

Multi-Expert Ensemble Strategy for Weakly Supervised Video Anomaly Detection

Hongyu Yang, Wanru Xu , Zhenjiang Miao, Yi Tian, Ping Guo,
and Ruiying Yao



School of Computer Science and Technology, Beijing Jiaotong University

Outline

- ◆ Introduction
- ◆ Related Work
- ◆ Our Framework
- ◆ Experiments
- ◆ Conclusion

Introduction

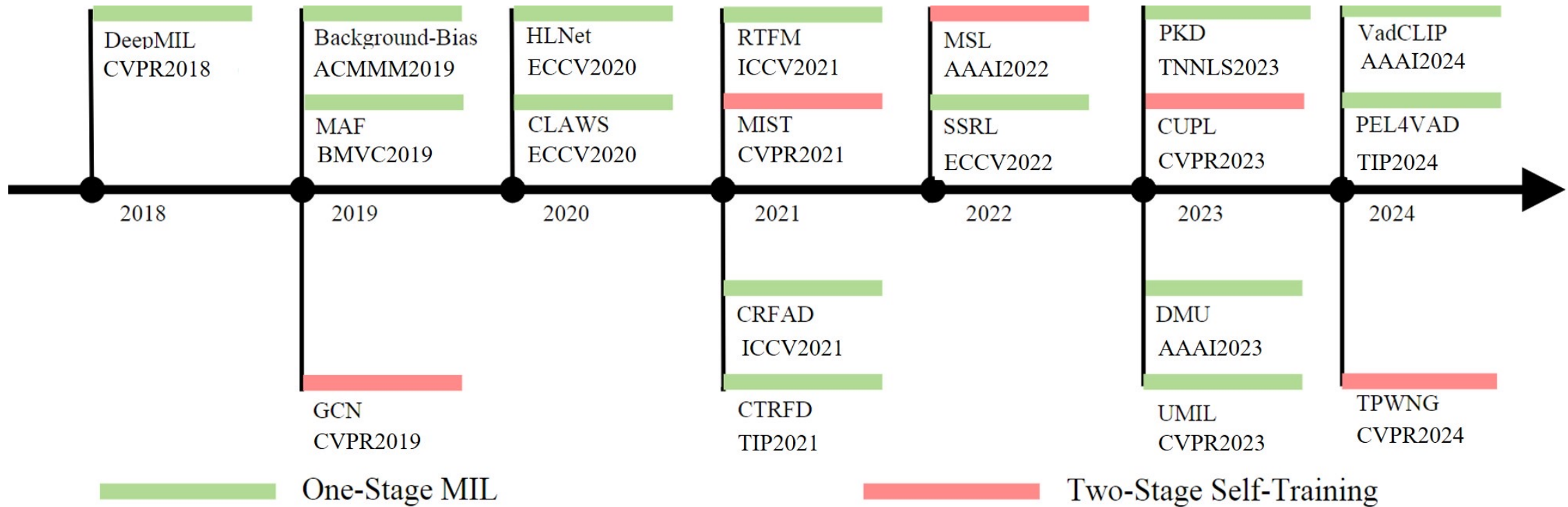
- **Definition :** Weakly supervised video anomaly detection identifies anomalous segments using only video-level labels.
- **Background:** Traditional WSVAD methods commonly rely on MIL to infer frame-level anomaly scores from video-level labels.
- **Challenge:** Real-world anomalies are diverse and complex, making it difficult for a single detector to remain robust across different anomaly patterns.

