Fine-VAD: Towards Fine-Grained Video Anomaly Detection via Progressive Cross-Granularity Learning

Supplementary Material

In this supplementary material, we provide additional discussions and analyses to support the claims made in the main paper:

- **Computational Efficiency.** Section A reports the computational efficiency of Fine-VAD, including inference speed (FPS) and GFLOPs, demonstrating the practicality and scalability of our framework.
- **Hyperparameter Sensitivity.** Section B presents sensitivity analyses on key hyperparameters, including the temporal window size, the top-T aggregation parameter in MIL-Align, and the prompt design used in the MLLM-based variant.
- **Ablation on Pseudo-Labeling Strategies.** Section C evaluates alternative pseudo-label construction strategies for generating macro-categories, confirming the robustness of our intermediate-level supervision.
- **Ablation on Intermediate-Level Depth.** Section D analyzes the effect of varying the depth of intermediate alignment, including inserting multiple pseudo-supervision stages.
- **Ablation on XD-Violence.** Section E provides additional ablation results on XD-Violence to further verify the effectiveness of the proposed design under a different dataset setting.
- **Additional Comparison with Related Work.** Section F provides a detailed discussion comparing our progressive cross-granularity learning paradigm with related strategies such as label distillation and curriculum learning, highlighting the unique aspects of our approach.
- **Additional Experimental Results.** Section G provides extended experiments, including (1) per-category anomaly detection performance, (2) results on ShanghaiTech, and (3) coarse-grained detection results.
- **Failure Cases and Analysis.** Section H presents representative failure examples and analyzes the main challenges that remain for Fine-VAD.

A. Computational Efficiency

We assess the computational efficiency of Fine-VAD in terms of inference time and computational cost. For a fair comparison, we evaluate our method against recent VLM-based approaches using the same CLIP backbone (ViT-B/16). As shown in Table 1, Fine-VAD achieves an inference time of 23.12 ms and a computational cost of 34.61G MACs, while substantially outperforming all baselines in fine-grained VAD performance.

Table 1. Comparison of inference time, computational cost (MACs), and average fine-grained mAP on the UCF-Crime dataset.

Method	Inference Time ↓	MACs ↓	Average mAP (%) ↑
VadCLIP [17]	22.30ms	29.17G	6.68
ExVAD [4]	15.37ms	12.04G	10.15
Fine-VAD (Ours)	23.12ms	34.61G	**14.99**

Although the computational cost of Fine-VAD is moderately higher than that of VadCLIP [17] and ExVAD [4], the overall overhead remains modest. This increase is justified by a clear performance gain, as Fine-VAD improves the average mAP from 10.15% to 14.99% over the strongest baseline. The efficiency of our design mainly comes from two aspects: the CLIP backbone remains frozen during both training and inference, and the additional modules are lightweight, consisting only of the video adapter and alignment layers.

Notably, an inference time of 23.12 ms corresponds to approximately 43.25 FPS, which is well above the frame rate of many real-world surveillance videos. Since practical surveillance systems typically operate at around 15–25 FPS, Fine-VAD is suitable for real-time deployment in practical scenarios.

B. Hyperparameter Sensitivity

We analyze the sensitivity of Fine-VAD to three factors: the temporal window size used in the video adapter, the top-T parameter in the MIL-Align strategy, and the prompt template used in the MLLM-based variant. These factors affect temporal context modeling, category-level similarity aggregation, and textual query formulation, respectively.

Analysis of Temporal Window Size. We first study the effect of the temporal window size in the video adapter. As shown in Table 2, performance generally improves as the window size increases to an appropriate range, while overly large windows may weaken local motion cues and hurt fine-grained discrimination. In the main experiments, we follow prior work [17] and set the window size to 8 for UCF-Crime and 64 for XD-Violence, which also correspond to the best average mAP in our experiments.

