WACV 2026
TUCSON, AZ 3/6 - 3/10

Learning Unified Spatio-temporal Representations for Efficient Compressed Video Understanding


Shristi Das Biswas


Efstathia Soufleri


Arani Roy

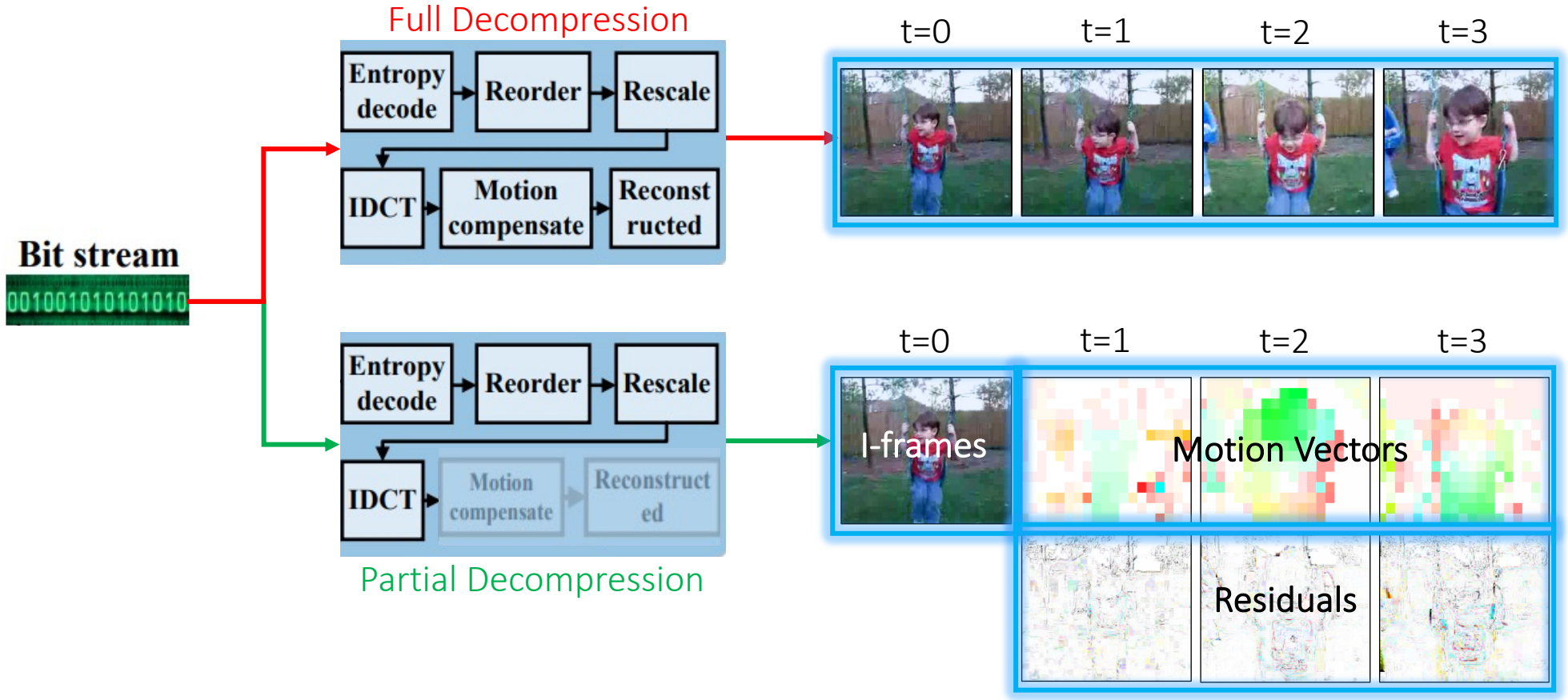

Kaushik Roy

PURDUE UNIVERSITY
C-BRIC Center for Brain Inspired Computing
SRC Semiconductor Research Corporation
IARPA
DARPA
NSF
U.S. DEPARTMENT OF ENERGY