Explosion



Traffic accidents

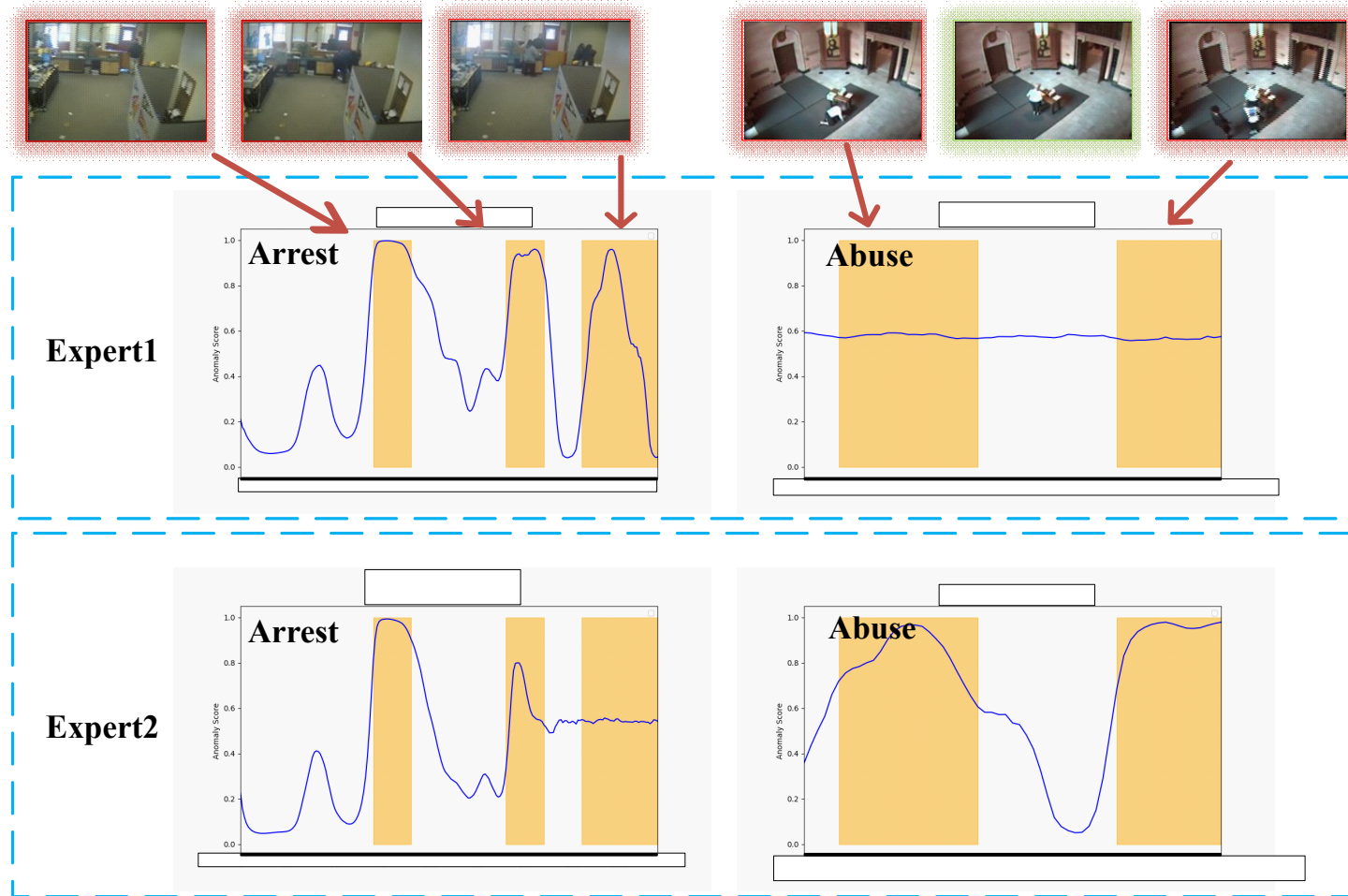
Related Work



- ❑ Single WSVAD detector is unreliable across diverse anomaly patterns
- ❑ Lack of sufficient fine-grained frame-level supervision

