coverage for spatial detail, potentially removing evidence needed for precise event boundaries. Separately, we observe that current VAU models can produce overly broad and low-dispersion anomaly intervals, while standard overlap metrics may not fully expose this degeneration. To address these challenges, we propose **AnomMerge**, a learnable token compression framework for VAU that reduces spatial redundancy while preserving observed temporal units and their positional correspondence. We further develop adaptive spatial merging with unit-specific similarity thresholds and spatial and feature consistency constraints to preserve fine-grained temporal evidence for anomaly localization. To better evaluate localization behavior, we complement standard overlap metrics with coverage-aware temporal IoU (CA-tIoU) and interval dispersion contraction (IDC), which measure excess temporal coverage and contraction of predicted interval dispersion, respectively. On Vad-Reasoning, AnomMerge reduces IDC by 44.3% relative to VAU-R1, while improving video-level anomaly recognition and retaining only 36.59% of the original visual tokens. The code will be publicly released.

1 INTRODUCTION

Video anomaly understanding (VAU) requires recognizing abnormal events, localizing their temporal extent, and explaining what happened. These capabilities are important for practical video monitoring, where an anomaly prediction is most useful when the relevant moment can be quickly identified and the underlying event can be understood and verified. By providing temporally grounded and semantically interpretable anomaly information, VAU can support efficient video review and downstream decision-making. Recent multimodal large language models (MLLMs) have substantially advanced these capabilities through structured reasoning and reinforcement fine-tuning (Huang et al., 2026a; Zhu et al., 2025b). Such fine-grained understanding, however, relies on retaining sufficient visual evidence across both time and space.

Current VAU systems remain constrained by limited visual input budgets. VAD-R1 uses 16 uniformly sampled frames in its main setting (Huang et al., 2026a), which can leave large temporal gaps in long videos, whereas VAU-R1 samples at one frame per second (FPS) (Zhu et al., 2025b), providing denser temporal coverage at higher visual-token cost. Under a fixed budget, this creates a trade-off between temporal coverage and spatial detail, as illustrated in Figure 1. Existing video token compression methods reduce this cost through query-guided token selection (Yu et al., 2026) or spatial and temporal token merging (Bolya et al., 2022; Tao et al., 2025; Fu et al., 2025; Fan et al., 2026). However, generative VAU poses distinct challenges: a general anomaly instruction provides limited guidance for selecting evidence of an unspecified event, while visually similar tokens may still represent different temporal states or contain distinct local anomaly cues. *The key challenge is therefore to reduce visual-token cost while preserving temporally distinct observations and local evidence required for anomaly understanding.*

To address this challenge, we propose **AnomMerge**, which preserves observed temporal units while compressing spatial redundancy within each unit. AnomMerge first contextualizes visual tokens with spatial and temporal information, and then predicts a unit-specific similarity threshold from

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.

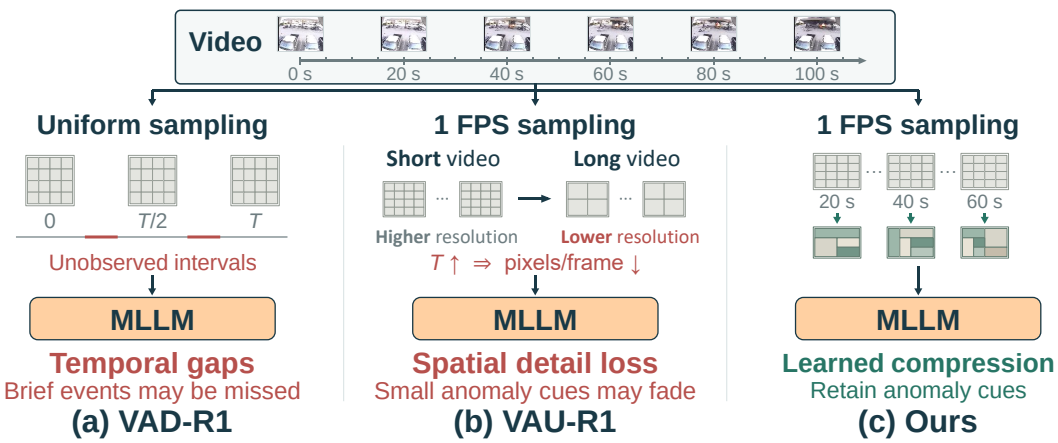

Figure 1: Visual-evidence trade-offs under a limited input budget. (a) Fixed-count sampling leaves temporal gaps. (b) 1 FPS sampling under a fixed pixel budget reduces spatial detail. (c) AnomMerge compresses spatial redundancy while preserving observed temporal units.

local token statistics. Within each temporal unit, spatially adjacent tokens are adaptively grouped under constraints on group size, spatial extent, and feature consistency, and each group is pooled into a compact token while retaining an original spatiotemporal coordinate for positional encoding. We further identify overly broad interval predictions in VAD-R1 and complement standard overlap metrics with coverage-aware temporal intersection-over-union (CA-tIoU) and interval dispersion contraction (IDC) to assess excess temporal coverage and contraction in predicted interval dispersion. On the Vad-Reasoning dataset, AnomMerge retains only 36.59% of the original visual tokens while improving F1 from 0.890 to 0.925 and CA-tIoU from 0.373 to 0.383, and reducing IDC from 0.255 to 0.142 compared with VAU-R1. These results demonstrate the effectiveness of preserving temporal observations while compressing spatial redundancy for VAU.

Our contributions are threefold:

- We observe that existing generative VAU models often predict overly broad temporal intervals with limited variation across videos, and introduce an evaluation framework in which CA-tIoU and IDC complement standard temporal overlap metrics by measuring excess temporal coverage and interval-dispersion contraction.
- We propose **AnomMerge**, a task-adapted token compression framework for video anomaly understanding that learns compact visual representations while preserving evidence relevant to anomaly recognition and temporal localization.
- We design adaptive spatial merging within each temporal unit, using learned similarity thresholds and spatial constraints to reduce redundancy while preserving temporal resolution and positional correspondence.

2 Related Work

Video Anomaly Understanding. VAU has evolved from anomaly scoring (Liu et al., 2018; Sultani et al., 2018; Zhou et al., 2023) to semantic interpretation and explanation (Zanella et al., 2024; Ye et al., 2025; Huang et al., 2025). Richer supervision further broadens the task to analysis and understanding across temporal granularities (Du et al., 2024; Zhang et al., 2025b). Recent approaches combine structured reasoning, reinforcement fine-tuning, and reflection to improve anomaly judgments and temporal grounding (Huang et al., 2026a; Zhu et al., 2025b; Zhao et al., 2026; Huang et al., 2026b). However, limited visual budgets constrain temporal coverage and spatial detail, motivating token compression that preserves evidence for recognizing and localizing anomalies.

Video Token Compression. Video token compression reduces redundancy through pruning or token selection (Shen et al., 2026), spatial or temporal merging (Bolya et al., 2022; Tao et al., 2025; Fu et al., 2025; Fan et al., 2026), and token-budget allocation (Liu et al., 2026). Compression can

also be guided by queries (Yu et al., 2026) or learned for downstream tasks (Ryoo et al., 2021; Syed Ariff et al., 2026). Most methods target general or query-conditioned video understanding, with task-specific compression explored for anomaly detection (Dai et al., 2026). Generative VAU poses a challenge: the anomaly is not specified in advance, limiting the usefulness of query guidance, while temporal compression may blur distinctions required for accurate localization. This motivates preserving temporal observations and concentrating compression within each temporal unit.