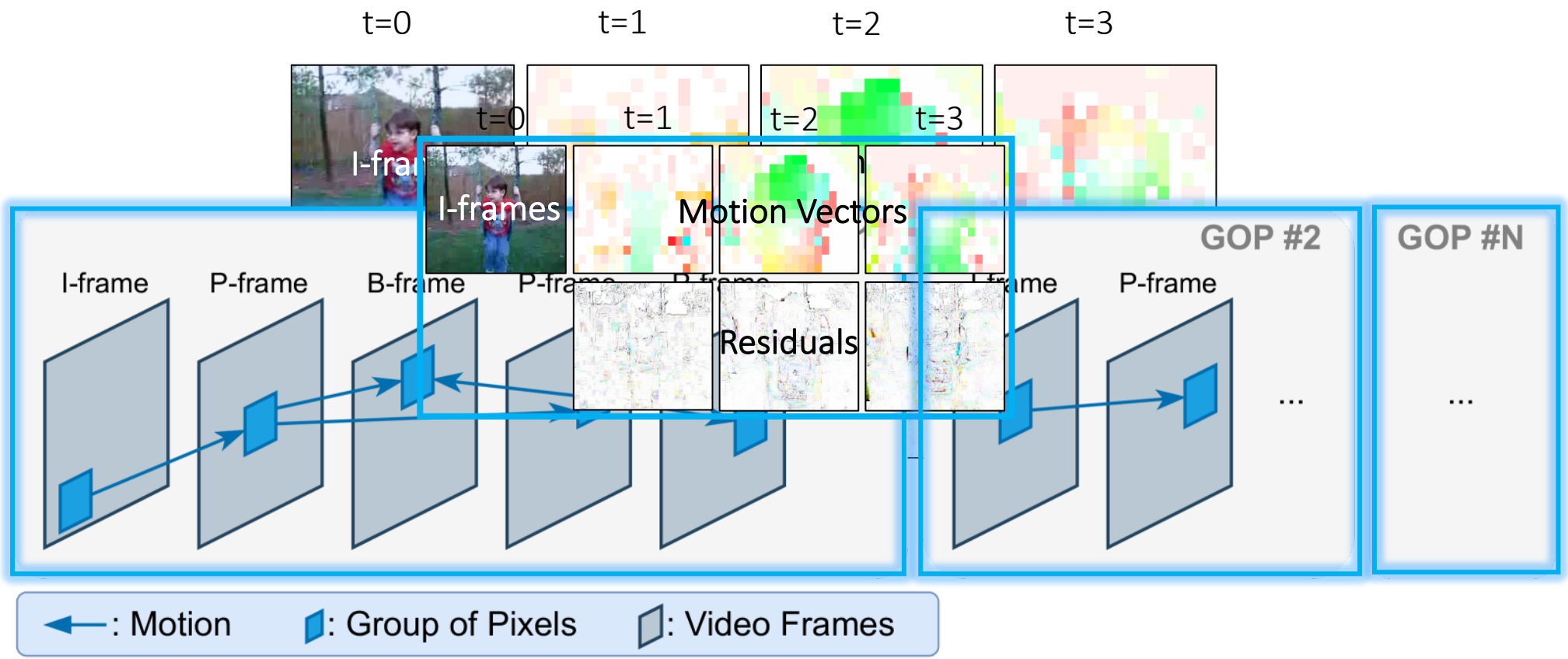
What is the task?

Action recognition is an important task in the field of computer vision that entails classifying human actions depicted in video frames. Think of it as the video counterpart of image classification.