Problems

Anomaly Detection Curve



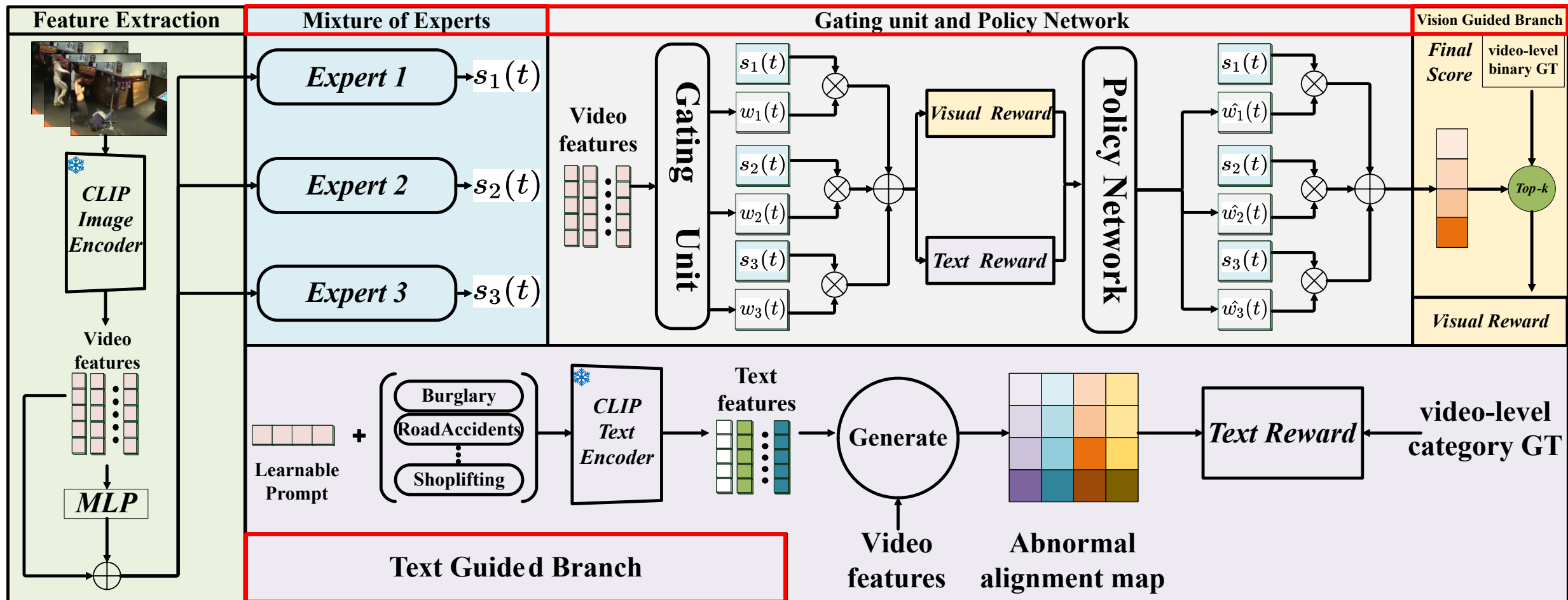
Compares the capabilities of different experts in modeling various abnormal situations

Motivation

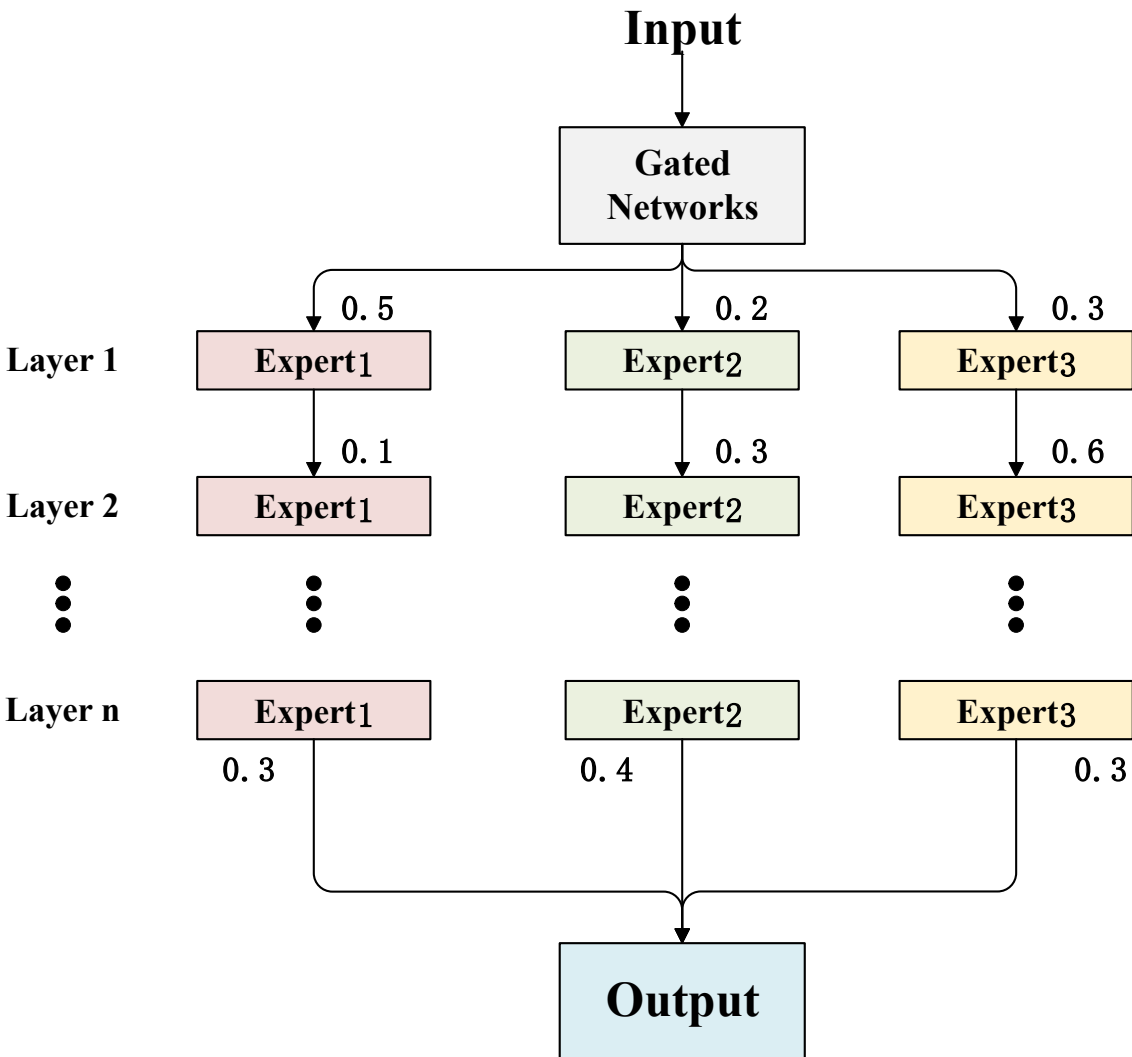
To address these problems :