Video Temporal Grounding. Video temporal grounding localizes video segments corresponding to a textual query. MLLM-based methods integrate temporal positions into visual representations and generate event boundaries as timestamps (Ren et al., 2024; Guo et al., 2025; Li et al., 2025a). Grounded tuning and adaptive frame scaling support grounding across video lengths (Zeng et al., 2025; Li et al., 2026), while reinforcement learning advances reasoning-guided localization and generalist interval-set prediction (Wang et al., 2026b; Zhu et al., 2026). Unlike grounding with a specified target event, generative VAU must first discover the abnormal event before localizing it. Meanwhile, existing VAU models can produce overly broad and low-dispersion anomaly intervals, a failure mode that standard overlap metrics do not fully characterize.

3 Evaluating Localization Beyond Overlap

Unlike query-conditioned video temporal grounding, where a textual query specifies the target event (Zhang et al., 2020; Lei et al., 2021), generative VAU typically relies on a general anomaly understanding system prompt (Huang et al., 2026a; Zhu et al., 2025b). The model must identify whether and what anomaly occurs and localize its temporal extent. We observe that existing generative VAU models can produce overly broad and similar anomaly intervals across videos, which standard overlap metrics do not fully characterize. We therefore introduce complementary measures for evaluating this localization behavior.

Task and notation. Given a video of duration D and a general system prompt s, the model predicts whether an anomaly occurs, its temporal extent if present, and a semantic explanation. For a ground truth (GT) anomaly interval $G = [a, b]$ and a valid prediction $P = [\hat{a}, \hat{b}]$, let $g = |G|$ denote the GT duration, $I = |P \cap G|$ the temporal overlap, and $e = |P \setminus G|$ the excess predicted duration. The temporal IoU (tIoU) can be formulated as:

$$J = \frac{|P \cap G|}{|P \cup G|} = \frac{I}{g + e}, \tag{1}$$

and we define $q = g/D$ as the fraction of the video occupied by the GT anomaly.

Limitations of overlap-based evaluation. Our evaluation of VAD-R1 (Huang et al., 2026a) shows that predicted anomaly intervals are often broad and highly similar across videos. As shown in Figure 2, VAD-R1 predictions are strongly concentrated near the full video duration, whereas GT interval lengths vary considerably. This reveals clear interval collapse in temporal localization.

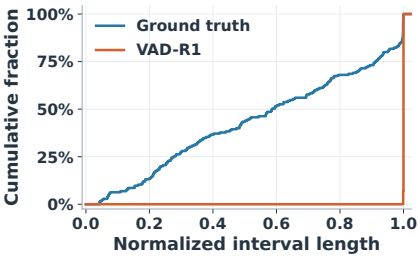

Figure 2: Cumulative distributions of GT and VAD-R1 predicted interval length on the Vad-Reasoning dataset.

Standard overlap metrics only partially characterize this collapse. If the prediction covers the entire video, $P = [0, D]$, temporal IoU reduces to $J = q = g/D$, so a long GT event can still obtain a high overlap score despite background coverage. Moreover, tIoU does not distinguish excessive coverage from missed event content. For example, for $G = [10, 90]$ in a 100-second video, both $P = [0, 100]$ and $P = [18, 82]$ obtain a tIoU of 0.8, although the former includes extra background while the latter misses part of the event. Aggregate overlap metrics also do not capture reduced variation of predicted intervals across videos. We therefore introduce two complementary measures: CA-tIoU for excess temporal coverage and IDC for contraction of interval dispersion.

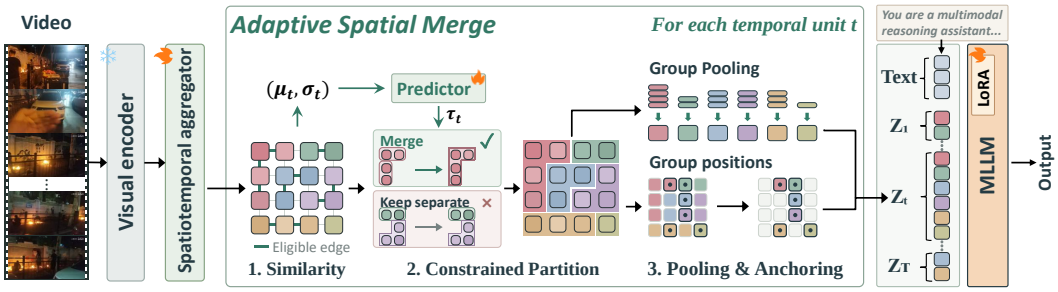

Figure 3: Overview of AnomMerge. First, the video is encoded into visual tokens and enriched with spatiotemporal context. Second, each temporal unit is compressed independently through learned threshold prediction and constrained spatial grouping. Finally, grouped tokens are pooled and anchored to original positions before being passed to the MLLM for anomaly understanding.

Coverage-aware temporal intersection-over-union (CA-tIoU). CA-tIoU augments tIoU with an explicit penalty on predicted duration outside the GT interval:

$$\text{CA-tIoU} = \frac{J}{1 + \lambda\, e/g}, \qquad \lambda \geq 0. \tag{2}$$

where e/g measures the excess predicted duration relative to the GT event length, and λ controls the penalty strength. When $\lambda = 0$, CA-tIoU reduces to tIoU. We set $\lambda = 1$; in this case, a prediction covering the entire video receives q^2 instead of q, imposing a stronger penalty on excessive temporal coverage and extra background.

Interval dispersion contraction (IDC). While CA-tIoU evaluates interval quality for individual videos, it does not capture whether predicted intervals collapse toward similar temporal ranges across videos. Let $\mathcal{A}$ denote the set of videos containing a GT anomaly, and let $\mathcal{V} \subseteq \mathcal{A}$ denote those with valid predicted intervals. Indexing videos by v, we define the normalized intervals as $P_v^{\text{norm}} = [\hat{a}_v/D_v, \hat{b}_v/D_v]$ and $G_v^{\text{norm}} = [a_v/D_v, b_v/D_v]$. The predicted and GT endpoint dispersions and IDC are defined as:

$$\begin{aligned} \mathcal{D}_P &= \sqrt{\text{Var}_{v\in\mathcal{V}}(\hat{a}_v/D_v) + \text{Var}_{v\in\mathcal{V}}(\hat{b}_v/D_v)}, \\ \mathcal{D}_G &= \sqrt{\text{Var}_{v\in\mathcal{V}}(a_v/D_v) + \text{Var}_{v\in\mathcal{V}}(b_v/D_v)}, \\ \text{IDC} &= 1 - \min\left(1, \frac{\mathcal{D}_P}{\mathcal{D}_G}\right). \end{aligned} \tag{3}$$

IDC measures the contraction of predicted endpoint dispersion relative to the GT dispersion. When GT intervals vary across videos, a high IDC indicates that predictions concentrate within a much narrower temporal range, reflecting interval collapse. In the extreme case where all videos receive the same normalized interval, $\text{IDC} = 1$, whereas $\text{IDC} = 0$ when $\mathcal{D}_P \geq \mathcal{D}_G$. Since IDC measures distributional contraction rather than localization accuracy, we interpret it together with CA-tIoU and standard overlap metrics to provide a more complete assessment of VAU.