Analysis of Top-T in MIL-Align. For the top-T parameter in MIL-Align, we adopt the default value $T = 16$ following prior work [17]. Since MIL-Align is a fixed inference mechanism inherited from earlier studies, our goal is to main-

Table 2. Sensitivity to temporal window size. The adopted settings, i.e., 8 for UCF-Crime and 64 for XD-Violence, are consistent with prior work and achieve the best average mAP (%) in our experiments.

Window Size	4	8	16	32	48	64	80
UCF-Crime	13.74	**14.99**	13.82	12.65	11.58	10.63	8.94
XD-Violence	22.63	23.41	26.98	29.59	30.66	**31.87**	27.14

Table 3. Sensitivity to prompt templates on UCF-Crime and XD-Violence.

Prompt Template	UCF-Crime	XD-Violence
a video of {category}	14.99	31.87
a video showing {category}	14.87	32.15
an anomalous event of {category}	15.05	31.68
Mean ± Std	**14.97 ± 0.10**	**31.90 ± 0.25**

Table 4. Preliminary sensitivity analysis of the top-T parameter in MIL-Align. Performance remains relatively stable across a range of values. $T = 16$ is highlighted as the default setting used in our main experiments and prior work.

Top-T	4	8	**16**	24	32
UCF-Crime (mAP %)	14.79	14.90	**14.99**	15.08	15.11
XD-Violence (mAP %)	31.58	31.96	**31.87**	31.82	31.75

tain fair comparability with existing baselines rather than exhaustively tune this parameter. Nevertheless, we conduct a small-scale preliminary analysis to examine its influence. As shown in Table 4, varying T within a reasonable range leads to only minor performance fluctuations, indicating that Fine-VAD is relatively robust to this choice. Although slightly better mAP values are observed at larger T on UCF-Crime, the default setting $T = 16$ provides strong and stable performance across both datasets, while remaining consistent with prior works [4, 17].

Analysis of Prompt Sensitivity. We further evaluate the sensitivity of Fine-VAD to different prompt templates in the MLLM-based variant. As shown in Table 3, the performance remains highly stable across three alternative prompt formulations, with very small standard deviations of 0.10% on UCF-Crime and 0.25% on XD-Violence. These results suggest that the proposed progressive cross-granularity paradigm is not tied to a specific prompt template and can generalize well across different textual expressions.

Overall, our analysis confirms that the hyperparameter choices used in the main experiments are both principled and empirically optimal, consistent with prior work and validated by ablation results.

Table 5. Ablation study on different pseudo-labeling strategies for the intermediate level on UCF-Crime and XD-Violence. Average mAP (%) is reported. Results demonstrate that our paradigm is robust to the specific choice of grouping method.

Pseudo-Labeling Strategy	**UCF**	**XD**
Hierarchical Clustering	14.81	31.04
Manual Grouping	**15.29**	**33.04**
Random Grouping	12.15	28.63
K-Means Clustering (Ours)	**14.99**	**31.87**

C. Ablation on Pseudo-Labeling Strategies

In our proposed Fine-VAD, we employ K-Means clustering on CLIP-derived text embeddings of ground-truth category labels to construct intermediate-level pseudo macro-categories. A natural question is how sensitive the overall performance of our paradigm is to the specific choice of this grouping strategy. To investigate this, we conduct an ablation study comparing our default K-Means approach with three alternative strategies for creating $K = 4$ macro-categories on the both datasets.

The compared strategies are as follows:

- **K-Means Clustering (Ours):** The default method used in the main paper, which automatically groups categories based on similarity in the CLIP embedding space.
- **Hierarchical Clustering:** An alternative method using agglomerative clustering with Ward linkage.
- **Manual Grouping:** A human-curated grouping based on expert knowledge. For example, *Robbery*, *Burglary*, *Shoplifting*, and *Stealing* are grouped as "Theft-related." This represents a form of idealized semantic clustering.
- **Random Grouping:** A strong negative baseline where the all categories are randomly partitioned into four groups. This setting tests the robustness of the paradigm under non-semantic intermediate supervision.

