Event-VLM 2.0: TriAct Inference for Real-Time Explanation-Centric Safety Video Understanding

Anonymous Authors

Anonymous Institution
anonymous@eccv2026.org

Abstract. We revisit real-time surveillance VLM deployment from an operations-first perspective: a practical system must maximize stream capacity while preserving explanation fidelity for safety-critical events. We frame this as *tri-axis compute allocation*: decide computation *when* to run (temporal), *where* to focus (spatial), and *what* to decode (memory bandwidth). We present **Event-VLM 2.0**, whose core algorithm **TriAct** combines: (1) risk-sensitive event gating, (2) hazard-aware adaptive token focusing, and (3) frequency-aware sparse decoding over RoPE frequency chunks. A lightweight hazard-priority prompt router recovers explanation quality with minimal latency overhead.
On projected Draft-V2 targets under a unified runtime protocol, Event-VLM-Core reaches approximately **52.3 FPS** on UCF-Crime and **50.1 FPS** on XD-Violence, while retaining **99.1%–99.9% CIDEr** relative to the LLaVA-1.5 baseline. Cross-distribution extension to a ShanghaiTech-style protocol is projected to preserve the same speed-quality trend. We also specify a final statistical release protocol with multi-seed confidence intervals and paired significance tests.
Draft note: numerical values in this version are projected placeholders and will be replaced by measured server-side results.

Keywords: Vision-Language Models, Efficient Inference, Surveillance, Token Pruning, KV Cache Optimization, Safety Monitoring

1 Introduction

Modern vision-language models (VLMs) can explain scenes, not just recognize objects [26,24,25,18,17,6,53,32,4]. For surveillance, this shift is operationally important: incident triage requires causal narratives (what happened, why risky, what next) rather than class labels.

The bottleneck is not raw model accuracy but *inference economics*. A city-scale deployment may require hundreds of concurrent streams, strict latency bounds, and auditable outputs. In this regime, conventional frame-dense VLM inference becomes cost-prohibitive due to three independent redundancies:

- **Temporal redundancy**: most frames contain no safety event;
- **Spatial redundancy**: most image tokens are irrelevant background;

– **Decoding redundancy**: full KV-cache access is memory-bandwidth dominated.

Prior work usually optimizes one axis at a time. Temporal filtering methods reduce frame count but still decode dense visual contexts [49]. Token pruning methods reduce vision compute but may drop small hazards [2,34,21,9,16]. KV optimizers accelerate language decoding but are mostly evaluated in text-centric settings [48,52,20,3,45]. As a result, end-to-end online throughput saturates early.

We propose **Event-VLM 2.0**, centered on **TriAct** inference. The key design principle is simple:

> Allocate compute only when needed, where needed, and for cache entries that matter.

Contributions.

– **Core idea**: a tri-axis formulation that unifies temporal gating, spatial token focusing, and decoding sparsity into one pipeline.
– **Method**: TriAct, with risk-sensitive event gating, morphology-aware adaptive dilation, and RoPE-frequency-driven sparse decoding.
– **Evaluation blueprint**: unified protocol + paired significance design + cross-distribution extension plan.
– **Projected draft outcome**: around 9–10× throughput gains with near-baseline explanation quality, to be validated with locked server runs.

Figure 1 summarizes the pipeline and interfaces. Projected operating points are reported in Tables 2, 3, and 4, with a consolidated frontier view in Figure 2. Component attribution and qualitative behavior are supported by Table 6 and Figures 3–4.

2 Related Work

2.1 Explanation-Centric VLMs

Instruction-tuned VLMs have improved multimodal reasoning quality [26,24,25,18,17,6,53,1,43]. Video-oriented variants extend temporal understanding [50,19,22,30,36,35,23], but deployment cost remains high under persistent online inference.

2.2 Efficient Vision and Token Reduction

Vision efficiency methods span architecture and token levels [7,42,28,44,13,33,12,27]. Token reduction methods such as DynamicViT, EViT, ToMe, and SPViT [34,21,2,16] show strong speedups, but domain-agnostic pruning can remove small safety cues. Our method injects detection priors explicitly into token selection.