4 AnomMerge: Spatially Bounded Token Merging

Overview. We sample frames from the input video and extract visual tokens with a frozen visual encoder. As shown in Figure 3, AnomMerge first enriches these tokens with spatiotemporal context, as described in Section 4.1. It then performs adaptive spatial merging independently within each temporal unit, including threshold prediction, constrained token grouping, and group pooling with positional anchoring, as detailed in Section 4.2. The resulting compact tokens preserve all temporal units and are passed with their positional coordinates and the system prompt s to the MLLM for anomaly recognition, localization, and explanation.

4.1 SPATIOTEMPORAL STATE-SPACE AGGREGATION

We adopt the spatial-then-temporal Mamba (Gu & Dao, 2023) aggregation of STAC (Syed Ariff et al., 2026) and extend its causal temporal scan to bidirectional temporal encoding. This allows each temporal unit to incorporate both preceding and subsequent observations, providing fuller temporal context for anomaly understanding and subsequent spatial grouping.

Given vision encoder features $\mathbf{F}_{\mathrm{e}} \in \mathbb{R}^{T\times N\times d}$, where T is the number of temporal units, N is the number of spatial tokens per unit, and d is the feature dimension, the aggregator produces contextualized features $\mathbf{F}_{\mathrm{st}} \in \mathbb{R}^{T\times N\times d}$ while retaining all input tokens. Appendix A.1 provides the Mamba formulation and aggregation details; Appendix B.2 reports ablations of the spatial and temporal contextualization modules.

4.2 ADAPTIVE SPATIAL MERGING

Let $\mathbf{f}_{t,n} = \mathbf{F}_{\mathrm{st}}[t, n, :] \in \mathbb{R}^d$ denote the contextualized visual token feature at temporal unit t and spatial location n. Adaptive spatial merging comprises threshold prediction, constrained token grouping, and token aggregation with source-anchored positions.

Threshold prediction. Let $\mathcal{E}$ be the undirected edge set of the spatial token grid, where each edge $(i, j) \in \mathcal{E}$ connects horizontally or vertically adjacent tokens indexed by i and j. For temporal unit t, we define $a_t(i, j) = \cos(\mathbf{f}_{t,i}, \mathbf{f}_{t,j})$ as the cosine similarity. Let μ_t and σ_t denote the mean and sample standard deviation of these edge similarities over $\mathcal{E}$. A shared multilayer perceptron (MLP) takes $[\mu_t, \sigma_t]$ as input and predicts a coefficient k_t, giving the merging threshold

$$\tau_t = \mu_t + k_t\sigma_t. \tag{4}$$

The threshold remains fixed during grouping within unit t and controls merge eligibility. Predictor details are provided in Appendix A.2.

Constrained token grouping. Within each temporal unit, we initialize singleton groups and process neighboring-token pairs in descending similarity order, considering only pairs with $a_t(i, j) \geq \tau_t$. For two distinct groups C, C', their union $U = C \cup C'$ is accepted only if

$$\begin{aligned} |U| &\leq m_{\max}, \\ \max\{h(U), w(U)\} &\leq b_{\max}, \\ \cos(\mathbf{m}_{t,C}, \mathbf{m}_{t,C'}) &\geq \tau_t, \end{aligned} \tag{5}$$

where $h(U), w(U)$ are the bounding-box height and width in grid cells, and $\mathbf{m}_{t,C}$ is the mean feature of group C. The limits $m_{\max}$ and $b_{\max}$ bound group size and spatial extent, respectively. Group centers are updated after each merge. The resulting partition $\mathcal{C}_t$ contains spatially connected groups within the same temporal unit, with unmerged tokens retained as singletons. Coordinate and bounding-box calculations are provided in Appendix A.3.

Token pooling and positional anchoring. For each final group $C \in \mathcal{C}_t$, we mean-pool its contextualized features into one token:

$$\mathbf{z}_{t,C} = \frac{1}{|C|} \sum_{n\in C} \mathbf{f}_{t,n}. \tag{6}$$

Each pooled token retains the original temporal coordinate ρ_t of its source temporal unit and uses an original group member as its spatial anchor. We sort the member indices and select the middle member, choosing the lower-indexed of the two middle members for an even-sized group; its index is denoted by $n^*_{t,C}$. With $r(n)$ and $c(n)$ denoting the row and column coordinates of token n, the coordinates used for positional encoding are

$$\mathbf{p}_{t,C} = \left(\rho_t,\, r(n^*_{t,C}),\, c(n^*_{t,C})\right). \tag{7}$$

Thus, the pooled feature represents the entire group while its position remains anchored to an original member. Groups are ordered by temporal unit and anchor index, stacking their features into $\mathbf{Z} \in \mathbb{R}^{K\times d}$ and their aligned coordinates into $\mathbf{P}$ of shape $K \times 3$, where K is the total number of groups. The compressed tokens and positional coordinates are passed with the system prompt s to the MLLM for anomaly recognition, localization, and explanation.

In Appendix A.3, Algorithm 1 summarizes the complete token compression procedure.

4.3 TRAINING OBJECTIVE AND OPTIMIZATION

We train AnomMerge through supervised fine tuning (SFT) with an autoregressive language modeling objective. Conditioned on the system prompt s, compressed visual tokens $\mathbf{Z}$, and their positions $\mathbf{P}$, the MLLM predicts each target token from the preceding ground truth tokens. For a target sequence $\mathbf{y} = (y_1, \ldots, y_L)$ containing the reasoning and answer, we minimize the average negative log likelihood and add a compression regularizer:

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{L}\sum_{\ell=1}^{L} \log p(y_\ell \mid y_{<\ell}, s, \mathbf{Z}, \mathbf{P}), \qquad \mathcal{L} = \mathcal{L}_{\mathrm{SFT}} + \lambda_c \mathcal{L}_{\mathrm{comp}}. \tag{8}$$

Here p denotes the MLLM's token distribution conditioned on preceding target tokens $y_{<\ell}$. The regularizer $\mathcal{L}_{\mathrm{comp}}$, defined in Appendix A.2, Equation 14, encourages compression with weight λ_c.

We jointly optimize the contextualizer, threshold predictor, and MLLM adaptation parameters. Because threshold comparisons are discrete, a straight-through estimator (STE) (Bengio et al., 2013) provides surrogate gradients for threshold prediction.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets. Vad-Reasoning (Huang et al., 2026a) provides structured annotations for video anomaly recognition, temporal localization, and semantic explanation. We use 1,755 videos from its SFT subset for training and 439 videos for evaluation, including 204 abnormal videos. VANE-Bench (Gani et al., 2025) assesses anomaly understanding through multiple choice questions on surveillance videos and videos generated by AI. We evaluate eight source subsets, excluding UCF-Crime (Sultani et al., 2018) due to overlap with the Vad-Reasoning training set.