As shown in Table 5, the manually defined semantic grouping achieves the highest mAP, serving as an idealized upper bound based on expert knowledge. While this result confirms the effectiveness of semantically coherent macro-categories, the performance gain over our default K-Means strategy is modest. We adopt K-Means in our main framework because it requires no manual annotations and still achieves near-optimal results. Hierarchical clustering yields similarly strong performance, indicating that the paradigm is not overly sensitive to the specific grouping algorithm, as long as the resulting macro-categories preserve some degree of semantic coherence.

The most surprising and insightful finding comes from the random grouping setting. Even when the intermediate-level groupings are entirely arbitrary and non-semantic, the model still significantly outperforms the variant without any

Table 6. Ablation on the depth of intermediate supervision on UCF. We report average mAP (%) against the number of intermediate layers and the number of clusters (K) used at each layer.

# of Intermediate Layers	K per Layer	UCF-Crime
0 (Coarse + Fine only)	–	10.93
1 (Default, main paper)	4	14.99
2	4, 6	15.14
3	4, 6, 8	15.18

Table 7. **Ablation study on hierarchical supervision levels** on the XD-Violence dataset.

ID	Hierarchical Level			AVG mAP%
	Coarse	Inter	Fine	
1			✓	26.96
2	✓		✓	29.53
3		✓	✓	30.24
4	✓	✓	✓	**31.87**

intermediate supervision. This strongly supports our core hypothesis: the intermediate stage primarily serves as a **structural regularizer**. The act of introducing coarse categorical boundaries, even noisy ones, helps pre-structure the feature space and provides an inductive scaffold that stabilizes learning under sparse fine-grained supervision.

In summary, this study demonstrates the robustness of our progressive paradigm. While semantically meaningful groupings can offer marginal improvements, the overall effectiveness of our approach stems from the cross-granularity learning itself.

D. Ablation on Intermediate-Level Depth

We further analyze our hierarchical design by studying the impact of the *depth* of intermediate supervision, which is the number of intermediate alignment layers. This study is conducted on UCF-Crime, where we compare our default single-layer model against variants with zero, two, and three intermediate layers. For deeper models, we create a progressively finer supervision path by incrementally increasing the number of clusters (K) at each added layer.

The results in Table 6 reveal a clear trend of diminishing returns. Introducing a single intermediate layer provides the most substantial benefit, boosting the AVG mAP by a remarkable **+4.06%** from 10.93% to 14.99%. However, adding a second layer yields a much smaller gain of **+0.15%**, and a third layer provides a negligible improvement of just **+0.04%**. This demonstrates that while a single intermediate stage plays an essential role in bridging the supervision gap between coarse and fine levels, additional layers provide limited incremental benefit. As deeper intermediate hierarchies tend to offer little new structural guidance and may introduce redundancy, the performance gain quickly saturates. Considering the substantial computational cost introduced by each extra layer, the trade-off becomes apparent. Therefore, we adopt a single intermediate layer in our final model, as it achieves an effective balance between performance and efficiency.

E. Ablation on XD-Violence

To complement the ablation study on UCF-Crime in the main paper, we further conduct the same analysis on XD-Violence. We primarily present ablations on UCF-Crime in the main paper because it contains a larger number of anomaly categories and thus provides a more challenging testbed for evaluating fine-grained supervision. As shown in Table 7, the observations on XD-Violence are highly consistent with those on UCF-Crime. Using only fine-level supervision yields an average mAP of 26.96%, while introducing coarse-level supervision improves the performance to 29.53%. Replacing coarse-level supervision with intermediate-level supervision further increases the average mAP to 30.24%, indicating that the pseudo macro-category alignment provides more effective guidance for fine-grained recognition. Combining all three levels achieves the best performance of 31.87%, confirming that coarse, intermediate, and fine supervision are complementary and mutually reinforcing. These results further verify that the effectiveness of the proposed progressive cross-granularity design generalizes across datasets.

F. Additional Comparison with Related Work

We provide a detailed discussion comparing our progressive cross-granularity learning paradigm with two related learning strategies: Label distillation and Curriculum learning.

