Beyond Euclidean: Dual-Space Representation Learning for Weakly Supervised Video Violence Detection

Trustworthy Visual Intelligence Group

Jiaxu Leng, Zhanjie Wu, Mingpi Tan, Yiran Liu, Ji Gan, Haosheng Chen, Xinbo Gao*

Chongqing University of Posts and Telecommunications

Speaker: Zhanjie Wu

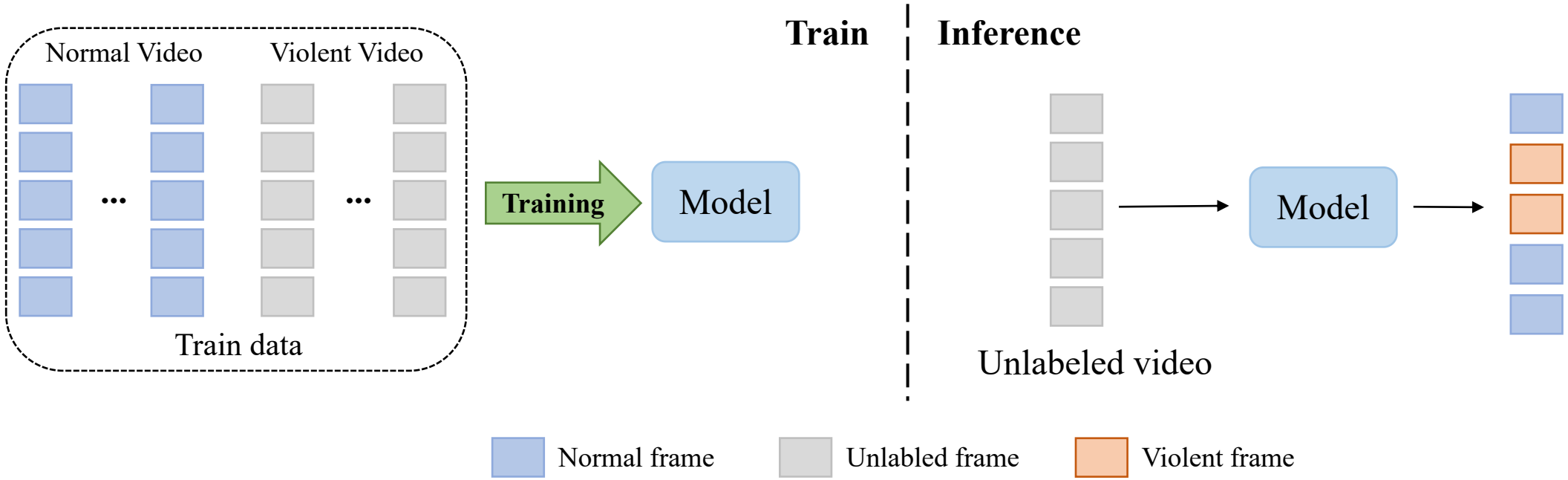
NEURAL INFORMATION PROCESSING SYSTEMS

Violence Detection

Outline

- Introduction
- Challenge
- Method
- Results
- Conclusion

Weakly Supervised Video Violence Detection (WSVVD)

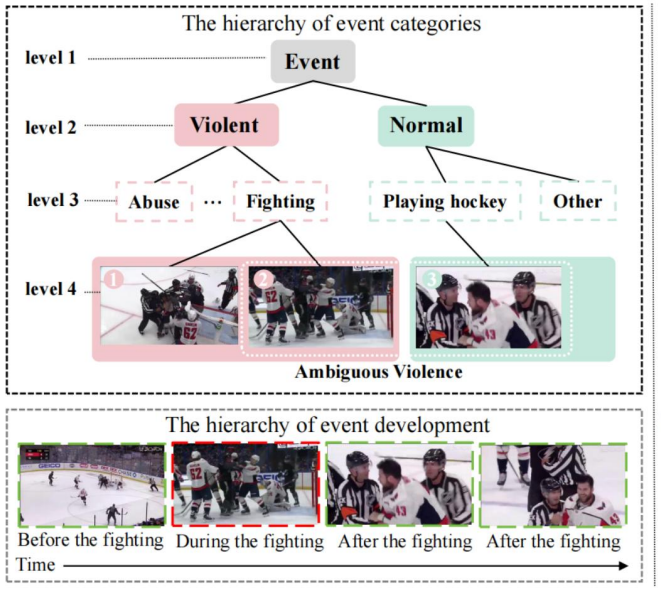


WSVAD involves training a binary classifier using video-level labels that include both normal and violent videos. An anomaly score is calculated for each frame to determine whether the frame contains a violent event.