Implementation details. *(1) Model configuration.* AnomMerge uses Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) as its backbone. The contextualizer contains one spatial and one temporal Mamba block, and spatial grouping uses $m_{\max} = 8$ and $b_{\max} = 4$. *(2) Input processing.* Videos are sampled at 1 FPS; sequences exceeding 768 frames are uniformly subsampled to 768 frames over the full video duration. The frame resolution is set to approximately $128 \times 28 \times 28$ pixels. Every two consecutive sampled frames form one temporal unit. *(3) Training and inference.* The visual encoder and base language model are frozen; the contextualizer, threshold predictor, and LoRA adapters (Hu et al., 2021) are jointly trained for five epochs on a single 48 GB GPU using AdamW (Loshchilov & Hutter, 2017), a batch size of 1 with gradients accumulated over 8 steps, and $\lambda_c = 10^{-3}$. Learning rates are 2×10^{-5} for the adapters and 5×10^{-5} for the compressor. We generate up to 1,024 tokens.

Compared methods. We compare three groups of methods: (1) general video MLLMs (Wang et al., 2025; Zhu et al., 2025a; Zhang et al., 2025a; 2024; Bai et al., 2025b); (2) video reasoning models (Wang et al., 2026a; Feng et al., 2026; Li et al., 2025c); and (3) specialized VAU methods (Zhang et al., 2025b; Tang et al., 2024; Huang et al., 2026a; Zhu et al., 2025b). The specialized methods focus on recognizing and explaining anomalous events, with recent approaches incorporating structured reasoning to improve anomaly judgments and temporal localization.

Evaluation metrics. *(1) Anomaly recognition.* We report accuracy and F1 on Vad-Reasoning and single choice accuracy on VANE-Bench. *(2) Temporal localization.* We report mIoU and recall at temporal IoU thresholds of 0.3 and 0.5, denoted by $\mathrm{R@0.3}$ and $\mathrm{R@0.5}$, respectively. Following the VAD-R1 protocol, correctly classified normal videos receive an IoU of 1. We additionally report CA-tIoU and IDC on GT-abnormal videos to assess temporal specificity, as defined in Section 3. *(3) Semantic explanation.* We report BLEU-2 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-2 recall (Lin, 2004). *(4) Compression and efficiency.* Visual-token retention is the ratio of total retained to total original visual tokens over the evaluated videos. We also report estimated FLOPs in TFLOPs and time to first token (TTFT) in seconds.

Table 1: Ablation of spatial and temporal merging on Vad-Reasoning. Ret. denotes visual-token retention (%). **Bold** and underlined values indicate the best and second-best results, respectively.

Variant	Spatial	Temporal	Ret.	Acc ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑
(a)	✗	✗	100.00	**0.938**	**0.933**	**0.412**	0.149	**0.732**
(b)	✗	✓	36.88	0.909	0.899	0.372	0.197	0.713
Ours	✓	✗	36.59	0.932	0.925	0.383	**0.142**	0.722

Table 2: Comparison of different token grouping strategies on Vad-Reasoning.

Grouping strategy	Ret.	Acc ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑
Spatial average pooling	40.04	0.923	0.915	0.385	0.167	0.724
STAC-S	50.38	0.923	0.915	**0.407**	0.150	**0.733**
Ours	36.59	**0.932**	**0.925**	0.383	**0.142**	0.722

5.2 Ablation Studies

We perform ablation studies on Vad-Reasoning to study the effects of the merging dimension, token grouping strategy, merge constraints, and model scale and backbone.

Spatial versus temporal merging. We first examine whether tokens should be merged across time or only within each temporal unit. As shown in Table 1, we compare three variants: no merging, temporal merging, and our adaptive spatial merging. Compared with no merging, our method reduces token retention from 100% to 36.59%, while Acc and F1 decrease only from 0.938 to 0.932 and from 0.933 to 0.925, respectively. CA-tIoU decreases from 0.412 to 0.383 and mIoU from 0.732 to 0.722, while IDC is slightly reduced from 0.149 to 0.142. At a comparable token budget, our spatial merging consistently outperforms temporal merging: Acc increases from 0.909 to 0.932 and F1 from 0.899 to 0.925, while CA-tIoU improves from 0.372 to 0.383, mIoU from 0.713 to 0.722, while IDC decreases by 27.9%. These results show that preserving the temporal units while compressing spatial redundancy yields more effective compression without sacrificing anomaly recognition and temporal localization to the same extent as temporal merging.

Comparison of token grouping strategies. We compare AnomMerge with spatial average pooling and STAC-S in Table 2. Spatial average pooling compresses each temporal unit using fixed non-overlapping spatial bins, whereas STAC-S Syed Ariff et al. (2026) merges temporally redundant tokens across frames at fixed spatial locations. In contrast, AnomMerge adaptively groups only spatially adjacent tokens within each temporal unit, preserving all observed temporal units. AnomMerge retains 36.59% of the visual tokens, compared with 40.04% for average pooling and 50.38% for STAC-S, while achieving the highest Acc, F1, and lowest IDC. Although STAC-S achieves higher CA-tIoU and mIoU, it retains substantially more tokens. These results show that adaptive spatial grouping within each temporal unit provides a favorable trade-off between token compression, anomaly recognition, and interval-dispersion control.

Merge constraints. We evaluate the spatial-extent and group-center consistency constraints in Table 3. Removing the extent constraint reduces F1 from 0.925 to 0.904, while removing center consistency increases IDC from 0.142 to 0.235. Removing both constraints yields more aggressive compression, reducing token retention from 36.59% to 33.41%, but lowers F1 to 0.916 and increases IDC to 0.219 despite a higher CA-tIoU of 0.403. Overall, these constraints limit excessive within-unit aggregation and mitigate the degradation caused by overly aggressive spatial compression.

Model scale and backbone. We further evaluate AnomMerge across different model sizes and Qwen backbone families in Table 4. Within Qwen2.5-VL, scaling from 3B to 7B consistently improves recognition and localization performance: F1 increases from 0.925 to 0.935, CA-tIoU improves by 9.7%, and IDC decreases by 16.9%. This indicates that AnomMerge can effectively leverage increased model capacity without changing its compression design.

Across backbone generations, however, improvements vary across different metrics. Compared with Qwen2.5-VL-3B, Qwen3-VL-2B increases CA-tIoU from 0.383 to 0.426 and reduces IDC from 0.142 to 0.056, but its F1 decreases from 0.925 to 0.905. Qwen3.5-2B achieves the highest CA-tIoU of 0.431 and mIoU of 0.765 while maintaining a comparable F1 of 0.922. These results

Table 3: Ablation of spatial extent and group center consistency constraints on Vad-Reasoning.

Variant	Extent	Center	Ret. (%)	Acc ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑
(a)	✗	✓	36.19	0.913	0.904	0.377	0.162	0.717
(b)	✓	✗	33.86	0.920	0.910	0.392	0.235	**0.735**
(c)	✗	✗	33.41	0.925	0.916	**0.403**	0.219	0.734
Ours	✓	✓	36.59	**0.932**	**0.925**	0.383	**0.142**	0.722

Table 4: AnomMerge across model sizes and backbone families.