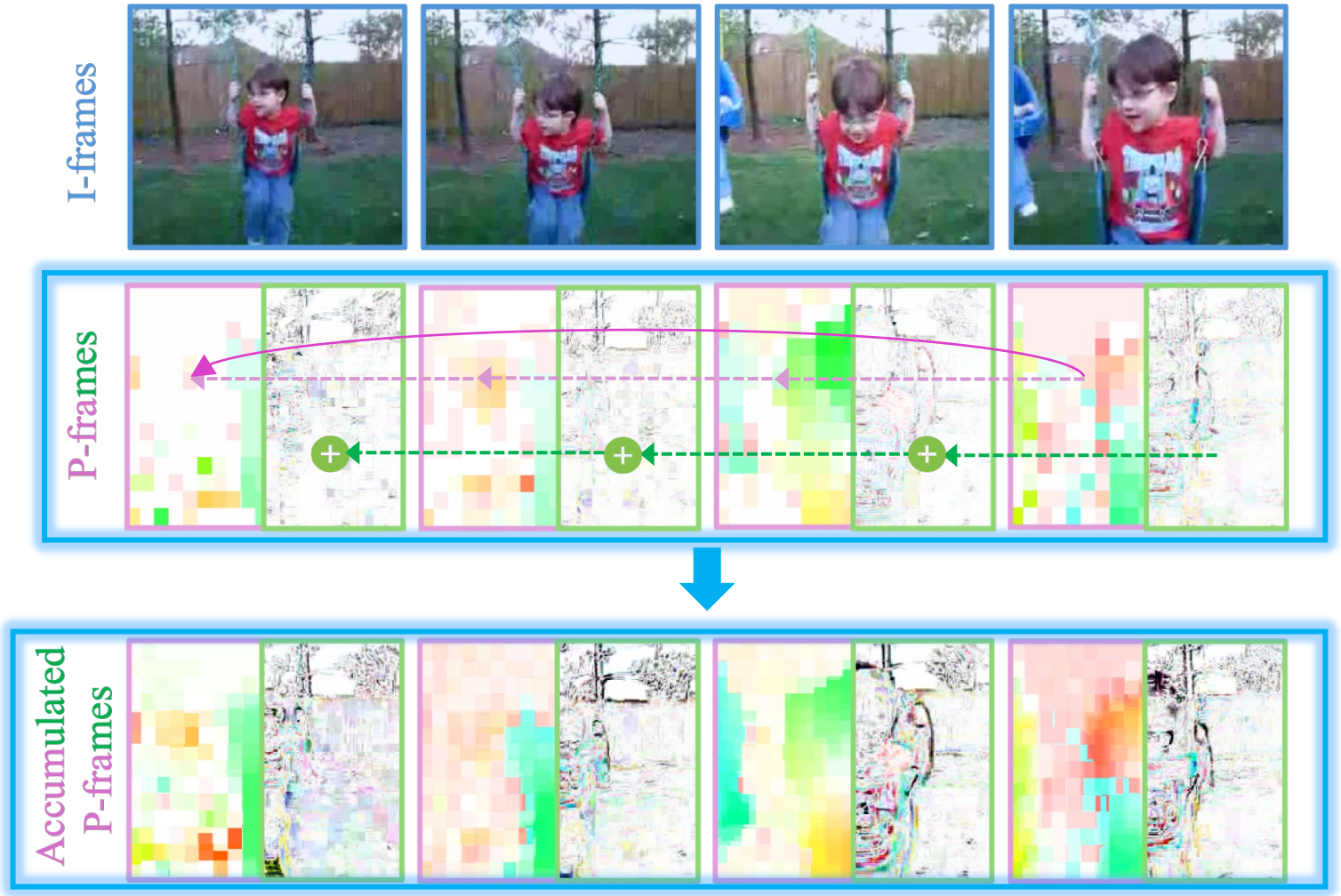
Understanding Compressed Video Modalities