Label Distillation vs. Ours. Label distillation [1, 3] is a knowledge transfer technique in which a compact *student* model is trained using probabilistic labels provided by a powerful pre-trained *teacher* model. While our use of coarse and intermediate supervision may appear analogous to offering softer guidance, the two approaches differ fundamentally in the source of supervision. Label distillation relies on an external teacher model to supply privileged information, or, in the case of self-distillation [5, 14], utilizes the model's own predictions as soft labels. In contrast, our multi-granularity supervision is derived intrinsically from the hierarchical semantic structure of the data itself. The coarse binary labels are ground truth, and the intermediate pseudo-labels are generated via unsupervised clustering over fine-grained labels. No external teacher model is involved. Moreover, the goals of the two methods are distinct: label distillation primarily targets model compression

Table 8. Per-category anomaly detection performance (mAUC) on UCF-Crime.

Method	Class													mAUC
	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fight	RoadAcc.	Robbery	Shoot	Shoplift	Steal	Vandalism	
CLIP [9]	57.37	80.65	93.72	80.83	74.34	90.31	83.54	87.46	70.22	63.99	71.21	45.49	66.45	74.28
ActionCLIP [12]	91.88	90.47	89.21	86.87	81.31	94.08	83.23	94.34	82.82	70.53	91.60	94.06	89.89	87.72
AnomalyCLIP [19]	75.03	94.56	96.66	94.80	90.08	94.79	88.76	93.30	86.85	87.45	89.47	97.00	89.78	90.66
Fine-VAD (Ours)	**92.15**	**95.32**	**97.12**	**95.67**	**91.45**	**95.34**	**89.73**	**95.12**	**89.76**	**88.54**	**92.34**	**97.45**	**93.12**	**93.32**

Table 9. Per-category anomaly detection performance (mAP) on XD-Violence.

Method	Class						mAP
	Abuse	CarAccident	Explosion	Fighting	Riot	Shooting	
CLIP zero-shot [9]	0.32	12.21	22.26	25.25	66.60	1.26	21.32
ActionCLIP [12]	2.73	25.15	55.28	58.09	87.31	12.87	40.24
AnomalyCLIP [19]	6.10	31.31	68.75	71.44	92.74	26.13	49.41
Fine-VAD (Ours)	**10.45**	**36.78**	**72.12**	**75.56**	**95.85**	**30.67**	**53.57**

or transfer, whereas our paradigm addresses the challenge of data scarcity in fine-grained tasks by constructing a structured learning path from generic to specific semantics.

Curriculum Learning vs. Our Paradigm. Curriculum learning [8, 13] aims to improve model training by organizing data in a order from easy to hard, based on task difficulty or sample uncertainty. The core idea is to gradually expose the model to increasingly complex examples, thereby facilitating stable and efficient convergence. At first glance, our progressive cross-granularity paradigm may appear similar, as it also introduces supervision in stages. However, the progression in our method is not defined by sample-level difficulty, but rather by the semantic granularity of the task itself. Specifically, our approach transitions from binary anomaly labels to intermediate pseudo-labels and then to fine-grained category labels, following a natural hierarchy embedded in the label space. Moreover, curriculum learning typically assumes a fixed target label space, whereas our method actively restructures the label space through unsupervised clustering to form intermediate-level guidance. Therefore, our paradigm constitutes a novel form of *task-level curriculum*, where the learning progression is driven by the hierarchical semantics of the task labels, rather than by sample-level difficulty.

G. Additional Experimental Results

G.1. Per-Category Anomaly Detection Performance

To provide a more detailed understanding of model behavior across different anomaly categories, we report per-category anomaly detection performance on UCF-Crime and XD-Violence in Tables 8 and 9, respectively. Following prior work [19], we adopt the mean Area Under the Curve (mAUC) for UCF-Crime and the mean Average Precision (mAP) for XD-Violence, as these metrics are consistent with the evaluation protocols commonly used in fine-grained VAD.

UCF-Crime contains a larger set of anomaly categories with dense frame-level annotations, making mAUC the standard metric for assessing per-category discrimination. In contrast, XD-Violence focuses on violence-related categories with highly imbalanced event durations, where mAP provides a more stable and interpretable measure of category-level detection quality.

