NEURAL INFORMATION PROCESSING SYSTEMS 2025

Project Page

PASS: Path-selective State Space Model for Event-based Recognition

Jiazhou Zhou[1], Kanghao Chen[1], Lei Zhang[2], Lin Wang[3]
(†Corresponding Author)

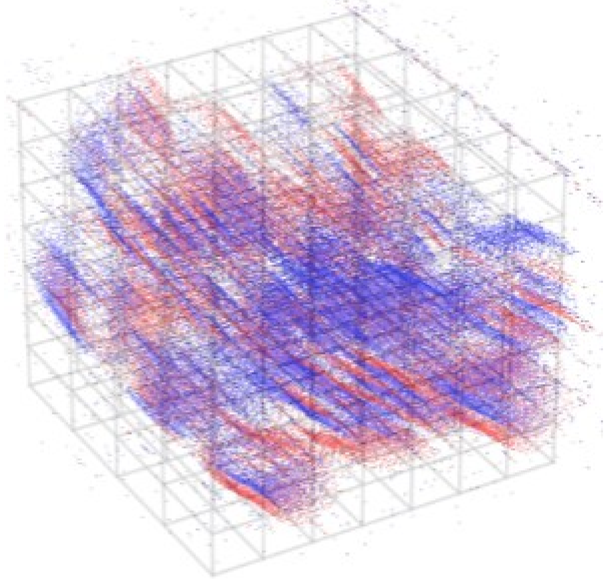
香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY
粤港澳大湾区数字经济研究院（福田）

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Project Page: https://jiazhou-garland.github.io/PASS_Homepage/

Background: Event Camera



Event camera perceive the **per-pixel brightness** changes **asynchronously.**

$$\varepsilon = \sum e_i(x_i, y_i, t_i, p_i)$$

ε encodes three critical pieces of information:

- **time** t_i
- **pixel location** (x_i, y_i)
- **polarity of intensity changes** p_i.

- It advances in:

High Temporal Resolution

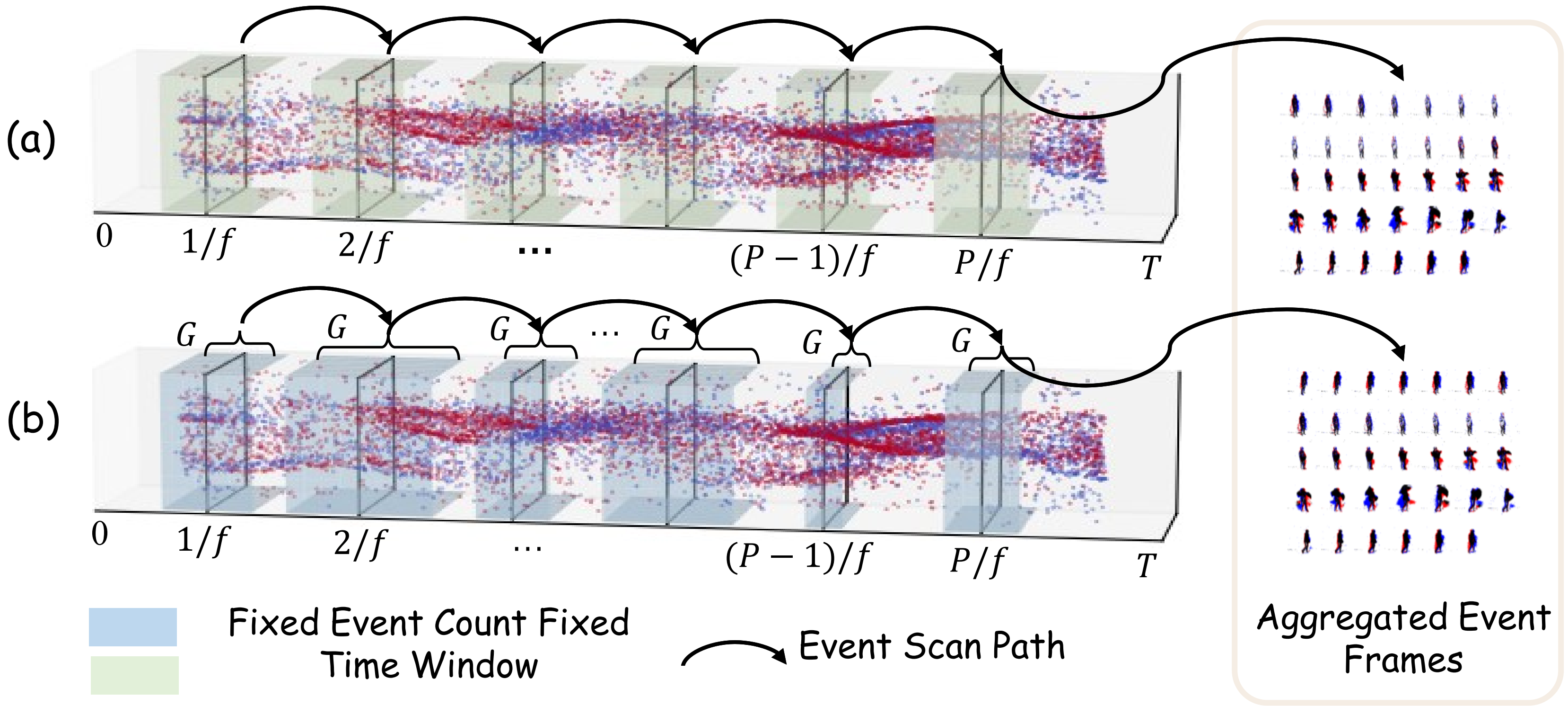
- Being resilient to:

Rapid Motion

Illumination Changes

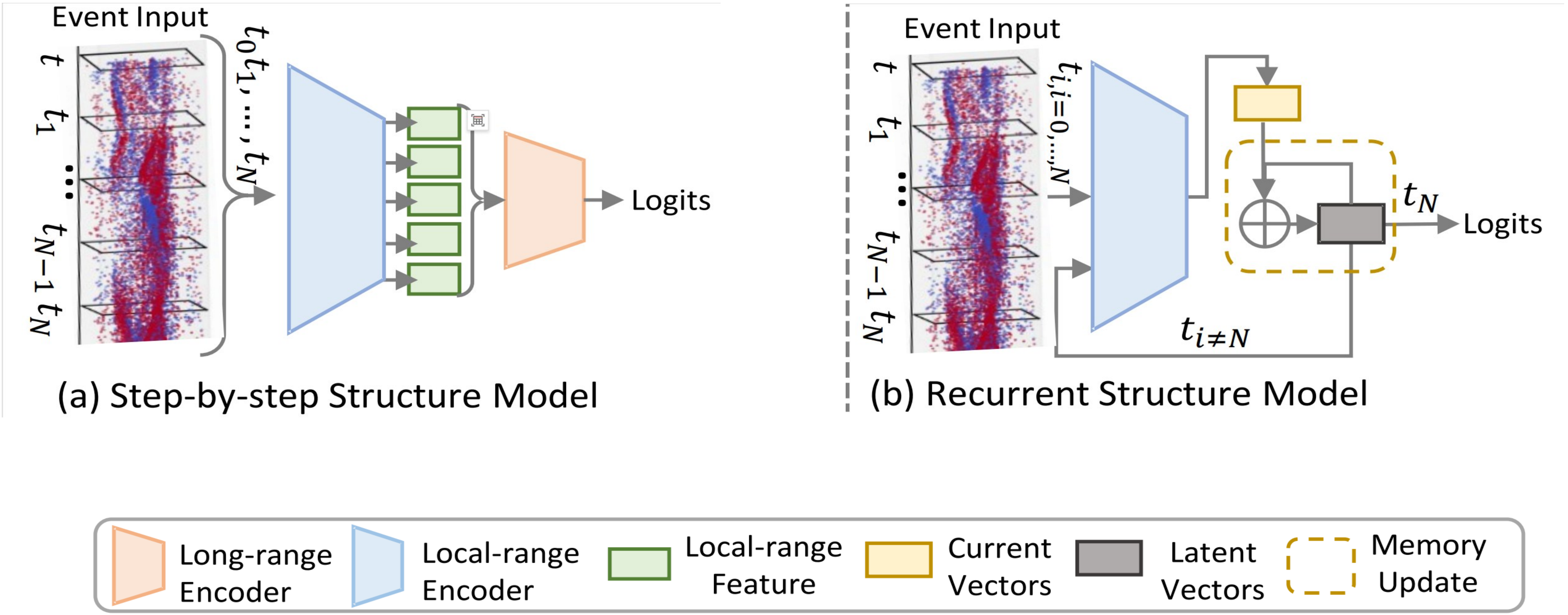
Related Work

Major Challenge: How to efficiently process and interpret event camera data characterized by high temporal density and spatial sparsity?