The Structure in Compressed Videos



Modeling Compressed Video Representations

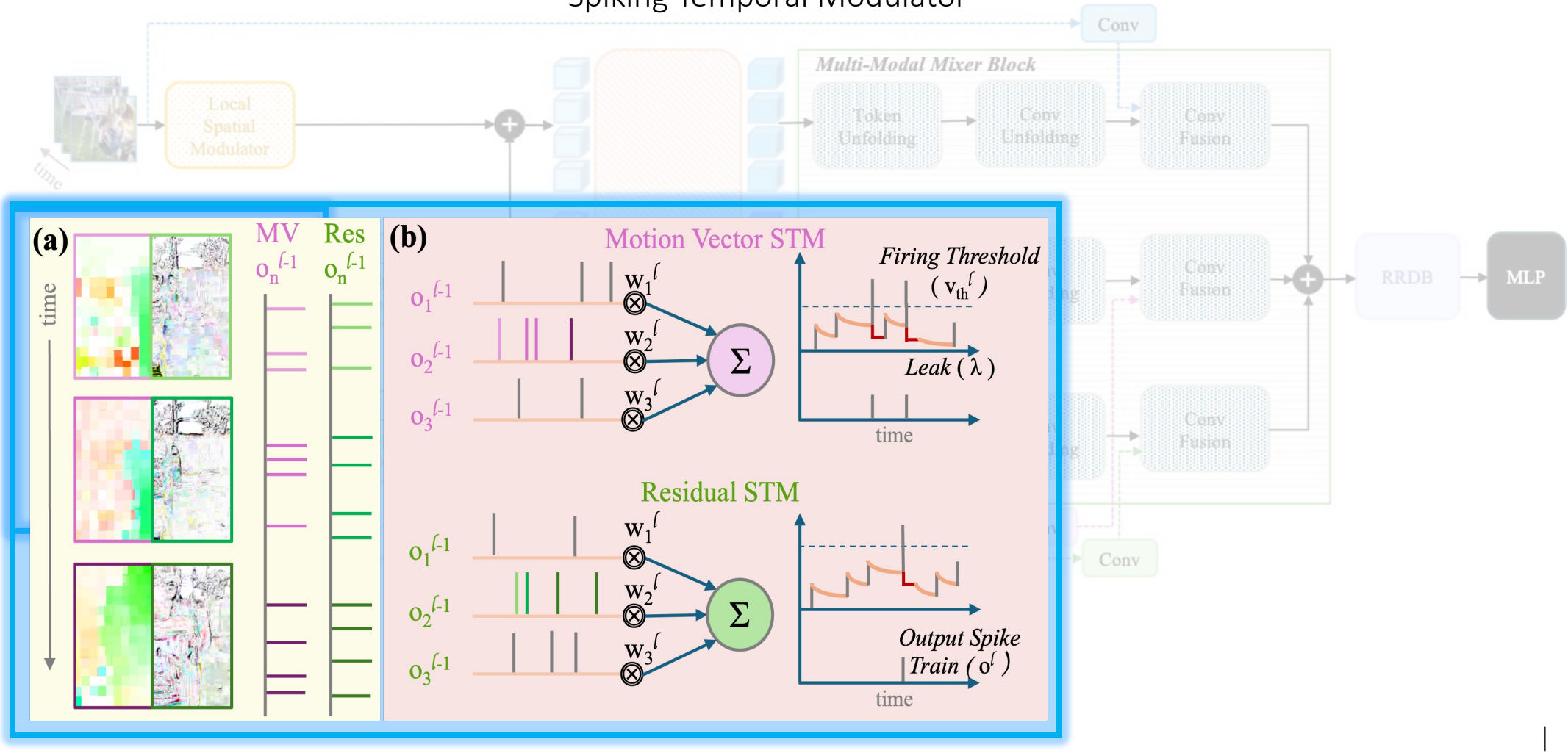


Original motion vectors and residuals capture only interframe changes, often with low signal-to-noise ratios, making them difficult to model.