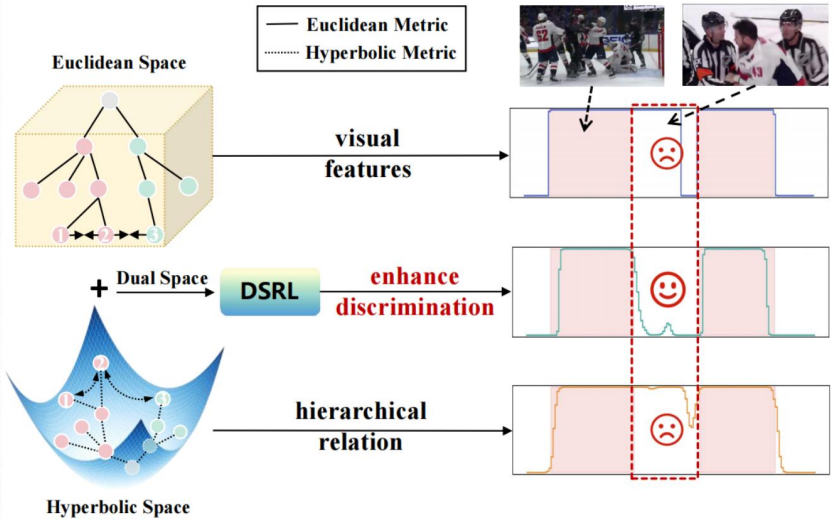
Ambiguous Violence: normal events that are visually similar to violent events

Existing Method:

- Using only Euclidean representations. (i.e. HL-Net, MACIL-SD)
- Using only hyperbolic representations. (i.e. HyperVD)