Backbone	Size	Acc ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑
Qwen2.5-VL	3B	0.932	0.925	0.383	0.142	0.722
Qwen2.5-VL	7B	**0.941**	**0.935**	0.420	0.118	0.748
Qwen3-VL (Bai et al., 2025a)	2B	0.916	0.905	0.426	**0.056**	0.746
Qwen3.5 (Qwen Team, 2026)	2B	0.932	0.922	**0.431**	0.117	**0.765**

show that AnomMerge generalizes across model scales and Qwen backbone families, while different backbones exhibit different strengths in anomaly recognition and temporal localization.

5.3 Main Results

Comparison with existing methods. Table 5 compares AnomMerge with general video MLLMs, video reasoning models, and specialized VAU methods on Vad-Reasoning. (1) Off-the-shelf video MLLMs and reasoning models show limited temporal localization performance, highlighting the importance of task-specific adaptation for generative VAU. (2) AnomMerge yields a substantially lower IDC than VAD-R1, decreasing from 0.999 to 0.142, indicating much less contraction of predicted interval dispersion. (3) Under the same 1-FPS setting as VAU-R1, AnomMerge improves Acc from 0.907 to 0.932, F1 from 0.890 to 0.925, and CA-tIoU from 0.373 to 0.383, while reducing IDC from 0.255 to 0.142. Although VAU-R1 improves temporal coverage through denser sampling, AnomMerge further preserves informative visual evidence through learned spatial compression, achieving stronger anomaly recognition and more specific temporal localization.

Comparison with video token compression methods. We compare AnomMerge with representative video token compression methods under two input protocols in Table 6. Under the 1-FPS protocol, FastVID (Shen et al., 2026), VidCom^2 (Liu et al., 2025), FlashVID (Fan et al., 2026), and OTT-Vid (Kang et al., 2026) are applied to the same task-finetuned Qwen2.5-VL-3B reference and calibrated to approximately 40% token retention. AnomMerge retains 36.59% of the tokens while achieving the highest Acc, F1, CA-tIoU, and lowest IDC among these methods. It also reduces estimated FLOPs from 103.65 TFLOPs for the uncompressed reference to 53.15 TFLOPs.

Under the 128-frame protocol, we further compare with FlexSelect (Zhang et al., 2025c) and QTSplus (Li et al., 2025b) under a unified input setting. Since these methods are originally designed for different video input protocols, we uniformly sample 128 frames for all methods, use the same Qwen2.5-VL-3B backbone and training protocol, and normalize predicted intervals to the range of $[0, 1]$ for evaluation. At 37.07% token retention, AnomMerge achieves an F1 of 0.913 and CA-tIoU of 0.391, compared with 0.684 and 0.200 for QTSplus, while improving mIoU from 0.588 to 0.711. Together, these results demonstrate that AnomMerge achieves effective token reduction while preserving anomaly recognition and temporal localization across different input protocols.

Evaluation on VANE-Bench. We further evaluate AnomMerge on VANE-Bench using eight source subsets, excluding UCF-Crime because it overlaps with the Vad-Reasoning training data. VAD-R1, VAU-R1, and AnomMerge are reevaluated under the same input and decoding protocol, while HAWK results are taken from VAD-R1. AnomMerge reduces the average number of visual tokens per clip from 633.43 to 234.12, corresponding to 36.96% retention.

As shown in Table 7, AnomMerge achieves the best accuracy on VideoLCM and remains competitive on several generated-video subsets despite substantial token reduction. Compared with VAU-R1 and VAD-R1, the performance gap is larger on several subsets because VANE-Bench videos are relatively short, allowing full-token methods to retain more visual details. AnomMerge performs lower

Table 5: Comparison with existing methods on the Vad-Reasoning dataset.

Method	Size	Frames	Anomaly Reasoning			Anomaly Detection						
			BLEU-2	METEOR	ROUGE-2	Acc	F1	CA-tIoU ↑	IDC ↓	mIoU	R@0.3	R@0.5
Open-Source video MLLMs												
InternVideo2.5	8B	16	0.114	0.264	0.109	0.677	0.708	0.134	0.850	0.366	0.408	0.374
InternVL3	8B	16	0.132	0.287	0.130	0.786	0.754	0.269	0.373	0.606	0.688	0.610
VideoLLaMA3	7B	16	0.066	0.160	0.072	0.774	0.768	0.049	0.289	0.435	0.458	0.428
LLaVA-NeXT-Video	7B	16	0.083	0.216	0.094	0.615	0.507	0.028	0.538	0.433	0.433	0.428
Open-Source video reasoning MLLMs												
Open-R1-Video	7B	16	0.067	0.189	0.085	0.617	0.546	0.059	0.426	0.420	0.435	0.424
Video-R1	7B	16	0.135	0.319	0.134	0.503	0.650	0.241	0.493	0.186	0.244	0.187
VideoChat-R1	7B	16	0.135	0.292	0.124	0.681	0.721	0.193	**0.126**	0.379	0.408	0.351
MLLM-based VAD methods												
Holmes-VAU	2B	16	0.005	0.046	0.011	0.788	0.716	–	–	–	–	–
HAWK	7B	16	0.045	0.162	0.049	0.469	0.637	–	–	–	–	–
Qwen2.5-VL	7B	1 FPS	0.137	0.297	0.127	0.788	0.773	0.293	0.214	0.595	0.661	0.604
+ VAD-R1	7B	16	0.236	0.411	0.196	0.879	0.868	0.377	0.999	0.713	0.770	0.706
+ VAU-R1	3B	1 FPS	0.246	0.432	0.208	0.907	0.890	0.373	0.255	**0.739**	**0.811**	**0.743**
+ Ours	3B	1 FPS	**0.254**	**0.443**	**0.214**	**0.932**	**0.925**	**0.383**	0.142	0.722	0.802	0.713

Table 6: Comparison with video token compression methods under 1-FPS and 128-frame protocols.

Method	Sampling	Ret.	Acc ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑	FLOPs ↓	TTFT ↓
1 FPS: training-free baselines and AnomMerge									
Qwen2.5-VL-3B (SFT)	1 FPS	100%	0.879	0.856	0.365	0.211	**0.725**	103.65	1.140
+ FastVID	1 FPS	$\sim$ 40%	0.850	0.831	0.372	0.217	0.688	54.12	0.745
+ VidCom^2	1 FPS	$\sim$ 40%	0.886	0.874	0.362	0.221	0.702	54.23	**0.728**
+ FlashVID	1 FPS	$\sim$ 40%	0.884	0.868	0.379	0.231	0.717	54.22	0.769
+ OTT-Vid	1 FPS	$\sim$ 40%	0.861	0.842	0.364	0.247	0.699	54.24	0.764
Ours	1 FPS	36.59%	**0.932**	**0.925**	**0.383**	**0.142**	0.722	**53.15**	0.784
128 frames: learned compression on Qwen2.5-VL-3B									
Qwen2.5-VL-3B	128	100%	0.542	0.664	0.286	**0.000**	0.253	105.92	1.048
+ FlexSelect	128	$\sim$ 40%	0.533	0.642	0.280	**0.000**	0.270	80.92	**0.832**
+ QTSplus	128	$\sim$ 40%	0.747	0.684	0.200	0.319	0.588	67.55	0.982
+ Ours	128	37.07%	**0.918**	**0.913**	**0.391**	**0.000**	**0.711**	**64.14**	0.985

Table 7: Performance comparison on VANE-Bench. † denotes results reported by VAD-R1.