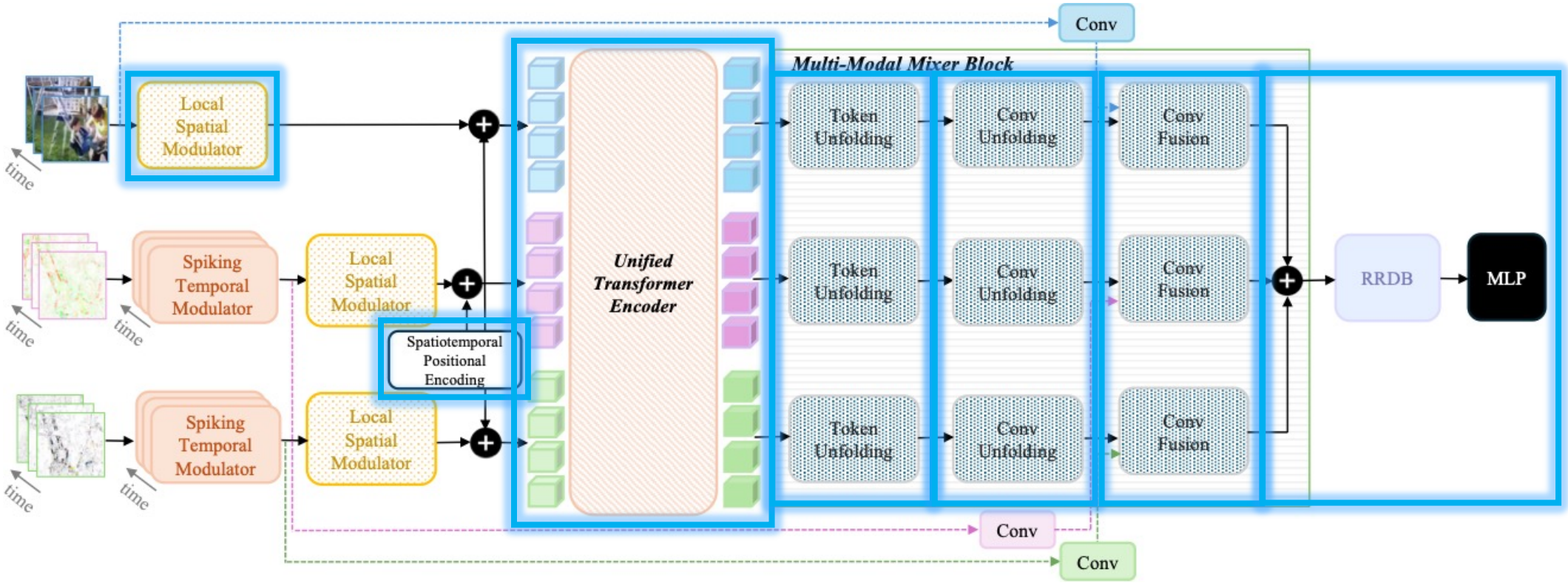
These observations motivate us to investigate accumulated P-frames to aggregate longer-term differences, allowing better visualization of pixel ownership for moving objects

Unified Spatio-Temporal Modeling of Compressed Modalities

Spiking Temporal Modulator



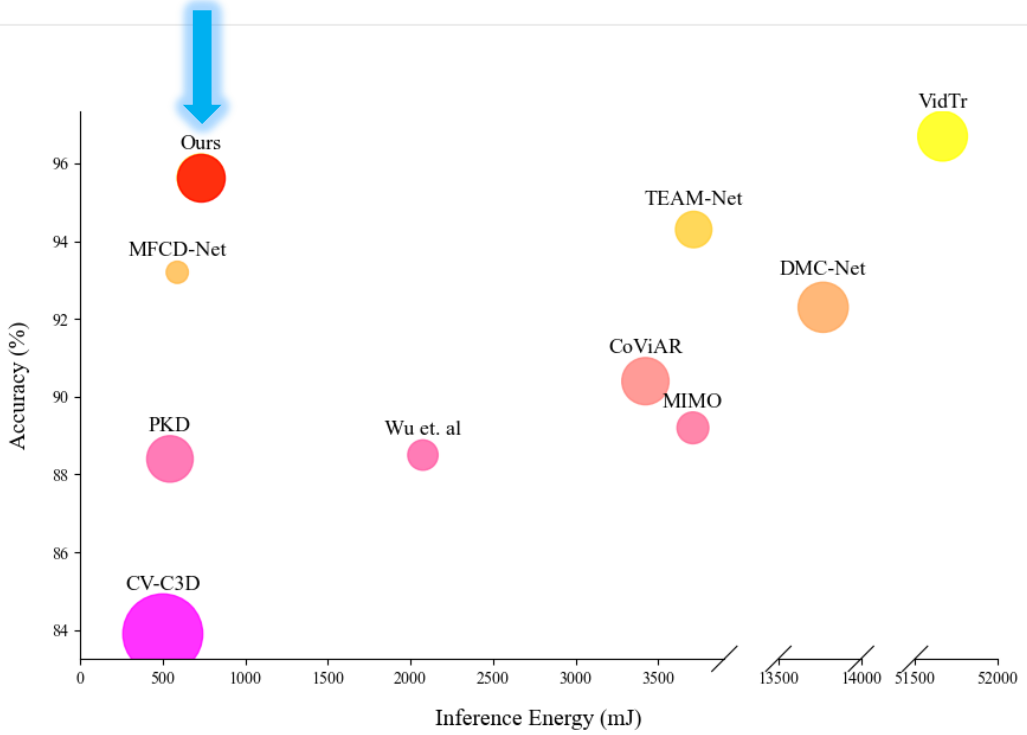
Unified Spatio-Temporal Modeling of Compressed Modalities