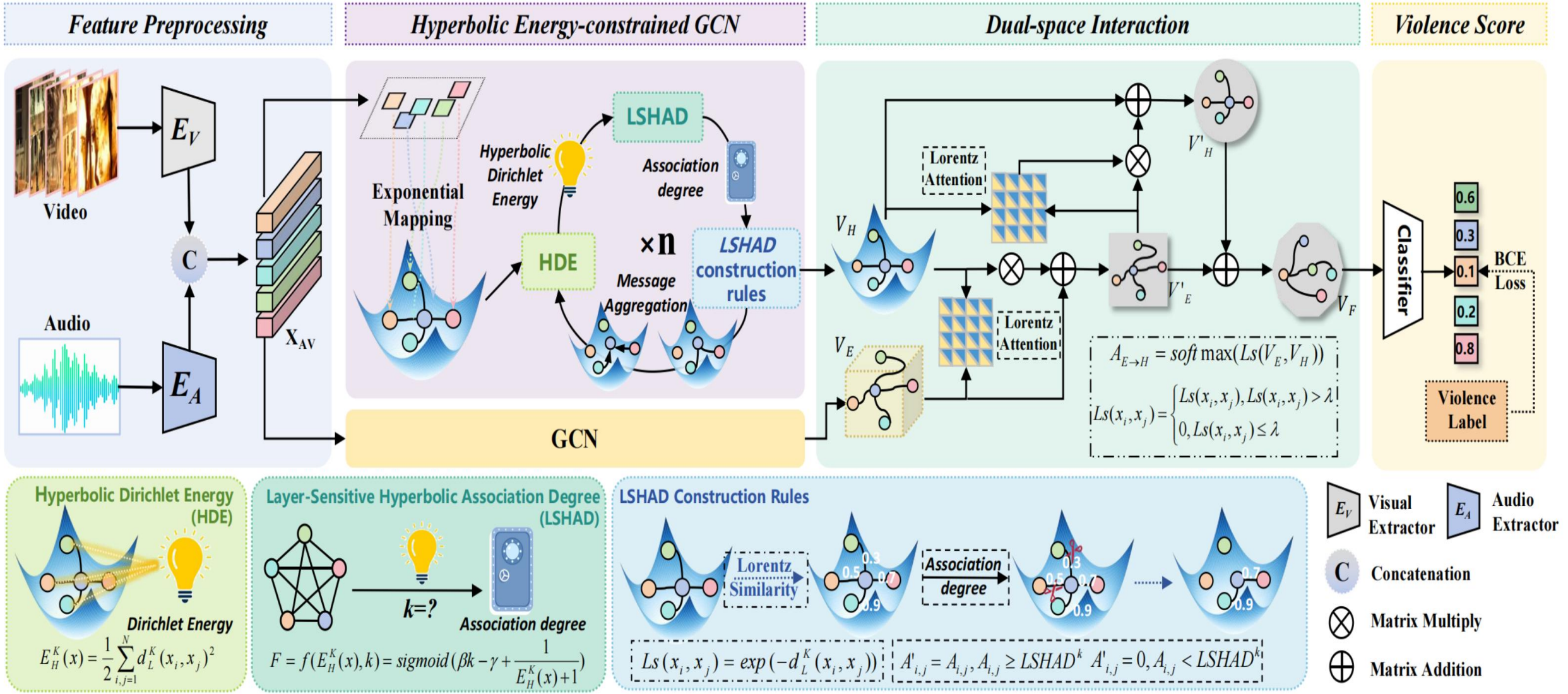
(a) Hierarchical Diagram in VVD



(b) Superiority of DSRL for Ambiguous Violence Detection

Movitation: hyperbolic representation enhances hierarchical event relations but weakens visual feature expression, while Euclidean representation emphasizes visual features but overlooks event relationships. DSRL effectively addresses ambiguous violence, which is challenging for either space alone.

Overview of the proposed DSRL framework



Comparisons with State-of-the-art Methods

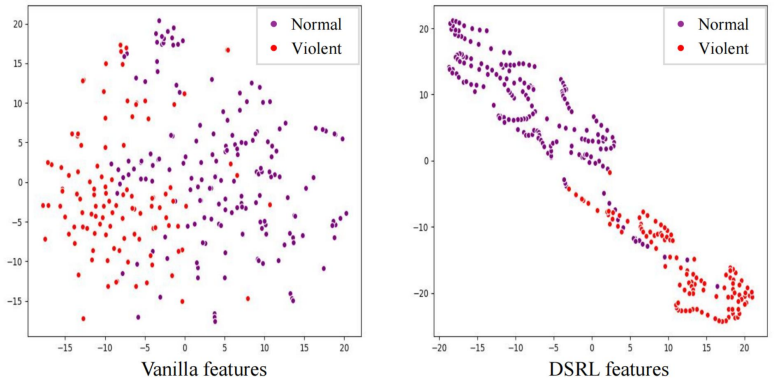
Methods	Input Setting	Feature Space	UCF-Crime	XD-Violence
Sultani et al. [29]	Unimodal	Euclidean	76.21	73.20
Wu et al. [32]	Unimodal	Euclidean	82.44	75.90
RTFM [30]	Unimodal	Euclidean	84.30	77.81
MSL [15]	Unimodal	Euclidean	85.30	78.28
MGFN [5]	Unimodal	Euclidean	**86.98**(1^{st})	79.19(3^{rd})
UMIL [17]	Unimodal	Euclidean	86.75(2^{nd})	81.66(2^{nd})
CU-Net [38]	Unimodal	Euclidean	86.22	78.74
Ours	Unimodal	Euclidean and Hyperbolic	86.38(3^{rd})	**82.01**(1^{st})
HL-Net [33]	Multimodal	Euclidean	-	78.64
Wu et al. [34]	Multimodal	Euclidean	-	78.64
Pang et al. [22]	Multimodal	Euclidean	-	79.37
UMIL [17]	Multimodal	Euclidean	-	81.77
Zhang et al. [38]	Multimodal	Euclidean	-	81.43
MACIL-SD [36]	Multimodal	Euclidean	-	83.40(3^{rd})
HyperVD [24]	Multimodal	Hyperbolic	-	85.67(2^{nd})
Ours	Multimodal	Euclidean and Hyperbolic	-	**87.61**(1^{st})