- ◆ we propose a hybrid expert model that leverages a collaborative framework for comprehensive anomaly pattern analysis across different dimensions.
- ◆ We introduce an enhanced multi-expert ensemble strategy that incorporates fine-grained frame-level supervision to dynamically select the optimal expert combination for different videos.

Our Framework



Mixture of Experts



Expert1: multi-instance binary classification expert

$$s_1(t) = FC(FFN(E_t))$$

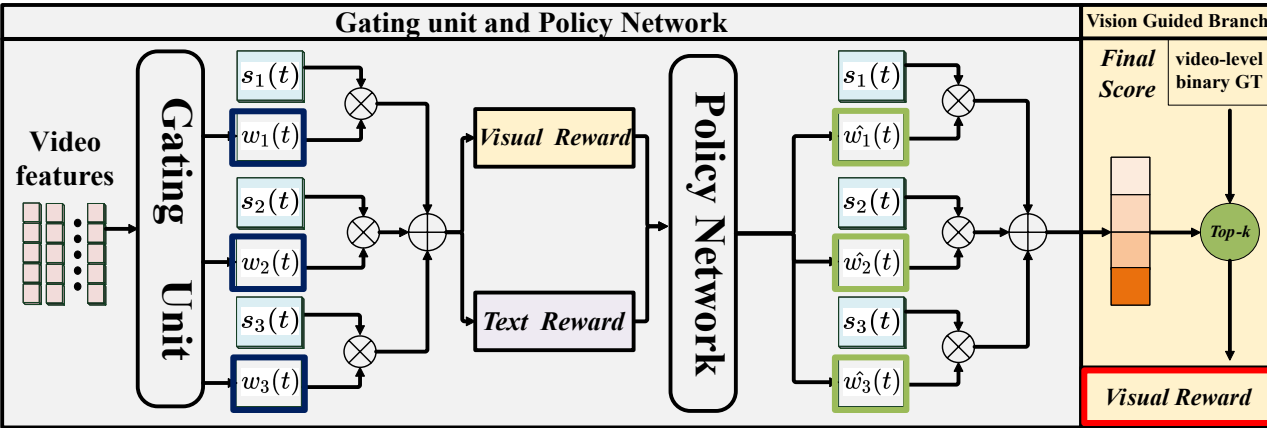
Expert2: anomaly attention mechanism based expert:

