Appendix

Contents

1 Implementation Details

We provide a live demonstration of our system at: `https://memovad2026.github.io/`.

1.1 Dataset Distributions

Figure 1: Visualization of sample frame sequences from the UCF-Crime and XD-Violence datasets. Anomalous regions are highlighted with red bounding boxes.

We evaluate our MemoVAD on two large-scale weakly supervised video anomaly detection benchmarks: UCF-Crime [Sultani *et al.*, 2018] and XD-Violence [Wu *et al.*, 2020]. The detailed statistics of the dataset splits and anomaly categories are provided below:

- **UCF-Crime:** The dataset consists of 1,900 untrimmed surveillance videos covering 13 real-world anomaly categories: *Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting,* and *Vandalism.* The dataset is divided into a training set with 800 normal and 810 anomalous videos, and a testing set with 150 normal and 140 anomalous videos.
- **XD-Violence:** This is a large-scale multi-modal dataset containing 4,754 videos collected from movies and games. It includes 6 anomaly categories: *Abuse, Car Accident, Explosion, Fighting, Riot,* and *Shooting.* The

Attribute	UCF-Crime	XD-Violence
Source Domain	Real-world Surveillance	Movies & Games
Modality	Visual Only	Audio-Visual
Anomaly Categories	13	6
Total Videos	1,900	4,754
Training Set	**1,610**	**3,954**
- *Normal Videos*	800	*Mixed / Unspecified*
- *Anomalous Videos*	810	
Testing Set	**290**	**800**
- *Normal Videos*	150	*Mixed / Unspecified*
- *Anomalous Videos*	140	

Table 1: Statistical comparison between UCF-Crime and XD-Violence datasets. Note that while XD-Violence provides audio signals, we only utilize the visual modality in this work.

training set comprises 3,954 videos, while the testing set contains 800 videos. Although the dataset includes audio, we only utilize visual data for fair comparison with vision-only baselines.

To facilitate a comprehensive understanding of the experimental setup, we summarize the key statistical specifications of both datasets in Table 1. The side-by-side comparison highlights the diversity in data sources (real-world surveillance vs. fictitious movies), the variation in dataset scale, and the distinct anomaly taxonomies inherent to each benchmark. For a qualitative perspective, we also provide visualizations of sample frame sequences with anomalous regions highlighted in Figure 1 of the Appendix.

1.2 Hyper-parameters

Figure 2: The physical testbed of the MemoVAD system deployed on NVIDIA Jetson AGX Orin Developer Kit.

We implement MemoVAD using PyTorch. Training is conducted on a high-performance NVIDIA A800 GPU, while inference is deployed on NVIDIA Jetson AGX Orin devices to validate efficiency on actual edge hardware, as illustrated in Fig. 2. We employ VideoMAE-Small [Tong *et al.*, 2022] pre-trained on Kinetics-400 as the backbone, keeping its parameters frozen during training. The input frames are resized to $H \times W = 224 \times 224$, and we extract $F = 16$

frames with a stride of $\tau = 4$ for each clip, resulting in a dimension of $D_{\text{stu}} = 384$. We use the AdamW optimizer with a weight decay of 10^{-4} and set the initial learning rate to 5×10^{-4} with a cosine annealing scheduler. The batch size and total epochs are set to 64 and 30, respectively. Regarding the DSM, we set the memory size $N = 2048$ and utilize Qwen3-VL-8B [Bai *et al.*, 2025] as the VLM to generate semantic information. For hyper-parameters, L in Eq. 4 is set to 9, while τ_{unc} and τ_{sim} in Eq. 13 are set to 0.5 and 0.75 respectively. m in Eq. 17 is set to 0.5 and λ_{sp} in Eq. 19 is set to 8×10^{-5}. The loss balancing weights in Eq. 21 are set to $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, and $\lambda_3 = 0.5$. For better readability, we summarize the key hyper-parameters in Table 2.

Hyper-parameter	**Value**
Input Resolution ($H \times W$)	224×224
Frames per Clip (F)	16
Temporal Stride (τ)	4
Student Dimension (D_{stu})	384
Context Window (L)	9
Optimizer	AdamW
Initial Learning Rate	5×10^{-4}
Weight Decay	10^{-4}
Batch Size	64
Total Epochs	30
Memory Size (N)	2048
Uncertainty Threshold (τ_{unc})	0.5
Similarity Threshold (τ_{sim})	0.75
Margin (m)	0.5
Sparsity Weight (λ_{sp})	8×10^{-5}
Loss Weights ($\lambda_1, \lambda_2, \lambda_3$)	1.0, 1.0, 0.5

Table 2: Summary of Hyper-parameters.