Ablation Studies

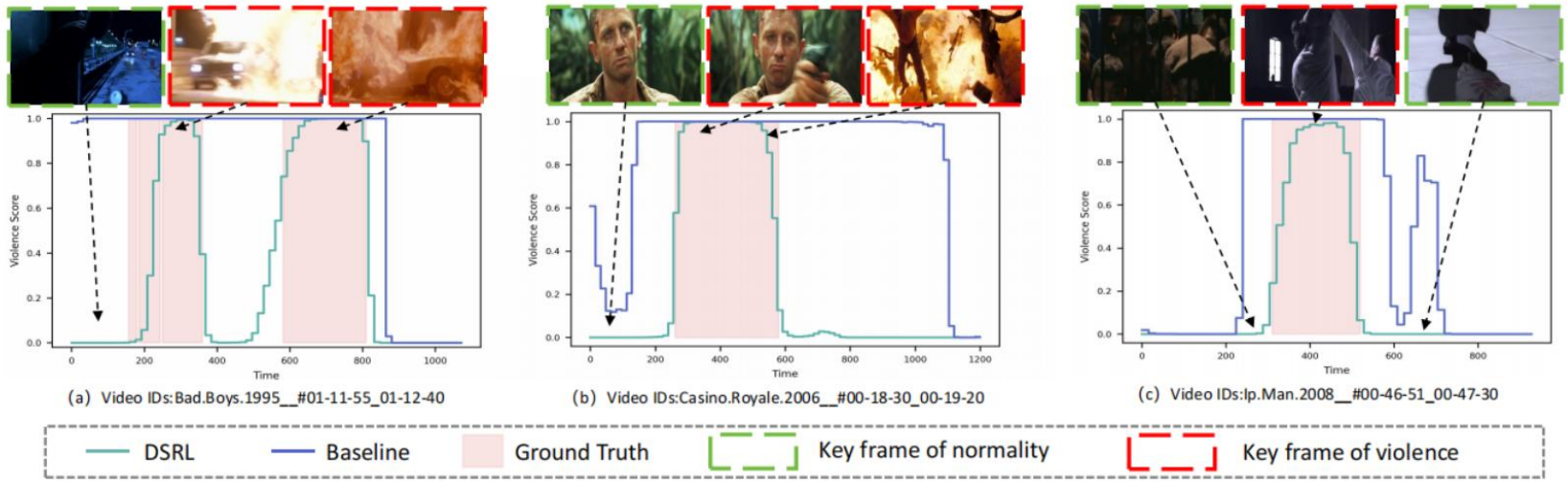
Table 2: Ablations on XD-Violence dataset.

Euclidean	Hyperbolic		DSI			XD-Violence	
GCN	HE-GCN	HGCN	Concat	Cosine Metric	Lorentzian Metric	Multimodal(%)	Unimodal(%)
✓						84.04	77.95
✓		✓	✓			85.01	77.93
✓	✓		✓			86.46	79.70
✓	✓			✓		86.91	80.72
✓	✓				✓	87.61	82.01