$$s_2(t) = As_1(t)$$

Expert3: Frame-prediction based expert

$$s_3(t) = |\widehat{v_{t+1}^{pred}} - v_{t+1}|_2^2$$

Reinforced Multi-Expert Ensemble Strategy

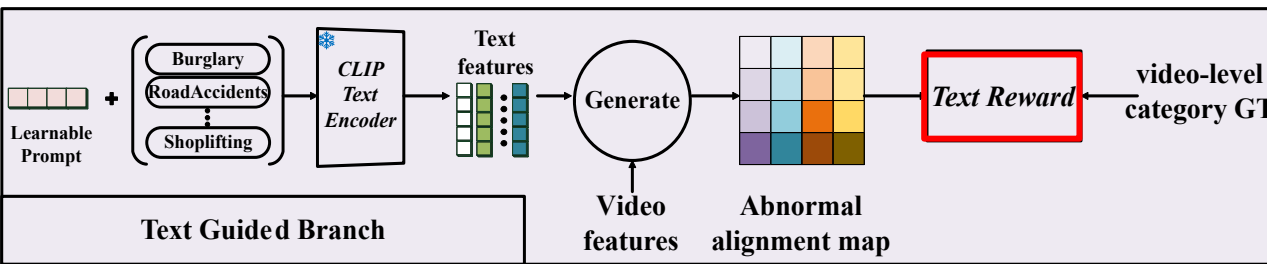


Global Visual Reward Function Based on Visual Guidance Branch:

$$R_{Visual} = -\frac{1}{N} \sum_{j=1}^N [y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j)]$$

Local Semantic Reward Function Based on Text Guidance Branch:

$$R_{Text} = -\sum_{i=1}^m y_i \cdot \log(p_i)$$



Experiments

Comparison with other methods on UCF-Crime

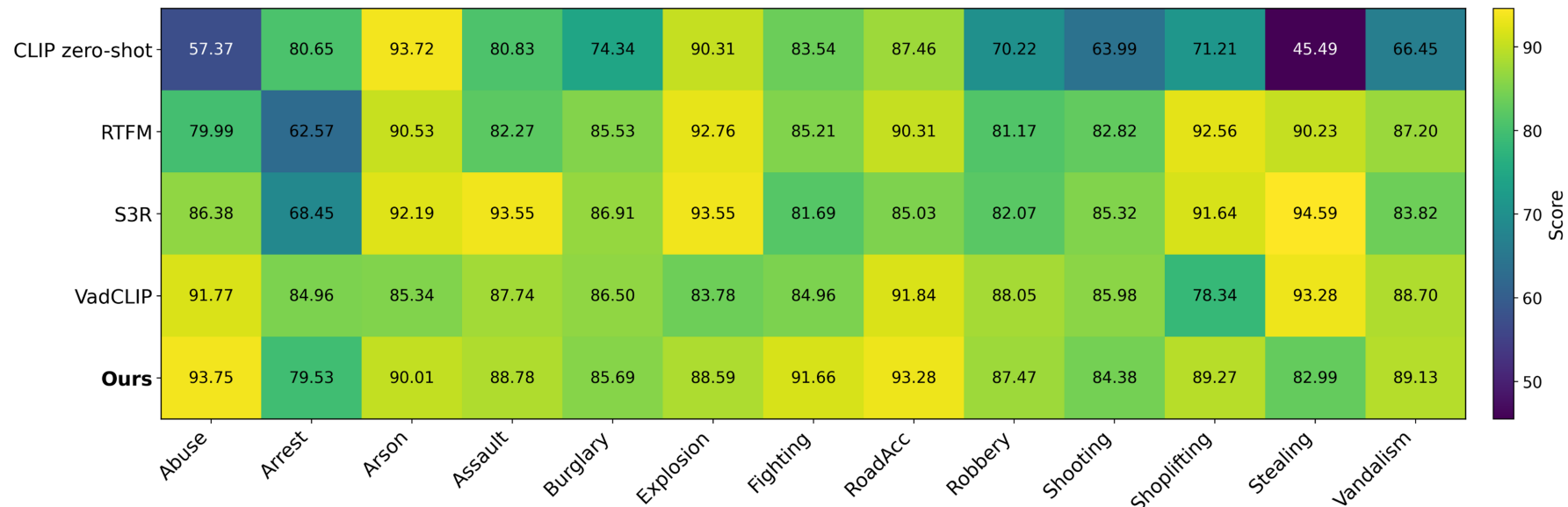
Method	Publication	Feature	AUC (%)
MSL	AAAI 2022	VideoSwin _{RGB}	85.62
S3R	ECCV 2022	I3D _{RGB}	85.99
SSPL	ECCV 2022	I3D _{RGB}	87.43
CMRL	CVPR 2023	I3D _{RGB}	86.10
CUPL	CVPR 2023	CLIP	86.22
MGFN	AAAI 2023	VideoSwin _{RGB}	86.67
UMIL	CVPR 2024	CLIP	86.75
DMU	AAAI 2024	CLIP	86.97
PE-MIL	CVPR 2024	I3D _{RGB}	86.83
TPWNG	CVPR 2024	I3D _{RGB}	87.79
VadCLIP	AAAI 2024	CLIP	88.02
STPrompt	ACMMM 2024	CLIP	88.08
Ours		CLIP	89.41