1.3 Baselines

We compare our proposed MemoVAD with the following state-of-the-art methods:

- **Sultani et al.** [Sultani *et al.*, 2018]: The pioneering work that formulates WSVAD as an MIL task, using C3D features and a ranking loss to distinguish normal and anomalous bags.
- **RTFM** [Tian *et al.*, 2021]: Proposes a feature magnitude learning approach that encourages large feature magnitudes for anomalies and small magnitudes for normal events.
- **CRFD** [Wu and Liu, 2021]: A method that improves feature discrimination by utilizing multi-scale temporal dependencies and feature relationships.
- **MSL** [Li *et al.*, 2022]: Introduces a multi-sequence learning framework to capture various temporal dependencies of anomalous events.
- **MGFN** [Chen *et al.*, 2023]: Combines magnitude-contrastive learning with a gloss feature module to better capture the motion patterns of anomalies.
- **UR-DMU** [Zhou *et al.*, 2023]: Utilizes dual memory units to separate the representations of normal and abnormal patterns, reducing the interference between them.
- **CLIP-TSA** [Joo *et al.*, 2023]: Leverages the zero-shot capability of CLIP with a temporal self-attention mechanism for anomaly detection.
- **TPWNG** [Yang *et al.*, 2024]: Enhances CLIP-based detection by generating text prompts that align with visual content using a prompt generation network.

- **VadCLIP** [Wu *et al.*, 2024]: Adapts CLIP for VAD by introducing a dual-branch adapter to align video features with anomaly-related text descriptions.
- **OVVAD** [Li *et al.*, 2025]: An open-vocabulary approach that attempts to detect unseen anomalies by generating descriptive text for video content.
- **EventVAD** [Shao *et al.*, 2025]: Focuses on modeling the complex temporal structure of anomalies as events rather than isolated frames.

2 Prompts and Demonstrations

2.1 Prompts

Role	Prompt Content
System	You are an intelligent video surveillance assistant designed to detect security anomalies.

User	Analyze the provided video frames.
	Focus strictly on these anomaly categories: [**Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Shoplifting, Stealing, Vandalism**].
	Reasoning Logic:
	1. If the event belongs to one of these categories, output 'true' for `is_anomaly` and provide a concise description.
	2. If the scene is normal (walking, working, standing), output 'false'.
	Output Format:
	Output strictly in JSON:
	{
	"is_anomaly": <boolean>,
	"description": "<string>"
	}

Table 3: The specific prompts used for the UCF-Crime dataset. The user prompt incorporates a category-guided constraint to reduce hallucinations.

Role	Prompt Content
System	You are an intelligent video surveillance assistant designed to detect security anomalies.

User	Analyze the provided video frames.
	Focus strictly on these anomaly categories: [**Abuse, Car Accident, Explosion, Fighting, Riot, Shooting**].
	Reasoning Logic:
	1. If the event belongs to one of these categories, output 'true' for `is_anomaly` and provide a concise description.
	2. If the scene is normal (walking, working, standing), output 'false'.
	Output Format:
	Output strictly in JSON:
	{
	"is_anomaly": <boolean>,
	"description": "<string>"
	}

Table 4: The specific prompts used for the XD-Violence dataset. The category constraint is adapted to the 6 distinct classes defined in the benchmark.

To fully leverage the generalized visual reasoning capabilities of the remote VLM (e.g., Qwen3-VL) while strictly adhering to the bandwidth and latency constraints of the edge-cloud collaborative environment, we employ a minimalist yet structured prompt engineering strategy.

Our prompt design follows three key principles:

1. **Role Contextualization:** We explicitly define the system persona as an "intelligent video surveillance assistant." It primes the VLM to adopt a professional tone and focus on security-relevant details rather than general image captioning.

2. **Closed-Set Category Constraint:** General-purpose VLMs often suffer from "hallucinations" or over-sensitivity (e.g., interpreting "running for a bus" as "escaping"). To mitigate this, we inject the specific anomaly definitions of the target dataset (UCF-Crime or XD-Violence) directly into the user instruction. It forces the model to perform a closed-set classification task combined with open-ended reasoning.

3. **Machine-Readable Formatting:** To facilitate seamless integration with the downstream Dynamic Semantic Memory (DSM), we enforce a strict JSON output format. It allows the edge system to deterministically parse the `is_anomaly` boolean flag for gating decisions and extract the `description` text for semantic embedding updates without complex regex post-processing.