Feature Discrimination Visualization

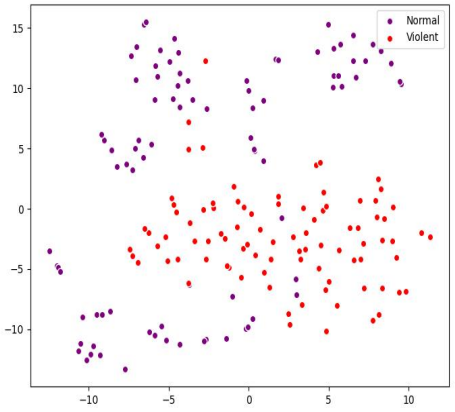


Qualitative Visualizations

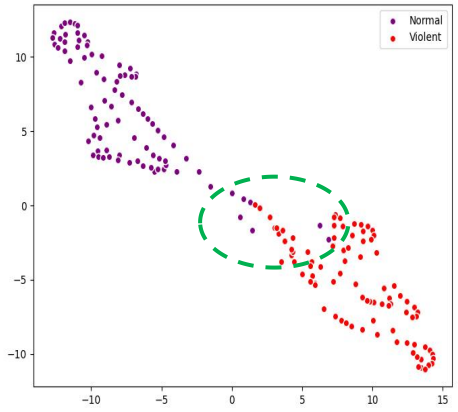


Visualizations of the ablation modules

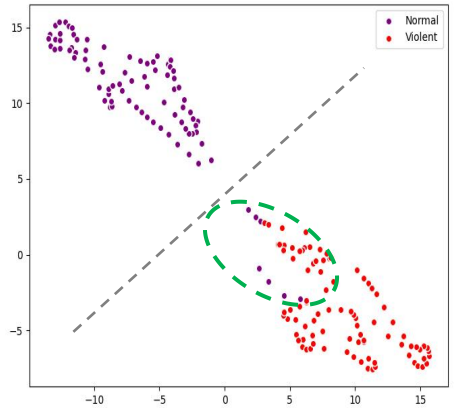
Feature level



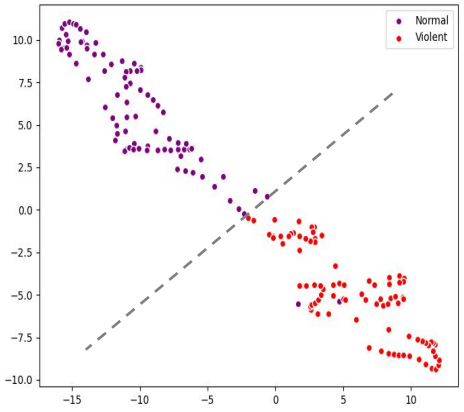
(a) Vanilla Features



(b) w/o HE-GCN and w/o DSI

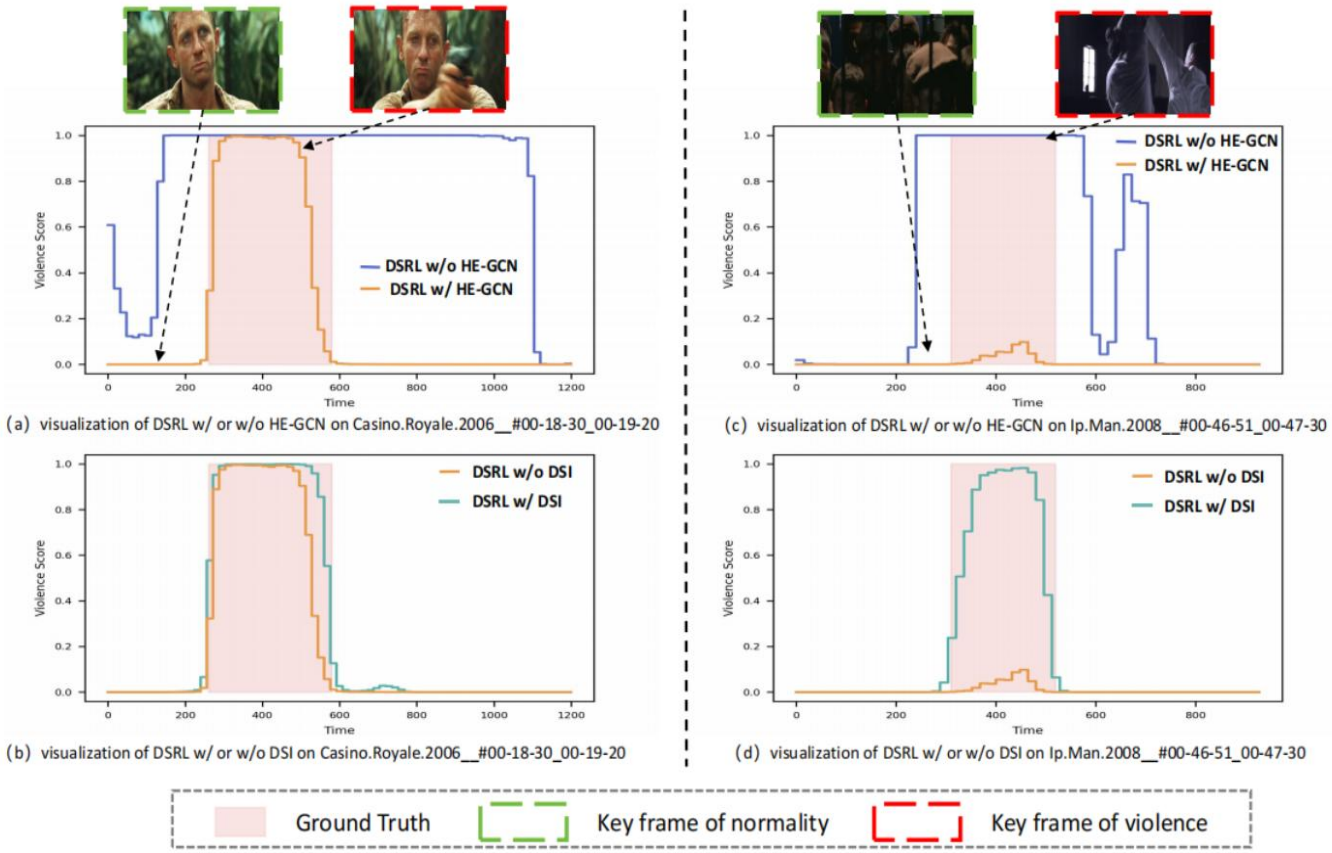