Method	Size	Ret.	SORA	OpenSORA	RG2	VideoLCM	MS-T2	Avenue	Ped1	Ped2
HAWK†	7B	–	24.64	52.00	44.00	36.54	50.00	36.36	36.67	38.89
VAD-R1	7B	100.00	50.00	**94.00**	68.00	75.00	**83.33**	78.79	63.33	**80.56**
VAU-R1	3B	100.00	**58.70**	82.00	**84.00**	75.96	81.25	**84.85**	**70.00**	72.22
Ours	3B	36.96	51.45	86.00	76.00	**76.92**	77.08	66.67	56.67	69.44

on real-world surveillance subsets, indicating that complex surveillance scenarios remain challenging under aggressive compression.

6 Conclusion

We present AnomMerge, a time-preserving token compression framework for generative video anomaly understanding. AnomMerge reduces spatial redundancy through adaptive spatial merging while preserving observed temporal units and positional correspondence. We further introduce CA-tIoU and IDC to complement conventional overlap metrics for evaluating temporal localization behavior. On Vad-Reasoning, AnomMerge retains only 36.59% of the original visual tokens while improving anomaly recognition and reducing interval-dispersion contraction compared with VAU-R1. These results demonstrate the effectiveness of preserving temporal observations while compressing spatial redundancy for efficient VAU.

Ethics Statement

We study video anomaly understanding on existing benchmarks. Surveillance applications raise concerns about privacy, bias, and false accusations. Predictions and explanations may be inaccurate; deployment requires human oversight and safeguards for privacy and fairness.

Reproducibility Statement

Section 5.1 describes the data protocol, model configuration, training settings, and evaluation metrics. Appendix A provides the aggregation, threshold prediction, optimization, and compression details. We will publicly release the AnomMerge code to support reproduction of our experiments.

AI Use Statement

Generative AI assisted with English editing, notation checks, and LaTeX formatting. The authors are responsible for the manuscript, code, and reported results, including their accuracy and originality.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *arXiv preprint arXiv:2602.19248*, 2026.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18793–18803. IEEE, 2024.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. *arXiv preprint arXiv:2602.08024*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22654–22663. IEEE, 2025.

Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational lmms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3123–3140, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 3302–3310, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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 *Forty-second International Conference on Machine Learning*, 2025.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *Advances in neural information processing systems*, 38:118486–118518, 2026a.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Li Shen, Wenqi Ren, Yong Xu, and Xiaochun Cao. Advancing adaptive multi-stage video anomaly reasoning: A benchmark dataset and method. *arXiv preprint arXiv:2601.10165*, 2026b.

Minseok Kang, Minhyeok Lee, Jungho Lee, Minjung Kim, Donghyeong Kim, Dayeon Lee, Heeseung Choi, Ig-jae Kim, and Sangyoun Lee. Ott-vid: Optimal transport temporal token compression for video large language models. *arXiv preprint arXiv:2605.11803*, 2026.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8592–8603. IEEE, 2025a.

Siyou Li, Huanan Wu, Juexi Shao, Yinghao Ma, Yujian Gan, Yihao Luo, Yuwei Wang, Dong Nie, Lu Wang, Wenqing Wu, et al. Seeing the forest and the trees: Query-aware tokenizer for long-video multimodal language models. *arXiv preprint arXiv:2511.11910*, 2025b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025c.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *Advances in Neural Information Processing Systems*, 38:64426–64455, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection-a new baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545. IEEE, 2018.

Xinxin Liu, Shiwei Gan, Xiao Liu, Yafeng Yin, Lei Xie, and Sanglu Lu. Infomerge: Information-aware token compression for efficient video large language models. *arXiv preprint arXiv:2606.02161*, 2026.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34:12786–12797, 2021.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488. IEEE, 2018.

Syed Hesham Syed Ariff, Yun Liu, Guolei Sun, Jing Yang, Henghui Ding, Xue Geng, and Xudong Jiang. Stac: Selective spatiotemporal aggregation and compression for video reasoning segmentation. In *European Conference on Computer Vision*, pp. 473–492. Springer, 2026.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001. IEEE, 2025.

Xiaodong Wang, Zhirong Wu, Langling Huang, Yuxi Zheng, and Peixi Peng. Incentivizing versatile video reasoning in mllms via data-efficient reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5444–5454, 2026a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026b.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8679–8688. IEEE, 2025.

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*, 2026.

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

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6543–6554, 2020.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853. IEEE, 2025b.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. Flexselect: Flexible token selection for efficient long video understanding. *arXiv preprint arXiv:2506.00993*, 2025c.

Zihao Zhao, Shengting Cao, and Muchao Ye. Srvau-r1: Enhancing video anomaly understanding via reflection-aware learning. *arXiv preprint arXiv:2602.01004*, 2026.

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.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025b.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, et al. Timelens2: Generalist video temporal grounding with multimodal llms. *arXiv preprint arXiv:2607.17423*, 2026.

A Additional Implementation Details

A.1 Spatiotemporal state-space aggregation details

We use the input and output notation of Section 4.1: $\mathbf{F}_{\mathrm{e}}, \mathbf{F}_{\mathrm{st}} \in \mathbb{R}^{T\times N\times d}$. Let $\mathbf{F}_{\mathrm{s}} \in \mathbb{R}^{T\times N\times d}$ denote the intermediate spatial features. The spatial and temporal parameter sets are θ_{s} and θ_{t}, with $\theta = (\theta_{\mathrm{s}}, \theta_{\mathrm{t}})$ denoting the full contextualizer parameters.

Selective state-space formulation. At scan step ν, for one channel of a Mamba mixer's expanded representation, let $u_\nu \in \mathbb{R}$ be the scan input and $\mathbf{h}_\nu \in \mathbb{R}^{d_{\mathrm{h}}}$ its state. The discrete selective recurrence is (Gu & Dao, 2023; Dao & Gu, 2024)

$$\mathbf{h}_\nu = \bar{\mathbf{A}}_\nu \mathbf{h}_{\nu-1} + \bar{\mathbf{b}}_\nu u_\nu, \qquad v_\nu = \mathbf{c}_\nu^\top \mathbf{h}_\nu + \eta u_\nu. \tag{9}$$

Here $\bar{\mathbf{A}}_\nu \in \mathbb{R}^{d_{\mathrm{h}}\times d_{\mathrm{h}}}$ is the structured discretized transition, $\bar{\mathbf{b}}_\nu, \mathbf{c}_\nu \in \mathbb{R}^{d_{\mathrm{h}}}$ are input/readout coefficients, and η is a learned skip coefficient. The discretized coefficients depend on the input through selective parameterization, while $\eta \in \mathbb{R}$ is independent of the input and $v_\nu \in \mathbb{R}$ is the channel output. Each scan starts from $\mathbf{h}_0 = \mathbf{0}$ and traverses N spatial positions or T temporal units. The state dimension d_{h} is distinct from the token feature dimension d. Equation 9 describes the state update; projections, convolutions, and gating complete the mixer.