Performance Evaluation

Smaller Benchmarks

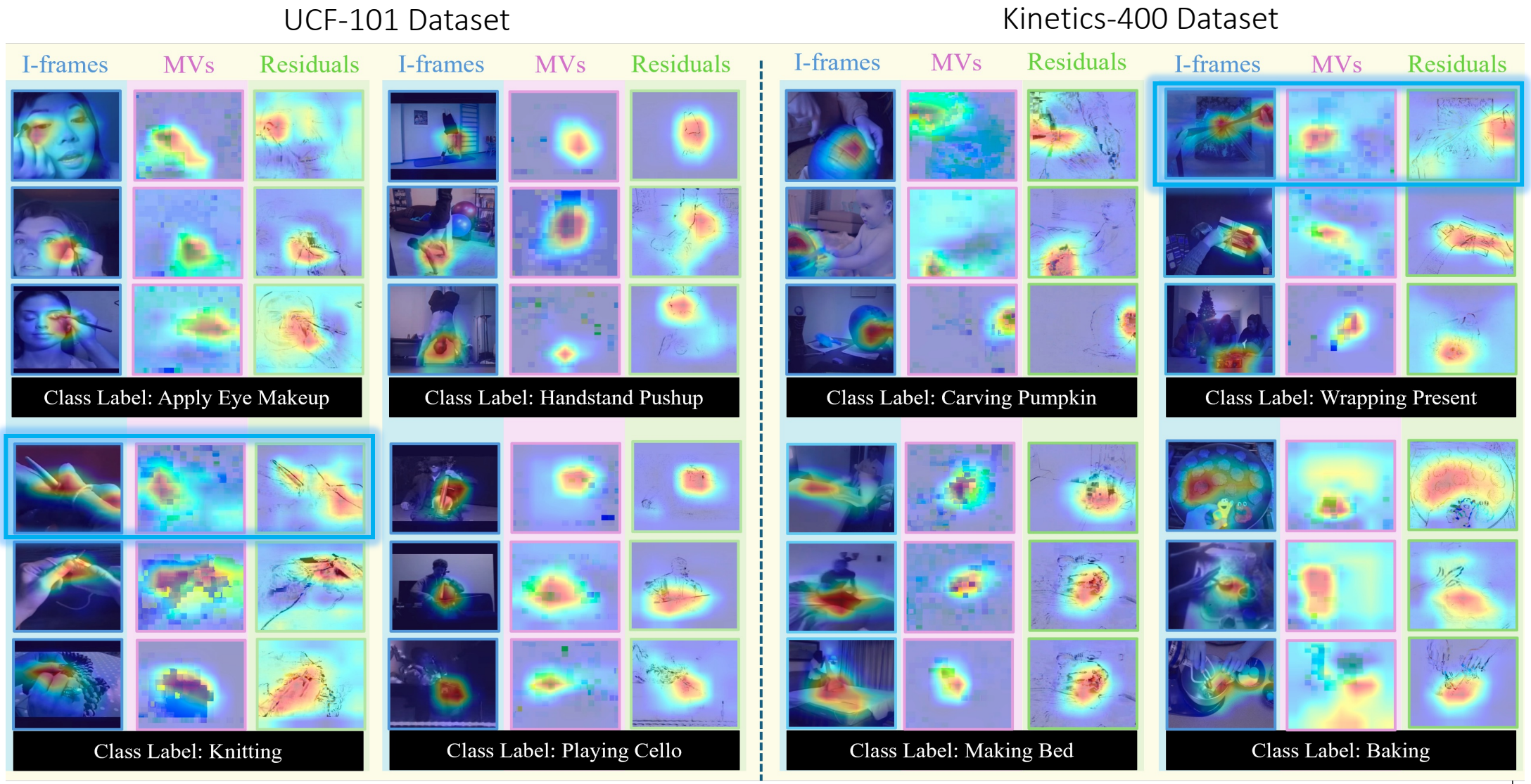
Methods	M	P_{Spike}	$E_{Total}(J)$	Speed (V/s)
ALT-L/14 [24]			242.733	-
VidTr [23]			51.180	0.284
Liu et. al [35]	RI		>3.422	-
MFCD-Net [58]	RI		0.589	0.352
DMC-Net [59]	CI		13.771	0.043
CV-C3D [32]			0.501	1.259
MIMO [34]			3.709	1.307
TEAM-Net [30]			3.714	1.537
CoViAR [6]			3.422	1.585
Wu et al. [31]			2.075	1.658
PKD [33]			0.544	-
MM-ViT [63]			82.801	-
Ours		99	**0.734**	**16.025**