(c) w/ HE-GCN and w/o DSI



(d) w/ HE-GCN and w/ DSI

Frame level



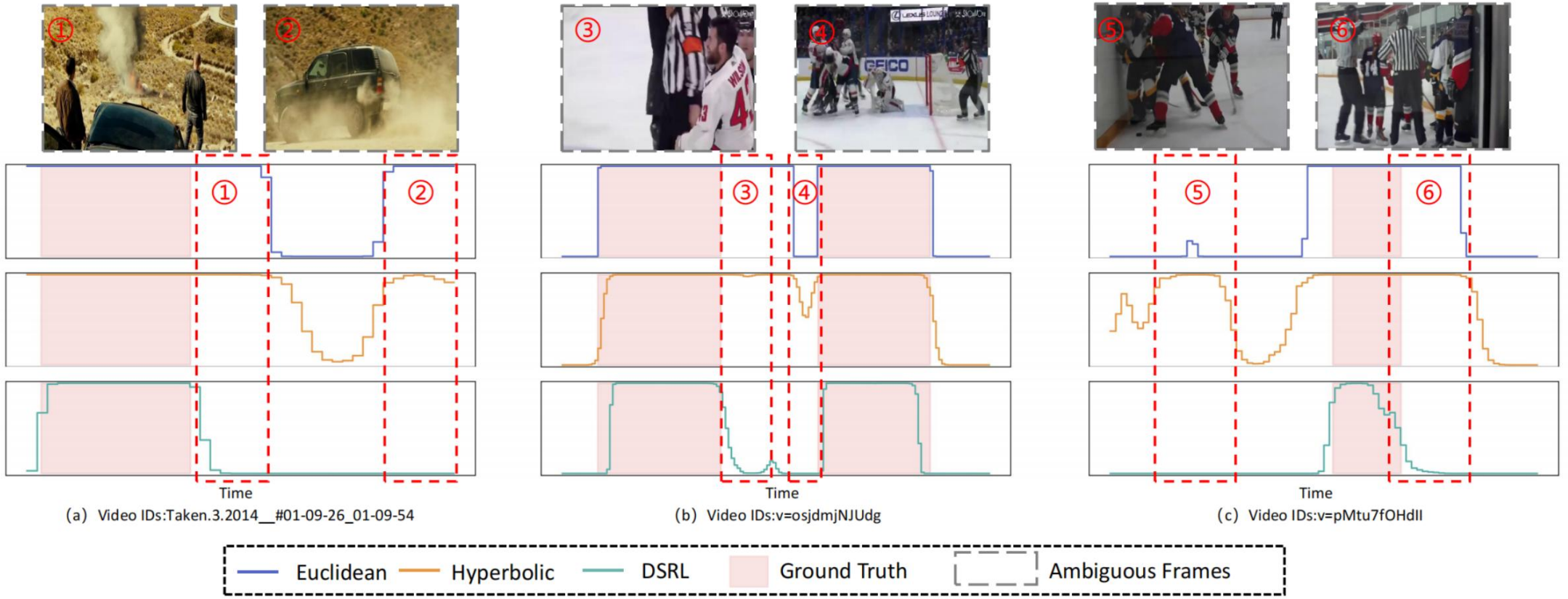
(a) visualization of DSRL w/ or w/o HE-GCN on Casino.Royale.2006__#00-18-30_00-19-20

(c) visualization of DSRL w/ or w/o HE-GCN on Ip.Man.2008__#00-46-51_00-47-30

(b) visualization of DSRL w/ or w/o DSI on Casino.Royale.2006__#00-18-30_00-19-20