Spatial aggregation. Let $\mathcal{R}$ reverse the sequence axis, preserving the channel order. For each unit t, flatten the native spatial grid in row-major order and apply RMSNorm (Zhang & Sennrich, 2019) before spatial aggregation:

$$\begin{aligned} \mathbf{U}_t &= \mathrm{RMSNorm}_{\theta_{\mathrm{s}}}(\mathbf{F}_{\mathrm{e}}[t,:,:]), \\ \mathbf{F}_{\mathrm{s}}[t,:,:] &= \mathbf{F}_{\mathrm{e}}[t,:,:] + [\mathcal{G}_{\mathrm{s}}^{\rightarrow}(\mathbf{U}_t) + \mathcal{R}(\mathcal{G}_{\mathrm{s}}^{\leftarrow}(\mathcal{R}(\mathbf{U}_t)))]\,\mathbf{W}_{\mathrm{s}}. \end{aligned} \tag{10}$$

The branch operators $\mathcal{G}_{\mathrm{s}}^{\rightarrow}, \mathcal{G}_{\mathrm{s}}^{\leftarrow}$ map $\mathbf{U}_t \in \mathbb{R}^{N\times d}$ to $\mathbb{R}^{N\times d_{\mathrm{inner}}}$, where $d_{\mathrm{inner}} = 2d$. The shared output projection is $\mathbf{W}_{\mathrm{s}} \in \mathbb{R}^{d_{\mathrm{inner}}\times d}$. They share the input projection but use separate convolution and state-space parameters. Each branch includes convolution, selective scanning, and gating. Aligned branch outputs are summed before the shared output projection and residual addition, following the released STAC implementation. These operations are repeated independently across temporal units.

Temporal aggregation. For a temporal sequence $\mathbf{V} \in \mathbb{R}^{T\times d}$, define the residual Mamba2 block $\mathcal{T}_{\theta_{\mathrm{t}}} : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times d}$ (Dao & Gu, 2024)

$$\mathcal{T}_{\theta_{\mathrm{t}}}(\mathbf{V}) = \mathbf{V} + \mathrm{Mamba2}_{\theta_{\mathrm{t}}}(\mathrm{RMSNorm}_{\theta_{\mathrm{t}}}(\mathbf{V}))\,. \tag{11}$$

STAC uses a causal temporal scan. For offline VAU, we instead apply this block in both directions at every fixed grid location n:

$$\begin{aligned} \mathbf{V}_n &= \mathbf{F}_{\mathrm{s}}[:,n,:], \\ \mathbf{F}_{\mathrm{st}}[:,n,:] &= \tfrac{1}{2}\left[\mathcal{T}_{\theta_{\mathrm{t}}}(\mathbf{V}_n) + \mathcal{R}(\mathcal{T}_{\theta_{\mathrm{t}}}(\mathcal{R}(\mathbf{V}_n)))\right]. \end{aligned} \tag{12}$$

The temporal passes share all block parameters and are averaged after their residual additions. Locations correspond by grid index, without object tracking. We use one spatial block and one temporal block, each with inner width $2d$ and state dimension $d_{\mathrm{h}} = 16$. The resulting $\mathbf{f}_{t,n} = \mathbf{F}_{\mathrm{st}}[t,n,:]$ supplies contextual features to the merging rule. The aggregation preserves all TN tokens and incorporates past and future observations before spatial grouping.

A.2 Threshold implementation and optimization

Implementation configuration. The shared coefficient predictor uses an MLP with dimensions $2 \to 64 \to 64 \to 1$ and ReLU hidden activations, with initial outputs near 0.1. LoRA adapters use rank 16 and a scaling factor of 32; training settings follow Section 5.1.

Threshold optimization. Discrete grouping decisions are detached from the computation graph. To train the threshold predictor with the SFT loss, we apply the STE to each pooled token:

$$\delta_t = \tau_t - \mathrm{sg}(\tau_t), \qquad \widetilde{\mathbf{z}}_{t,C} = \mathbf{z}_{t,C} + \delta_t \mathbf{1}_d, \tag{13}$$

where sg denotes stop-gradient and $\mathbf{1}_d$ is the all-ones vector. Since $\delta_t = 0$ in the forward pass, token values are preserved while surrogate gradients reach the predictor through τ_t.

Compression regularization. Let $K_t = |\mathcal{C}_t|$ be the retained token count in temporal unit t. Using the similarity statistics μ_t and σ_t from Section 4.2, we define

$$\mathcal{L}_{\text{comp}} = \frac{1}{T}\sum_{t=1}^{T}\left[1 - \frac{1}{2}(1 - \mu_t + \sigma_t)\right]\left(\frac{\text{sg}(K_t)}{\sqrt{N}} + \delta_t\right). \tag{14}$$

This term encourages compression through surrogate gradients while treating K_t as fixed. Setting $\lambda_c = 0$ disables the regularizer but retains the STE for SFT.

A.3 Spatial grouping and compression algorithm

Let H_g and W_g denote the spatial grid height and width, with $N = H_g W_g$. Tokens are indexed in row-major order, so the zero-based row and column coordinates of token $n \in \{1, \ldots, N\}$ are

$$r(n) = \lfloor (n-1)/W_g \rfloor, \qquad c(n) = (n-1) \bmod W_g. \tag{15}$$

For a candidate union U, its bounding-box height and width in grid cells are

$$h(U) = \max_{n\in U} r(n) - \min_{n\in U} r(n) + 1, \qquad w(U) = \max_{n\in U} c(n) - \min_{n\in U} c(n) + 1. \tag{16}$$

The group center is $\mathbf{m}_{t,C} = |C|^{-1}\sum_{n\in C}\mathbf{f}_{t,n}$ and is recomputed from all member features after each accepted merge. Bounding the spatial extent constrains the enclosing box without requiring the group itself to be rectangular.

Algorithm 1 AnomMerge Forward Token Compression

Input: encoder features $\mathbf{F}_{\text{e}} \in \mathbb{R}^{T\times N\times d}$, original temporal coordinates $\rho_{1:T}$, grid dimensions H_g, W_g, group limits $m_{\max}, b_{\max}$, contextualizer and predictor parameters θ, ϕ.
Output: compressed tokens $\mathbf{Z}$ and positional coordinates $\mathbf{P}$.