Across both datasets, Fine-VAD achieves superior or best-in-class performance on nearly all categories, demonstrating the effectiveness of the proposed progressive cross-granularity learning paradigm in enhancing fine-grained anomaly discrimination. Notably, Fine-VAD substantially boosts performance on visually similar anomaly types (e.g., *Arson* vs. *Explosion*, *Fighting* vs. *Riot*), which aligns with our design goal of reducing inter-class confusion through hierarchical supervision.

G.2. Results on the ShanghaiTech Dataset

To further assess the generalization ability of Fine-VAD, we evaluate our framework on the ShanghaiTech dataset [6], a widely used benchmark originally designed for semi-supervised anomaly detection in campus scenes. Although ShanghaiTech is primarily used in a semi-supervised setting, it also provides category annotations for each anomalous event. This enables a per-category evaluation protocol similar to prior work [19]. Following this setting, we report the mAUC for each activity class.

Table 10. Per-category anomaly detection performance (mAUC) on the ShanghaiTech dataset.

Method	Class															mAUC
	Car	Chase	Circuit	Fall	Fight	Jump	Monocycle	Push	Robbery	Run	Skateboard	Stoop	Throw	Vaudeville	Vehicle	
CLIP [9]	61.65	77.88	5.95	61.73	79.37	23.68	77.78	63.36	37.71	54.39	76.15	8.47	44.10	65.97	27.08	51.02
ActionCLIP [12]	98.50	93.86	98.59	16.38	97.45	89.63	98.05	8.14	67.36	78.25	97.10	0.76	97.70	98.65	93.97	75.63
AnomalyCLIP [19]	98.08	96.66	97.97	96.69	98.03	95.48	86.89	97.99	95.00	97.95	97.29	98.62	96.50	96.97	96.79	96.46
Fine-VAD (Ours)	**99.12**	**97.45**	**99.05**	**97.88**	**98.76**	**96.72**	**98.34**	**98.12**	**96.85**	**98.45**	**98.67**	**99.10**	**97.88**	**98.45**	**97.92**	**97.85**

Table 11. **Comparison of coarse-grained video anomaly detection performance**. Methods are grouped by paradigm from top to bottom: weakly supervised, fine-grained, MLLM-based, and our proposed variants.

Method	**UCF (AUC%)**	**XD (AP%)**
UR-DMU [21] AAAI'23	86.75	82.41
MGFN [2] AAAI'23	86.98	80.11
PEL4VAD [7] TIP'24	86.76	85.59
FedCLIP [11] AAAI'25	85.07	74.23
RealAD [10] CVPR'18	84.14	75.18
AVVD [15] TMM'23	82.44	78.64
VadCLIP [17] AAAI'24	88.02	84.51
AnomalyCLIP [19] CVIU'24	86.36	78.51
OVVAD [16] CVPR'24	86.40	66.53
LAVAD [20] CVPR'24	80.28	62.01
ExVAD [4] ICML'25	<u>88.29</u>	<u>86.52</u>
VERA [18] CVPR'25	86.55	70.54
Fine-VAD (Ours)	**88.79**	**87.32**

ShanghaiTech covers a diverse range of activities, such as *Chasing*, *Fighting*, and *Throwing*, offering a complementary testbed for examining the robustness of fine-grained anomaly understanding under varied motion patterns and campus-specific contexts. As shown in Table 10, Fine-VAD achieves strong performance across almost all categories and consistently surpasses existing baselines, yielding the highest overall mAUC. These results further verify that the proposed progressive cross-granularity learning paradigm generalizes effectively across datasets with different scene distributions and anomaly features.

G.3. Coarse-Grained Detection Results.

To complement the fine-grained results reported in the main paper, we provide the coarse-grained anomaly detection results in Table 11. Fine-VAD also achieves the best performance under the coarse-grained setting, reaching 88.79% AUC on UCF-Crime and 87.32% AP on XD-Violence. These results indicate that the proposed progressive cross-granularity paradigm not only improves fine-grained anomaly classification, but also preserves strong coarse-grained anomaly detection capability. This suggests that the gains of Fine-VAD do not come at the expense of coarse-grained detection performance.