Longer Benchmarks

Method					
ViViT-L FE [37]					
VidTr [23]					
I3D [10]	RD	50.0	-		100.67
VideoMAE V2-g [25]	RD	77.0	88.5	88.8	2633.04
VideoSwin-L [22]	RD	69.6	84.9	86.1	484.61
Video-FocalNet-B [26]	RD	71.1	83.6	86.7	822.48
Liu et. al [35]	RD + CD	-	73.5	-	>3.42
MFCD-Net [58]	RD + CD	-	68.3	-	0.59
CoViAR [6]	CD	-	69.1	-	3.42
TEAM-Net [30]	CD	-	72.2	-	3.71
MM-ViT [63]	CD	64.9	-	81.5	82.80
Ours	**CD**	**62.5**	**74.2**	**80.3**	**0.73**
Ours-L	**CD**	**64.2**	**75.9**	**81.1**	**1.40**

Where was the model "looking" to make its decision?



Ablation Insights – Investigating Modality/Architecture Choices

I	MV	R	Top-1 [%]	Top-5 [%]
✓	✓	✗	93.21	97.46
✓	✗	✓	92.93	97.01
✗	✓	✓	84.06	88.32
✓	✓	✓	95.63	99.53

Variant	Top-1 [%]
LS ⇒ GS ⇒ T	90.72
LS ⇒ T ⇒ GS	91.64
T ⇒ LS ⇒ GS	95.63

(a) Contextualization Order

Variant	Top-1 [%]	$E_{Total}(J)$
Cross-att.	94.72	0.93
Separate	94.18	0.95
Unified	95.63	0.73

(b) Transformer Encoder Analysis

Variant	Top-1 [%]
W/o STM	88.61
Common STM	93.45
STM → LSTM	93.92
STM (fixed v_{th}, λ)	93.91
STM	95.63

Variant	Top-1 [%]
W/o P-frame Accumulation	89.65
W/o Multi-Modal Mixer	90.84
W/o skips	89.02
W/o RRDB	91.77
GAP → Class Token	94.31

Summary

- To efficiently learn robust spatio-temporal representations that can effectively model both local and global contexts, this paper re-examines the design of established video understanding backbones.
- Motivated by the practical observation that decompressing videos is not only an overhead but also an inconvenience representations become less robust and increases dimensionality to make training computationally challenging, we propose a lightweight yet powerful factorized end-to-end framework to unify the advantages of compressed video modalities in a compact way.
- The efficacy of our design choices in effectively modeling shared spatiotemporal statistical patterns in the compressed representation is evidenced by our strong performance on five public benchmarks while offering sizeable benefits in computational cost and inference latency.
- The resulting design is an excellent choice for resource-constrained edge applications and hopes to inspire future work toward efficient video understanding systems not requiring decoded videos.

Thank You!

PURDUE UNIVERSITY®