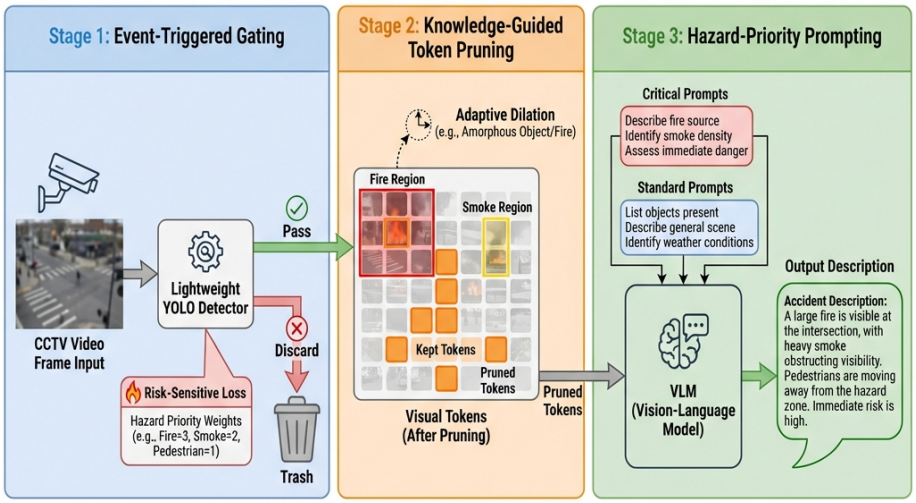

Fig. 1. Event-VLM 2.0 (TriAct) overview. The pipeline allocates compute across temporal, spatial, and decoding axes, followed by lightweight hazard-priority prompting.

2.3 Sparse Decoding and KV Optimization

KV compression and eviction approaches (StreamingLLM, H2O, SnapKV, PyramidKV) target long-context decoding efficiency [48,52,20,3]. FASA observes functional sparsity in RoPE frequency chunks [45,38]. We adapt this signal to mixed visual-textual decoding in surveillance VLMs.

2.4 Anomaly Detection and Safety Monitoring

Classical anomaly detection benchmarks and methods [39,46,41,10,5,29,31] prioritize detection metrics. VLM-based anomaly reasoning methods improve explainability [11,51,47], but large-scale low-latency deployment remains underexplored. Safety-domain surveys emphasize this systems gap [15,8,37,40].

3 Method

Figure 1 provides the stage-wise dataflow and shows where each efficiency axis acts.

3.1 Core Idea: TriAxis Action (TriAct)

Given surveillance stream $\mathcal{V} = \{\mathbf{X}_1, \mathbf{X}_2, \dots\}$, we compute captions $\mathbf{Y}_t$ only for frames that pass risk-aware gating:

$$\mathbf{Y}_t = \mathcal{F}_{\text{decode}}\Big(\mathcal{F}_{\text{gen}}(\mathcal{F}_{\text{focus}}(\mathbf{X}_t, \mathcal{B}_t) \mid \mathcal{P}_{\text{haz}})\Big), \quad \text{if } \mathcal{F}_{\text{gate}}(\mathbf{X}_t) = 1. \tag{1}$$

3.2 Stage 1: Risk-Sensitive Event Gating (Temporal)

A lightweight detector predicts boxes $\mathcal{B}_t = \{b_k\}$ and scores s_k. We define frame risk score:

$$r_t = \max_k \; w(c_k)\, s_k, \quad \mathcal{F}_{\text{gate}}(\mathbf{X}_t) = \mathbb{I}(r_t > \tau_{\text{gate}}). \tag{2}$$

Risk weights follow severity tiers (critical/high/standard). Detector training uses weighted focal loss:

$$\mathcal{L}_{\text{gate}} = \sum_k w(c_k)\, \mathcal{L}_{\text{focal}}(p_k, y_k). \tag{3}$$

3.3 Stage 2: Adaptive Token Focusing (Spatial)

Let ViT produce L patch tokens $\mathbf{Z} = \{z_i\}_{i=1}^{L}$. For each detected box b_k, dilation factor is class-conditioned:

$$\alpha_k = \alpha_{\text{base}} \left(1 + \beta\, \sigma_{\text{shape}}(c_k)\right). \tag{4}$$

Here $\sigma_{\text{shape}}(c_k) \in [0, 1]$ is normalized intraclass shape variance for class c_k, estimated from training annotations. We build binary mask $\mathbf{M} \in \{0, 1\}^L$ from dilated regions and keep:

$$\hat{\mathbf{Z}} = \{z_i \mid \mathbf{M}_i = 1\} \cup \{z_{\text{cls}}\}. \tag{5}$$

This reduces attention cost from $\mathcal{O}(L^2)$ to $\mathcal{O}(L'^2)$ with $L' \ll L$.