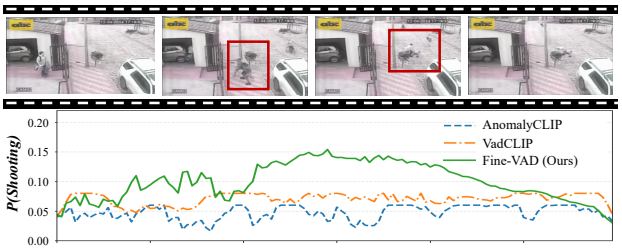

Figure 1. Representative failure case on a *shooting* event. All methods incorrectly classify the event as *fighting*, indicating the difficulty of recognizing short-duration and abrupt anomalies with subtle discriminative cues.

H. Failure Cases and Analysis

Despite the strong overall performance of Fine-VAD, some challenging cases remain difficult to classify correctly. As already reflected in the per-category results in Section F.1, the anomaly category *shooting* shows relatively lower performance than several other categories. We attribute this limitation partly to the fixed temporal sampling strategy, which may miss the key frames of short-duration and abrupt anomalies. As a result, the most discriminative visual evidence may not be fully captured, making fine-grained recognition more difficult.

Figure 1 presents a representative failure case on a *shooting* event. In this example, both baseline methods [1,2] and Fine-VAD incorrectly classify the event as *fighting*, while assigning a predicted probability below 0.15 to the correct *shooting* category. This example suggests that, although Fine-VAD improves fine-grained anomaly recognition overall, it still faces challenges when the target anomaly occurs abruptly and is visually similar to other human-interaction categories. These observations indicate that future improvements may benefit from more adaptive temporal sampling or stronger modeling of brief yet highly discriminative event moments.

References

[1] Peijun Bao, Zihao Shao, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, and Alex C. Kot. Omnipotent distillation with llms for weakly-supervised natural language video localization: When divergence meets consistency. In *AAAI*, pages 747–755, 2024. 3

[2] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton W. T. Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI*, 2023. 5

[3] Md. Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *ICLR*, 2025. 3

[4] 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 *ICML*, pages 25750–25761, 2025. 1, 2, 5

[5] Hyeonsu Jeong and Hye Won Chung. Rethinking self-distillation: Label averaging and enhanced soft label refinement with partial labels. In *ICLR*, 2025. 3

[6] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *ICCV*, pages 341–349, 2017. 4

[7] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.*, 33:4923–4936, 2024. 5

[8] Xiaofan Que and Qi Yu. Dual-level curriculum meta-learning for noisy few-shot learning tasks. In *AAAI*, pages 14740–14748, 2024. 4

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 4, 5

[10] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 5

[11] Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *AAAI*, pages 21017–21025, 2025. 5

[12] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *CoRR*, abs/2109.08472, 2021. 4, 5

[13] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4555–4576, 2022. 4

[14] Xucong Wang, Pengkun Wang, Shurui Zhang, Miao Fang, and Yang Wang. Multi-label self knowledge distillation. In *AAAI*, pages 21330–21338, 2025. 3

[15] Peng Wu, Xiaotao Liu, and Jing Liu. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25:1674–1685, 2023. 5

[16] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *CVPR*, pages 18297–18307, 2024. 5

[17] 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 *AAAI*, pages 6074–6082, 2024. 1, 2, 5

[18] Muchao Ye, Weiyang Liu, and Pan He. VERA: explainable video anomaly detection via verbalized learning of vision-language models. In *CVPR*, pages 8679–8688, 2025. 5

[19] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into CLIP latent space for video anomaly recognition. *Comput. Vis. Image Underst.*, 249:104163, 2024. 4, 5

[20] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *CVPR*, pages 18527–18536, 2024. 5

[21] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI*, pages 3769–3777. AAAI, 2023. 5