The specific prompt templates adapted for the UCF-Crime and XD-Violence datasets are detailed in Table 3 and Table 4, respectively.

2.2 Demonstrations

Figure 3 presents qualitative results from the VLM inference phase, visualizing the model's response to specific clips from the UCF-Crime and XD-Violence benchmarks. These demonstrations validate the effectiveness of the prompt engineering strategies detailed in Section 3.1, particularly the adherence to closed-set constraints and machine-readable formatting.

As shown in the top row (UCF-Crime: *Arrest024_x264*), the VLM demonstrates robust contextual understanding. It initially characterizes the scene as "normal traffic" successfully ignoring non-threatening dynamic background elements. However, once the event of interest occurs, the model immediately toggles the *is_anomaly* flag to *true* and provides a granular description of the "arrest situation," noting specific details such as "law enforcement officers" and "weapons drawn."

Similarly, the bottom row (XD-Violence: *v=251...*) demonstrates the capacity of the VLM to disambiguate between high-motion normal activities and actual violence. The VLM correctly distinguishes the allowable aggression of a hockey game from a violent altercation ("Fighting occurring"), generating semantic descriptions that are rich enough to update the downstream Dynamic Semantic Memory (DSM).

3 Additional Experiments

3.1 Effects with Different VLMs

Teacher VLM	AUC (%)↑	Latency (s)↓	Type
LLaVA-Next-7B	88.65	0.85	Open-Source
InternVL-Chat-V1.5	89.10	1.20	Open-Source
GPT-4o-mini	90.12	0.77	Close-Source
GPT-4o	**90.25**	2.55	Close-Source
Qwen3-vl-32B	89.63	1.33	Open-Source
Qwen3-vl-30B-A3B	89.70	1.27	Open-Source
Qwen3-vl-235b-a22b	90.18	2.26	Open-Source
Qwen3-VL-8B (Ours)	89.45	**0.65**	Open-Source

Table 5: Comparison of different Teacher VLMs on UCF-Crime.

Visualization of VLM inference results for the **Arrest024_x264** clips from the UCF-Crime benchmark.	
	{
	"is_anomaly": false,
	"description": "The scene shows normal traffic with multiple vehicles, including a police car, but no visible security anomalies from the specified categories.
	}
	 
	{
	"is_anomaly": true,
	"description": "Arrest situation involving multiple law enforcement officers surrounding vehicles with weapons drawn."
	}

Visualization of VLM inference results for the **v=251___mEwZA__#1_label_B1-0-0** clips from the XD-Violence benchmark.	
	{
	"is_anomaly": false,
	"description": "Normal scene of a hockey game with players skating and playing."
	}
	 
	{
	"is_anomaly": true,
	"description": "Fighting occurring during a hockey game, involving multiple players and referees."
	}

Figure 3: **Visualization of VLM Inference Results.** Qualitative examples from the UCF-Crime (top) and XD-Violence (bottom) benchmarks.

While we employ Qwen3-VL-8B [Bai *et al.*, 2025] as our default teacher model due to its balance of performance and efficiency, MemoVAD is VLM-agnostic. To verify the robustness of our framework and justify our choice, we evaluate the performance of MemoVAD when powered by different State-of-the-Art VLMs on the UCF-Crime dataset.

As shown in Table 5, we compare Qwen3-VL against LLaVA-Next [Liu *et al.*, 2023] (open-source), InternVL-Chat [Chen *et al.*, 2024] (open-source), and GPT-4o [Achiam *et al.*, 2023] (closed-source, accessed via API).

The results indicate that:

1. **Performance Consistency:** Across the board, all evaluated modern VLMs provide robust semantic guidance, consistently propelling the student model performance to exceed 88% AUC. This universality provides compelling evidence that the efficacy of MemoVAD is intrinsic to its collaborative architecture rather than being contingent upon the capabilities of a specific VLM. It confirms that our framework is robust and adaptable to various foundation models.

2. **Superiority of GPT-4o:** Unsurprisingly, the much larger GPT-4o achieves the highest AUC of 90.25%. Notably, the GPT-4o-mini variant also delivers a competitive AUC of 90.12% with a latency of 0.77 seconds. However, the inevitable network latency caused by data transmission to OpenAI servers and the associated API costs render these closed-source models unsuitable for real-time edge applications.