Experiments

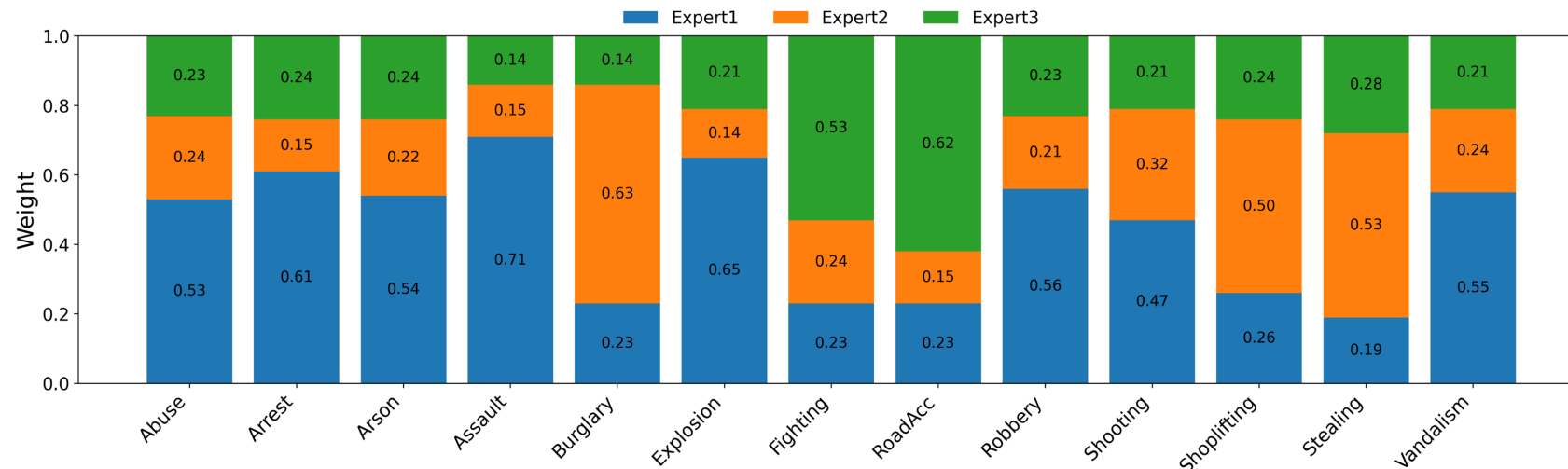
Comparison with other methods on XD-Violence

Method	Publication	Feature	AP (%)
MSL	AAAI 2022	VideoSwin _{RGB}	78.59
S3R	ECCV 2022	I3D _{RGB}	80.26
CMRL	CVPR 2023	I3D _{RGB}	81.30
CUPL	CVPR 2023	CLIP	81.43
MGFN	AAAI 2023	VideoSwin _{RGB}	80.11
DMU	AAAI 2024	CLIP	81.66
TPWNG	CVPR 2024	I3D _{RGB}	83.68
VadCLIP	AAAI 2024	CLIP	84.51
Ours		CLIP	86.07

Experiments



(a) Performance comparison on various anomalies on UCF-Crime



(b) Comparison of weight distribution of various expert

Ablation Studies

The effectiveness of three experts on UCF-Crime

Expert 1	Expert 2	Expert 3	AUC (%)
✓	✗	✗	87.63
✓	✓	✗	88.71
✓	✗	✓	88.96
✓	✓	✓	89.41

Ablation Studies

Effectiveness of Reinforced Multi-Expert Ensemble Strategy on UCF-Crime

Global Visual Reward		Local Semantic Reward		AUC (%)
	<i>x</i>		<i>x</i>	87.89
	✓		<i>x</i>	89.14
	<i>x</i>		✓	88.57
	✓		✓	89.41

Conclusion

- We introduce and effectively integrate three corresponding expert models, while designing appropriate training strategies for each.
- We introduce a frame-level gating unit based on reinforcement learning. By combining global video prediction results with text-visual alignment rewards, we provide guidance for the frame-level expert weight selection process.
- We validate the effectiveness of our method on two commonly used datasets in this field.

Thanks