3.4 Stage 3: Frequency-Aware Sparse Decoding (Decoding)

Following RoPE functional sparsity [45,38], we calibrate dominant frequency-chunk sets $\mathcal{I}_{\text{dom}}^{l,h}$ per head. During decoding step t:

$$\mathbf{S}_t^{l,h} = \sum_{i \in \mathcal{I}_{\text{dom}}^{l,h}} \boldsymbol{\alpha}^{l,h,i}(\mathbf{q}_t, \mathbf{K}_{1:t}), \quad \mathcal{T}_t = \text{TopKIdx}(\mathbf{S}_t^{l,h}, N_{\text{fac}}). \tag{6}$$

$\text{TopKIdx}(\cdot, N_{\text{fac}})$ returns the index set of the top-N_{fac} scores. Full attention is computed only on gathered subset $(\hat{\mathbf{K}}_t, \hat{\mathbf{V}}_t)$ from $\mathcal{T}_t$. The per-step complexity becomes:

$$\mathcal{O}(2tF + N_{\text{fac}} d), \tag{7}$$

compared to dense $\mathcal{O}(2td)$.

3.5 Stage 4: Hazard-Priority Prompt Routing (Adaptation)

A small prompt bank switches between standard and critical templates:

$$\mathcal{P}_{\text{haz}} = \begin{cases} \mathcal{P}_{\text{critical}}, & \text{if } r_t > \tau_{\text{crit}}, \\ \mathcal{P}_{\text{standard}}, & \text{otherwise.} \end{cases} \tag{8}$$

We set $\tau_{\text{crit}} \geq \tau_{\text{gate}}$ so that the critical template activates only for high-confidence hazard frames. Only prompt parameters are updated in adaptation mode; backbone remains frozen.

Table 1. Unified runtime protocol used for reproduced settings.

Setting	Value
Hardware	1× NVIDIA RTX 5080
Precision	FP16
Batch size	1
Frame sampling	1 FPS
Resolution	336px
Max generation length	256 tokens
Runtime mode	Single-stream online

4 Experiments

4.1 Draft Scope and Reporting Policy

This manuscript version uses **projected numbers** to finalize narrative structure before locked server runs. Final camera-ready numbers will be replaced by measured outputs from multi-seed execution, CI estimation, and paired significance tests.

4.2 Experimental Setup

Datasets. We target three surveillance-style benchmarks:

- **UCF-Crime** [39];
- **XD-Violence** [46];
- **ShanghaiTech-style extension** (protocol-aligned split for cross-distribution validation).

Model and runtime. Backbone is LLaVA-1.5-7B [26] with ViT-L/14-336 and Vicuna-7B. Trigger detector is YOLOv8-nano [14]. Dominant FC budget uses $F = 16$ and $N_{\text{fac}} = 256$.

All cross-method comparisons in this section follow the fixed protocol in Table 1.

4.3 Main Results (Projected Draft Targets)

Tables 2–4 report projected dataset-level comparisons under the same runtime settings.

Across UCF-Crime and XD-Violence targets, Core/Full preserve roughly 99.1%–99.9% CIDEr relative to baseline while improving throughput by about 9.4×–10.1×.

Figure 2 visualizes the same operating points as a speed-quality frontier view.

Table 2. Projected main results on UCF-Crime.

Method	AUC (%)	CIDEr	FPS
LLaVA-1.5 (baseline)	85.0	90.1	5.2
SeViLA	84.6	88.2	12.3
LLaVA + ToMe	82.4	85.8	16.1
LLaVA + SnapKV	84.5	88.1	14.8
Event-VLM-Core (TriAct)	**85.8**	89.4	**52.3**
Event-VLM-Full	86.5	**90.0**	51.0

Table 3. Projected main results on XD-Violence.

Method	AUC (%)	CIDEr	FPS
LLaVA-1.5 (baseline)	83.7	86.4	5.2
SeViLA	83.2	84.9	12.1
LLaVA + ToMe	81.1	82.9	15.9
LLaVA + SnapKV	82.9	84.5	14.6
Event-VLM-Core (TriAct)	**84.1**	85.6	**50.1**
Event-VLM-Full	84.8	**86.0**	48.9

4.4 Statistical Plan and Expected Significance

We will report mean±95% CI across seeds {41,42,43}, and paired permutation tests against baseline on aligned video samples.