Compute $\mathbf{F}_{\text{st}}$ with contextualizer θ and set $\mathbf{f}_{t,n} \leftarrow \mathbf{F}_{\text{st}}[t, n, :]$.
List grid neighbor edges $\mathcal{E}$: horizontal then vertical, each in row-major order.
for $t = 1, \ldots, T$ **do**
  Initialize singleton groups $\mathcal{C}_t \leftarrow \{\{1\}, \ldots, \{N\}\}$ with centers $\mathbf{m}_{t,\{n\}} \leftarrow \mathbf{f}_{t,n}$.
  if $\mathcal{E} \neq \varnothing$ **then**
    Compute neighbor similarities $a_t(i,j) \leftarrow \cos(\mathbf{f}_{t,i}, \mathbf{f}_{t,j})$ for $(i,j) \in \mathcal{E}$.
    Compute their mean μ_t and sample standard deviation σ_t.
    Predict k_t from $[\mu_t, \sigma_t]$ using the shared MLP with ϕ; set $\tau_t \leftarrow \mu_t + k_t\sigma_t$.
    for $(i,j) \in \mathcal{E}$ in stable descending order of $a_t(i,j)$ **do**
      if $a_t(i,j) < \tau_t$ **then continue**.
      Find $C, C' \in \mathcal{C}_t$ with $i \in C$ and $j \in C'$.
      if $C = C'$ **then continue**; $U \leftarrow C \cup C'$.
      if $|U| > m_{\max}$ or $\max\{h(U), w(U)\} > b_{\max}$ **then continue**.
      if $\cos(\mathbf{m}_{t,C}, \mathbf{m}_{t,C'}) < \tau_t$ **then continue**.
      Replace C, C' with U in $\mathcal{C}_t$ and recompute its center from all member features.
    end for
  end if
  for $C \in \mathcal{C}_t$ **do**
    Select the lower-middle member by sorted index as spatial anchor $n^*_{t,C}$.
    Mean-pool the group: $\mathbf{z}_{t,C} \leftarrow |C|^{-1}\sum_{n\in C}\mathbf{f}_{t,n}$.
    Preserve source coordinates: $\mathbf{p}_{t,C} \leftarrow (\rho_t, r(n^*_{t,C}), c(n^*_{t,C}))$.
  end for
end for
Sort groups by $(t, n^*_{t,C})$; stack $\mathbf{z}_{t,C}$ and $\mathbf{p}_{t,C}$ in this order into $\mathbf{Z}$ and $\mathbf{P}$.
return $\mathbf{Z}, \mathbf{P}$.

B Additional Experiments

B.1 Effect of input frame count

Using Qwen2.5-VL-3B-Instruct with VAD-R1-style SFT, we compare 8, 16, 32, 64, and 128 uniformly sampled frames without token compression. Each model is trained and evaluated with the same frame count to examine its effect on anomaly understanding and memory demand. Checkpoint selection uses the evaluation set, so these results serve as an exploratory comparison.

Table 8: Effect of input frame count on Qwen2.5-VL-3B with VAD-R1-style SFT.

Frames	Acc. ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑	R@0.5 ↑
8	0.818	0.820	0.394	0.590	0.642	0.629
16	0.850	0.849	0.395	0.553	0.668	0.656
32	0.861	0.861	0.400	0.691	0.676	0.663
64	0.863	0.864	**0.405**	0.503	0.675	0.661
128	**0.875**	**0.872**	0.400	**0.399**	**0.689**	**0.679**

As shown in Table 8, increasing the frame count from 8 to 128 raises F1 from 0.820 to 0.872 and reduces IDC from 0.590 to 0.399. Figure 4(a) also shows fewer predictions spanning nearly the entire video, although their proportion remains high at 83.6% with 128 frames. Denser sampling therefore improves recognition and partially alleviates interval collapse, but does not resolve it. CA-tIoU peaks at 64 frames, showing that additional frames do not consistently improve localization.

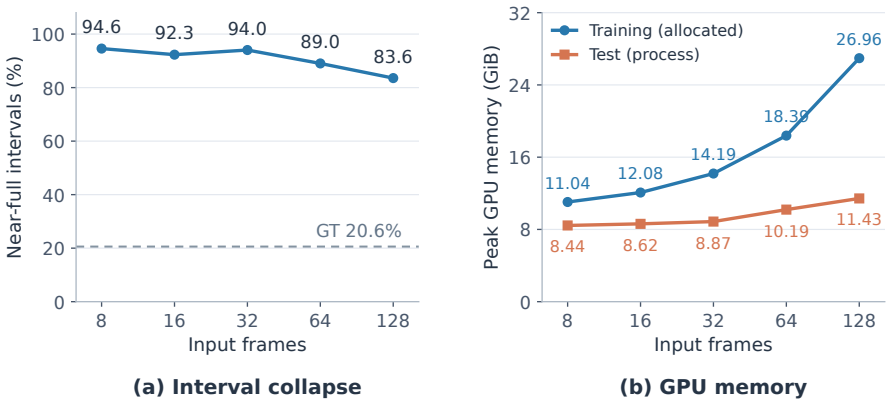

Figure 4: Effect of frame count on (a) predictions spanning nearly the full video and (b) GPU memory. The dashed line shows the ground truth proportion.

Figure 4(a) counts intervals starting within the first 5% and ending within the last 5% of the video, among valid predicted abnormal intervals. In Figure 4(b), training memory rises from 11.04 to 26.96 GiB and inference memory from 8.44 to 11.43 GiB. Both use cached visual features and exclude visual encoding; training reports peak allocated memory, whereas inference reports sampled process memory. These measurements show the growing memory cost of denser sampling and motivate compressing spatial redundancy while preserving temporal coverage.

B.2 Mamba contextualization

Using Qwen2.5-VL-3B, we compare AnomMerge without Mamba, with spatial or temporal contextualization alone, and with both. Each enabled axis uses bidirectional scans, while all variants retain the same adaptive spatial merging and positional anchoring. Table 9 reports the resulting token retention and anomaly understanding performance.

Spatial contextualization raises F1 from 0.846 to 0.911 and CA-tIoU from 0.352 to 0.370. Adding temporal context further improves them to 0.925 and 0.383 at similar retention (36.18% versus 36.59%). Temporal contextualization alone improves recognition but provides no localization gain in CA-tIoU or mIoU. The full model achieves the best recognition and localization scores and the lowest IDC, supporting the use of both spatial and temporal context before merging.

Table 9: Effect of spatial and temporal Mamba contextualization in AnomMerge.

Variant	**Spatial**	**Temporal**	Ret.	Acc ↑	F1 ↑	CA-tIoU ↑	IDC ↓	mIoU ↑
(a)	✗	✗	39.37	0.870	0.846	0.352	0.285	0.714
(b)	✓	✗	36.18	<u>0.920</u>	<u>0.911</u>	<u>0.370</u>	<u>0.153</u>	<u>0.720</u>
(c)	✗	✓	36.75	0.891	0.872	0.350	0.198	0.713
Ours	✓	✓	36.59	**0.932**	**0.925**	**0.383**	**0.142**	**0.722**

C QUALITATIVE COMPARISONS

Figure 5 compares VAD-R1, VAU-R1, and AnomMerge on two anomaly understanding cases. In the wrong-way driving example, both baselines describe normal traffic and miss the oncoming vehicle entering the camera vehicle's lane. AnomMerge identifies the vehicle moving against traffic and returns an anomaly interval, distinguishing the abnormal event from routine traffic.

In the falling-objects example, the baselines detect an anomaly but misinterpret its cause: VAD-R1 describes a weapon, while VAU-R1 attributes the debris to vandalism. AnomMerge instead identifies falling objects, consistent with the annotation. Together, these cases illustrate improvements in anomaly recognition and event interpretation, complementing the quantitative results.



Figure 5: Qualitative results on wrong-way driving (top) and falling objects (bottom).