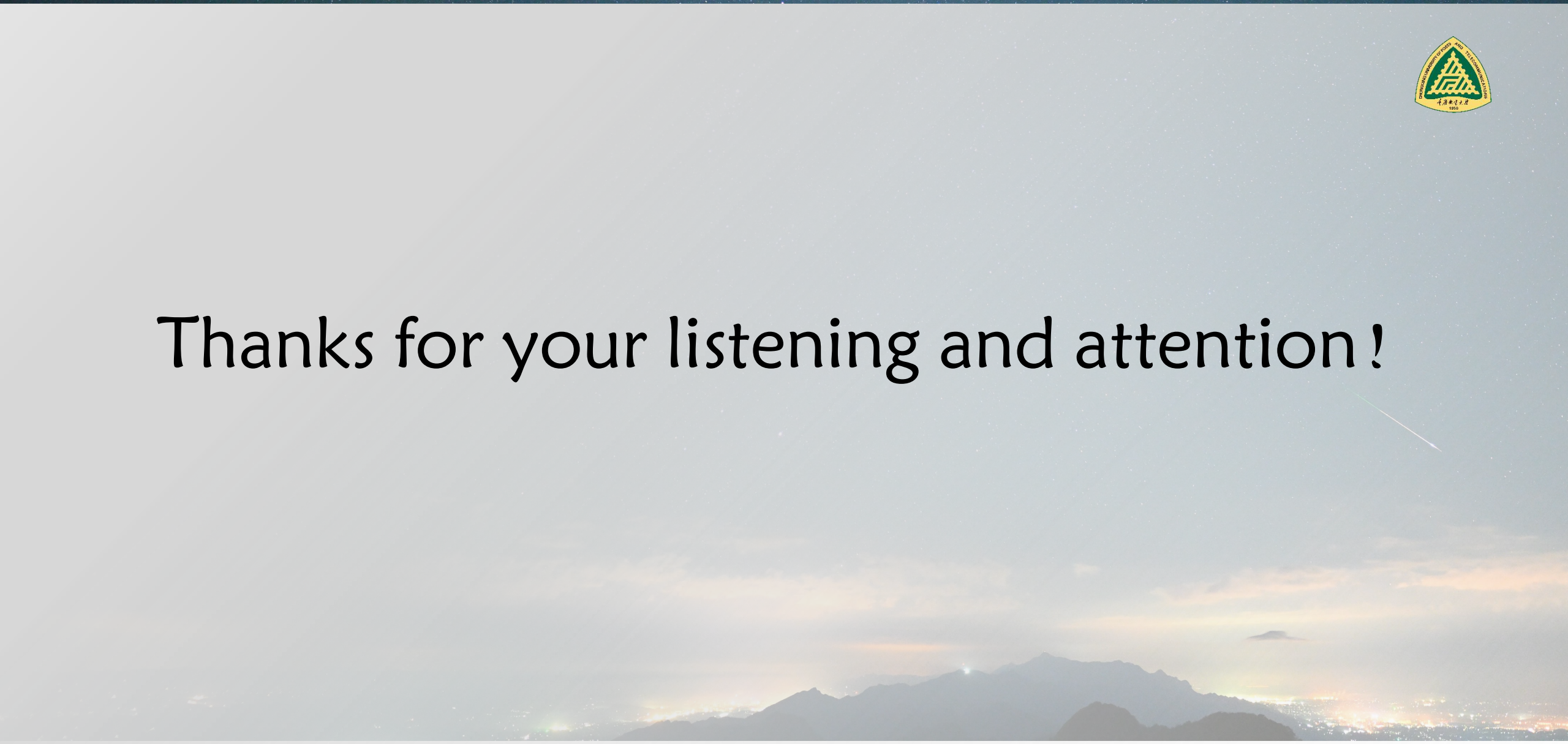
(d) visualization of DSRL w/ or w/o DSI on Ip.Man.2008__#00-46-51_00-47-30

Qualitative Visualizations of DSRL in the context of ambiguous violence



This supports our motivation: hyperbolic representation enhances hierarchical event relations but weakens visual feature expression, while Euclidean representation emphasizes visual features but overlooks event relationships. DSRL effectively addresses ambiguous violence, which is challenging for either space alone.

- We propose a comprehensive geometric representation learning method, Dual-Space Representation Learning (DSRL) which integrates the benefits of Euclidean and hyperbolic geometries to improve the discrimination of ambiguous violence.
- Hyperbolic Energy-constrained Graph Convolutional Network (HE-GCN) is designed to better capture the hierarchical context of events.
- Additionally, Dual-Space Interaction (DSI) is designed to facilitate information interactions.
- Our method achieves SOTA performance on the XD-Violence dataset in both unimodal and multimodal settings, especially excelling in resolving ambiguous violence.



Beyond Euclidean: Dual-Space Representation Learning for Weakly Supervised Video Violence Detection

NEURAL INFORMATION PROCESSING SYSTEMS