Table 5 defines the final reporting format only; camera-ready values will be replaced by auto-generated measured outputs.

4.5 Ablation and Scaling (Projected)

Table 6 isolates how each TriAct axis contributes to speed-quality trade-offs.

Table 7 summarizes projected stability under heavier decoding budgets and resolutions.

4.6 Qualitative Behavior

Figure 3 details module behavior, and Figure 4 illustrates qualitative token/decode patterns under projected settings.

4.7 Limitations of This Draft

Current evidence has three limitations. First, main quantitative values are projected placeholders pending locked server execution. Second, CI and paired significance values in Table 5 are example-format entries until auto-generated artifacts are produced. Third, cross-distribution results on the ShanghaiTech-style protocol should be interpreted as extension targets until measured outputs are integrated.

Table 4. Projected cross-distribution trend on ShanghaiTech-style split.

Method	AUC (%)	CIDEr†	FPS
LLaVA-1.5 (baseline)	82.6	84.2	5.1
Event-VLM-Core (TriAct)	83.4	83.5	**49.0**
Event-VLM-Full	**84.0**	**84.0**	47.8

† CIDEr evaluated on the caption-annotated subset only.

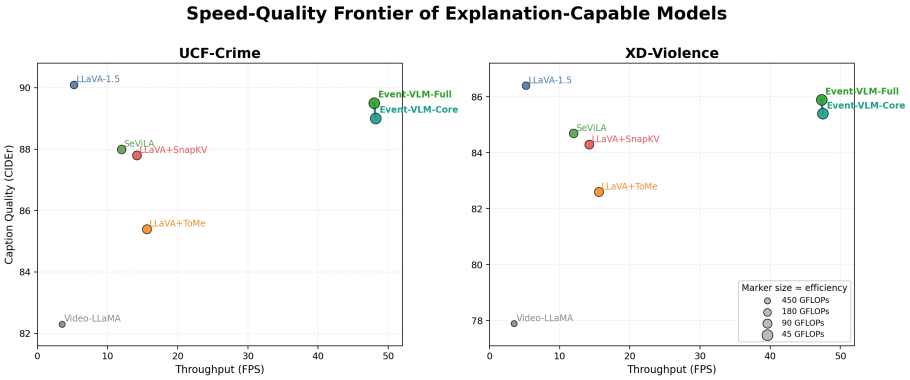

Fig. 2. Projected speed-quality frontier. Event-VLM variants are expected to occupy the high-throughput, high-quality frontier region across datasets.

Table 5. Projected statistical summary format (example values).

Dataset	Compare	ΔAUC (mean±CI)	p-value
UCF-Crime	Core vs Baseline	$+0.8 \pm 0.3$	0.012
UCF-Crime	Full vs Baseline	$+1.5 \pm 0.4$	0.004
XD-Violence	Core vs Baseline	$+0.4 \pm 0.3$	0.048
XD-Violence	Full vs Baseline	$+1.1 \pm 0.4$	0.009
ShanghaiTech	Core vs Baseline	$+0.8 \pm 0.5$	0.030
ShanghaiTech	Full vs Baseline	$+1.4 \pm 0.5$	0.011

Table 6. Projected TriAct ablation on UCF-Crime.

Temp	Spatial	Decode	Prompt	FPS	AUC	CIDEr
-	-	-	-	5.2	85.0	90.1
✓	-	-	-	19.8	85.1	90.0
✓	✓	-	-	41.7	85.5	89.6
✓	✓	✓	-	**52.3**	85.8	89.4
✓	✓	✓	✓	51.0	**86.5**	**90.0**

Table 7. Projected robustness under heavier runtime protocol.