3. **Efficiency of Qwen3-VL:** Our selected Qwen3-VL-8B model strikes an exceptional balance between performance

and computational cost. It achieves a competitive AUC of 89.45% while maintaining the lowest latency of 0.65 seconds among the evaluated baselines. The superior efficiency-to-performance ratio makes it the optimal candidate for our edge-cloud collaborative framework, ensuring rapid anomaly detection without the heavy resource burden typically associated with larger VLMs.

3.2 Parameter Sensitivity Analysis

To investigate the interplay between the loss components in Eq. 21, we conduct a grid search over the distillation weight λ_1 and smoothness weight λ_2. We explicitly visualize the joint sensitivity of these two parameters because they represent competing objectives: semantic fidelity (distillation) versus temporal consistency (smoothness), making their equilibrium critical for model performance. In contrast, the gating weight λ_3 functions as a relatively independent uncertainty regularizer. Since preliminary experiments indicated that λ_3 is robust around 0.5, we fix $\lambda_3 = 0.5$ to isolate the analysis of the trade-off between λ_1 and λ_2.

Figure 4 illustrates the resulting AUC performance surface. We observe that the performance initially improves as both weights increase, confirming that semantic knowledge from the teacher and temporal consistency constraints are essential for accurate detection. The AUC peaks at 89.45% when $\lambda_1 = 1$ and $\lambda_2 = 1$. Notably, the performance drops when the weights deviate significantly from this equilibrium. A small λ_1 fails to transfer sufficient semantic guidance, while an excessively large λ_1 causes the student to over-mimic the teacher, potentially overriding the primary MIL classification objective. Similarly, while λ_2 effectively suppresses noise, an overly large λ_2 leads to over-smoothing, causing the model to miss short-duration anomalous events. The 3D surface demonstrates that our method achieves an optimal trade-off at the selected values.

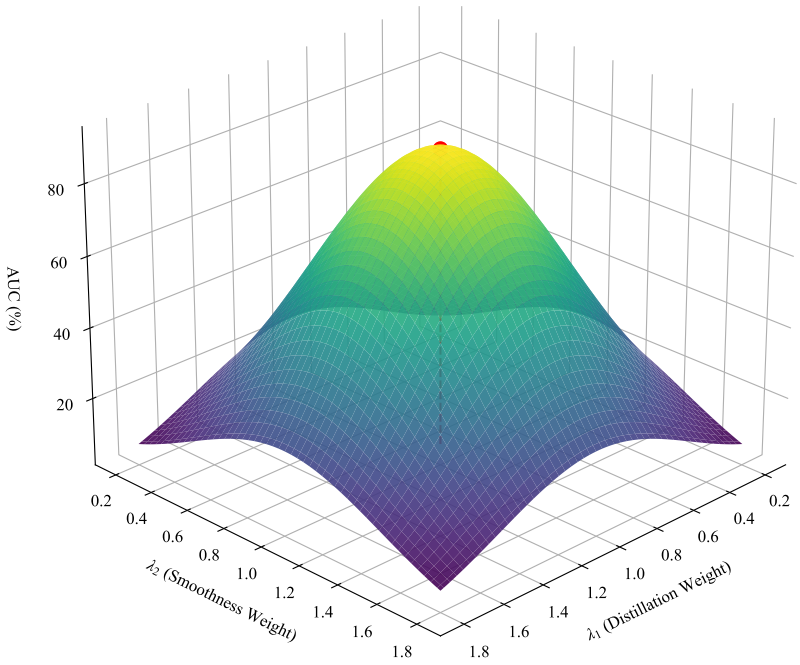

Figure 4: Joint sensitivity analysis of λ_1 and λ_2 on UCF-Crime. The z-axis represents the AUC score. We fix the gating weight $\lambda_3 = 0.5$ and vary the distillation weight λ_1 and smoothness weight λ_2.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[Chen *et al.*, 2023] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 387–395, 2023.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[Joo *et al.*, 2023] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023.

[Li *et al.*, 2022] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403, 2022.

[Li *et al.*, 2025] 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 Computer Vision and Pattern Recognition Conference*, pages 29203–29212, 2025.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Shao *et al.*, 2025] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2586–2595, 2025.

[Sultani *et al.*, 2018] 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*, pages 6479–6488, 2018.

[Tian *et al.*, 2021] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[Wu and Liu, 2021] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021.

[Wu *et al.*, 2020] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[Wu *et al.*, 2024] 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, pages 6074–6082, 2024.

[Yang *et al.*, 2024] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18899–18908, 2024.

[Zhou *et al.*, 2023] 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, pages 3769–3777, 2023.