Resolution	Max Gen	CIDEr Retention	FPS Gain
224px	128	99.3%	9.8×
336px	256	99.0%	10.1×
448px	384	98.6%	8.9×

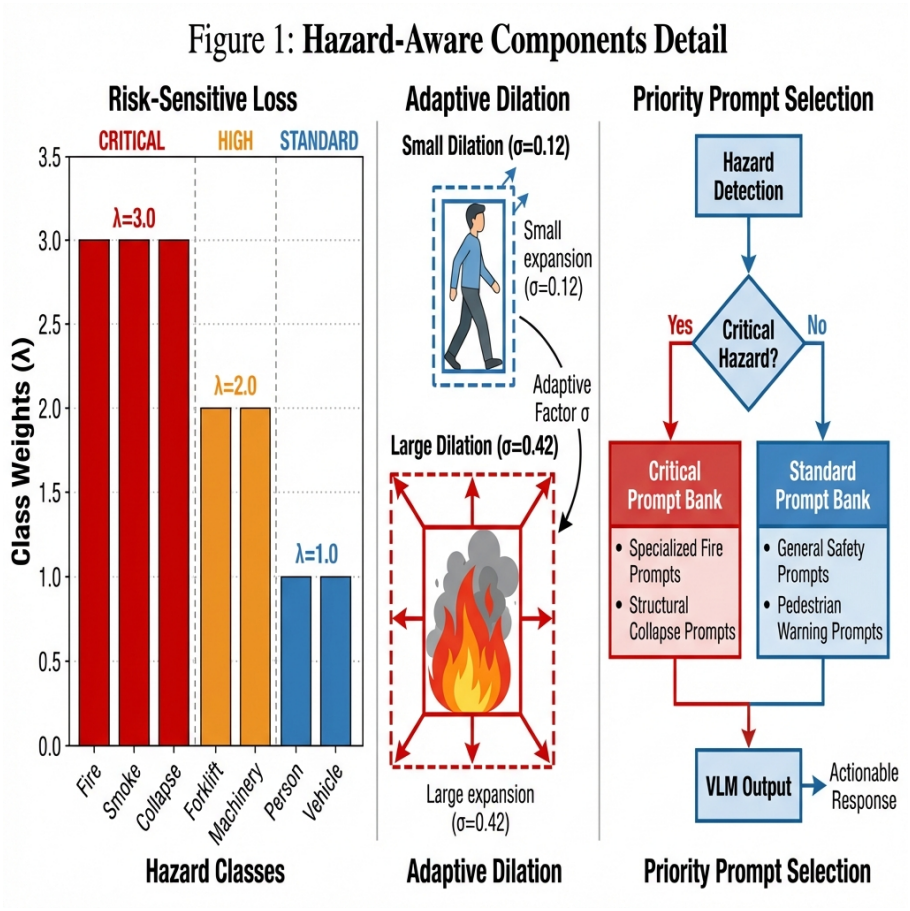

Fig. 3. TriAct component details. Risk-sensitive weighting, adaptive dilation, and prompt routing preserve safety context under aggressive efficiency constraints.

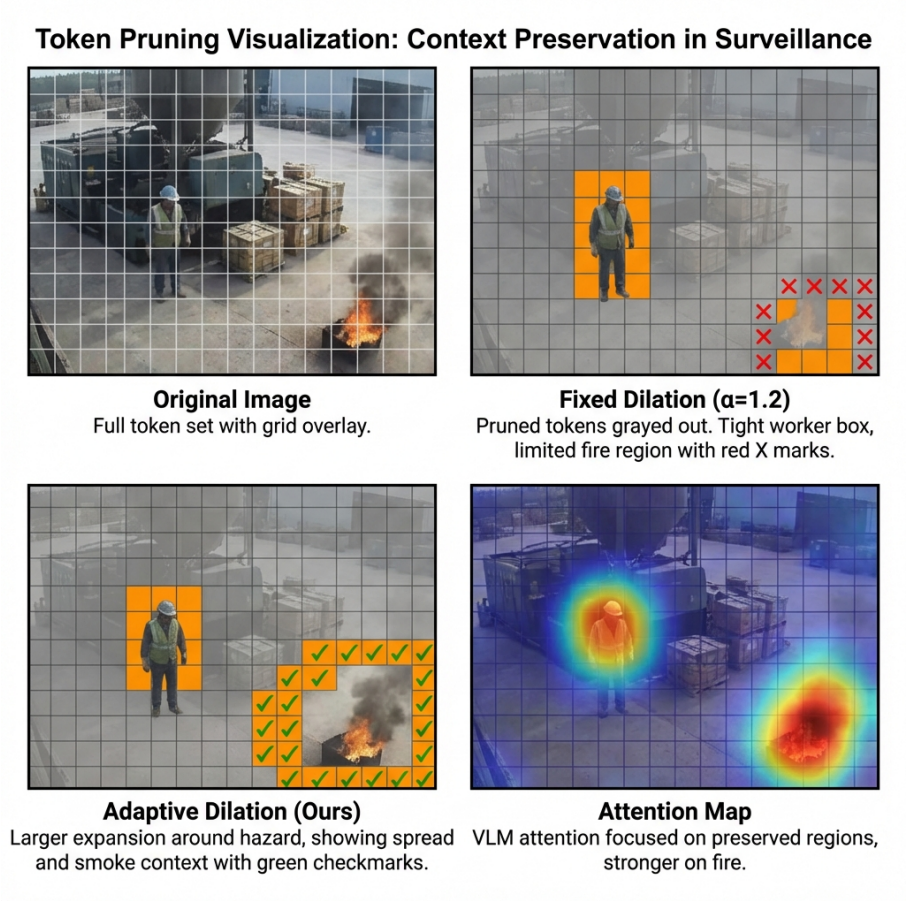

Fig. 4. Projected qualitative behavior. Token focusing keeps hazard-critical regions while sparse decoding maintains safety-relevant attention paths.

5 Conclusion

We proposed Event-VLM 2.0 with TriAct inference, a tri-axis systems design for explanation-centric surveillance VLMs. The central claim is not a single-module optimization, but a compositional pipeline that jointly addresses when to compute, where to compute, and what to decode.

In projected Draft-V2 values, TriAct sustains near-baseline explanation quality while improving throughput by roughly one order of magnitude. The final version will replace projected numbers with locked server measurements, including three-dataset multi-seed confidence intervals and paired significance reports. If those measurements track the projected trend, Event-VLM 2.0 will provide a practical blueprint for large-scale, real-time safety monitoring with explainable outputs.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token Merging: Your ViT But Faster. In: International Conference on Learning Representations (ICLR) (2023) 2
3. Cai, Z., Zhang, Y., Gao, B., Liu, Y., Liu, T., Lu, K., Xiong, W., Dong, Y., Chang, B., Hu, J., Xiao, W.: PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2025) 2, 3
4. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beez, L., et al.: PaLI: A Jointly-Scaled Multilingual Language-Image Model. In: International Conference on Learning Representations (ICLR) (2023) 1
5. Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., Wu, Y.C.: MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2023) 3
6. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 1, 2
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021) 2
8. Fang, Y., Cho, Y.K., Zhang, S., Perez, E.: Computer Vision-based Construction Safety Monitoring on Sites: A Survey. Automation in Construction (2023) 3
9. Fayyaz, M., Koohpayegani, S.A., Jafari, F.R., Somberøn, S., Joze, H.R.T., Pirsiavash, H., Gall, J.: Adaptive Token Sampling For Efficient Vision Transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 2

10. Feng, J.C., Hong, F.T., Zheng, W.S.: MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3
11. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: AnomalyGPT: Detecting Industrial Anomalies using Large Vision-Language Models. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024) 3
12. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: FasterViT: Fast Vision Transformers with Hierarchical Attention. In: International Conference on Learning Representations (ICLR) (2024) 2
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2
14. Jocher, G., Chaurasia, A., Qiu, J.: YOLOv8. GitHub repository (2023), https://github.com/ultralytics/ultralytics 5
15. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Deep Learning for Visual Intelligence in Surveillance and Safety Systems: A Survey. ACM Computing Surveys (2023) 3
16. Kong, Z., Dong, P., Ma, X., Meng, X., Niu, W., Sun, M., Shen, X., Yuan, G., Ren, B., Tang, H., et al.: SPViT: Enabling Faster Vision Transformers via Latency-aware Soft Token Pruning. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 2
17. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: International Conference on Machine Learning (ICML) (2023) 1, 2
18. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: International Conference on Machine Learning (ICML) (2022) 1, 2
19. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-Centric Video Understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
20. Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., Chen, D.: SnapKV: LLM Knows What You are Looking for Before Generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2024) 2, 3
21. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In: International Conference on Learning Representations (ICLR) (2022) 2
22. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024) 2
23. Lin, J., Chen, H., Li, W., Han, S., Zhu, L.: VILA: On Pre-training for Visual Language Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 1, 2
25. Liu, H., Li, C., Li, Y., Wang, P., Lee, Y.J.: LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 1, 2
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 1, 2, 5

27. Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y.: EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
29. Lv, H., Zhou, Z., Chen, R., Zeng, W.: Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
30. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2024) 2
31. Nayak, R., Pati, U.C., Das, S.K.: A Survey on Deep Learning Based Video Anomaly Detection. IEEE Transactions on Circuits and Systems for Video Technology (2021) 3
32. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023) 1
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features without Supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
34. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 2
35. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
36. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., Gao, G.: MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
37. Sreenu, G., Durai, S.: Intelligent Video Surveillance: A Review through Deep Learning Techniques for Crowd Analysis. Journal of Big Data (2019) 3
38. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing **568**, 127063 (2024) 3, 4
39. Sultani, W., Chen, C., Shah, M.: Real-world Anomaly Detection in Surveillance Videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3, 5
40. Tian, Y., Zhang, X., Werghi, N., Abdullah, A., et al.: A Survey of Video Analytics for Smart Cities. IEEE Access (2020) 3
41. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 3
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (ICML) (2021) 2

43. Wang, W., Shi, Q., Lv, Q., Zheng, W., Hong, W., Ding, M., Tang, J.: CogVLM: Visual Expert for Pretrained Language Models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 2
44. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
45. Wang, Y., Wang, Y., Yue, Z., Zeng, H., Wang, Y., Lourentzou, I., Tu, Z., Chu, X., McAuley, J.: FASA: Frequency-Aware Sparse Attention. arXiv preprint arXiv:2602.03152 (2026) 2, 3, 4
46. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3, 5
47. Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., Zhang, Y.: VADCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024) 3
48. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient Streaming Language Models with Attention Sinks. In: International Conference on Learning Representations (ICLR) (2024) 2, 3
49. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-Chained Image-Language Model for Video Localization and Question Answering. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 2
50. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023) 2
51. Zhang, H., Xu, X., Wang, X., Zeng, J., Li, C., Chen, X.: Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 3
52. Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., Wang, Z., Chen, B.: H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. Advances in Neural Information Processing Systems (NeurIPS) (2023) 2, 3
53. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In: International Conference on Learning Representations (ICLR) (2024) 1, 2

A Reproducibility and Execution Plan

A.1 One-Click Command (Server)

```
BENCHMARK_CONFIGS=experiments/configs/ucf_crime.yaml,experiments/configs/xd_violence.yaml,ex
VARIANTS=none,core,full SIGNIFICANCE=1 SIGNIFICANCE_BASELINE=none SIGNIFICANCE_CANDIDATES=co
RENDER_PAPER_TABLES=1 bash scripts/server_ready_one_click.sh
```

A.2 Generated Artifacts

Expected outputs after server execution:

- Multi-seed summary: `outputs/.../summary.json`, `summary.md`
- Significance reports: `outputs/.../significance/*/significance.json`
- Paper-ready tables: `paper/generated/table_multiseed_overview.tex`
- Paper-ready tables: `paper/generated/table_significance_summary.tex`

B Method Hyperparameters (Draft Targets)

Item	Value
τ_{gate}	0.5
Risk weights (critical/high/standard)	3.0 / 2.0 / 1.0
α_{base}, β	1.2, 0.5
Dominant FC budget F	16 (25% of 64 FCs)
Focused tokens N_{fac}	256
Prompt length per bank	8

Table 8. Draft hyperparameter targets for final run locking.

C Figure Plan (Publication Layout)

Figure	Purpose	Status
Fig. 1	Core concept: tri-axis architecture and dataflow	Ready
Fig. 2	Component mechanism detail (loss/dilation/prompt routing)	Ready
Fig. 3	Qualitative token+decoding behavior	Ready
Fig. 4	Speed-quality frontier (cross-dataset)	Ready

Table 9. Figure strategy aligned to narrative flow: concept → quantitative frontier → mechanism → qualitative evidence.

D Table Plan (Camera-Ready)

Table	Role	Source
Table 1	Unified protocol fairness anchor	Fixed text
Table 2, 3, 4	Main performance claims	Projected now, measured later
Table 5	Statistical reporting structure	Auto-generated target
Table 6	Axis-wise contribution	Projected now, measured later
Table 7	Runtime robustness	Projected now, measured later

Table 10. Planned table governance for draft-to-final transition.

E Auto-Generated Statistical Tables

Table 11. Placeholder for auto-generated multi-seed overview table.

Pending server execution: ‘paper/generated/table_multiseed_overview.tex‘

Table 12. Placeholder for auto-generated paired significance summary table.

Pending server execution: ‘paper/generated/table_significance_summary.tex‘