Related Work

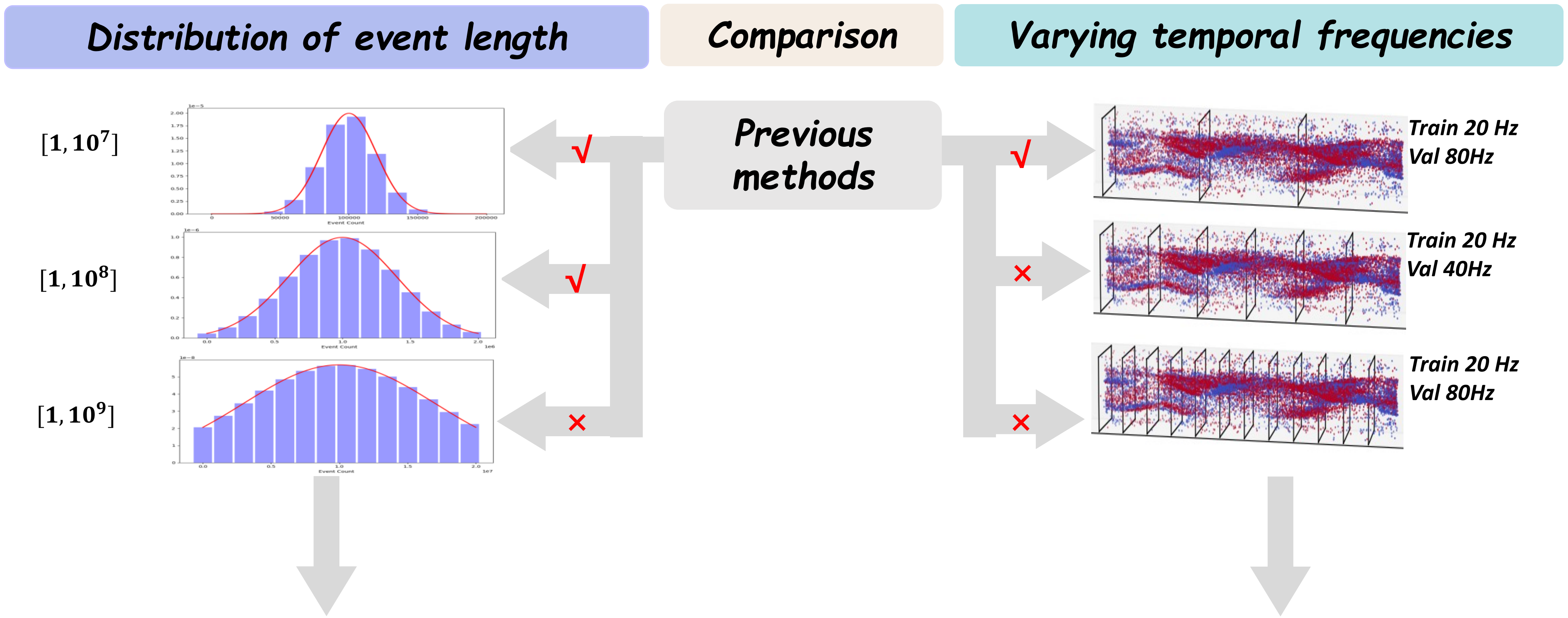
Previous methods generally fall into two categories:



(a) Step-by-step Structure Model

(b) Recurrent Structure Model

Limitations of Existing Methods



Distribution of event length

Comparison

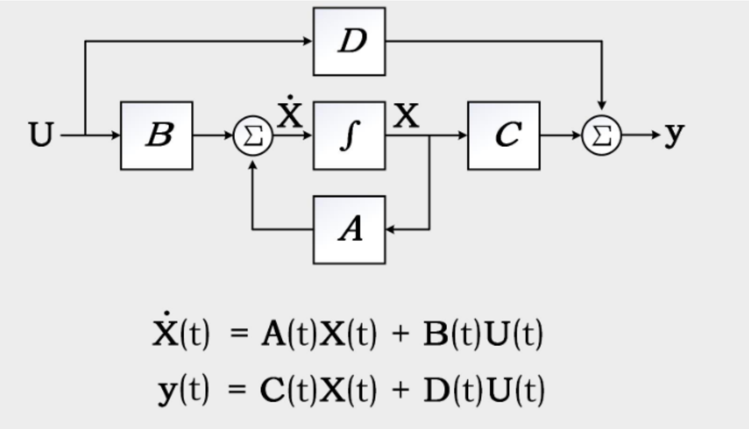
Varying temporal frequencies

Previous methods

Limited handling of event length distribution

Poor inference frequency generalization

Research Motivation

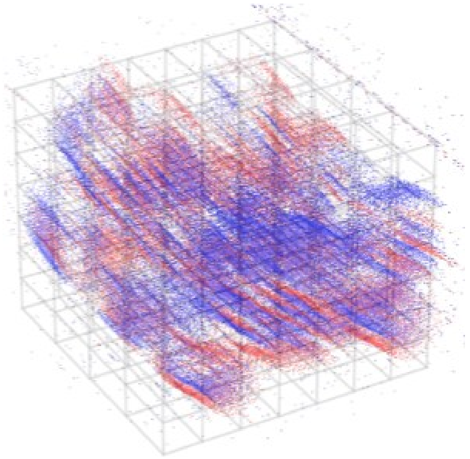


$$\dot{X}(t) = A(t)X(t) + B(t)U(t)$$
$$y(t) = C(t)X(t) + D(t)U(t)$$

State Space model: Linear complexity

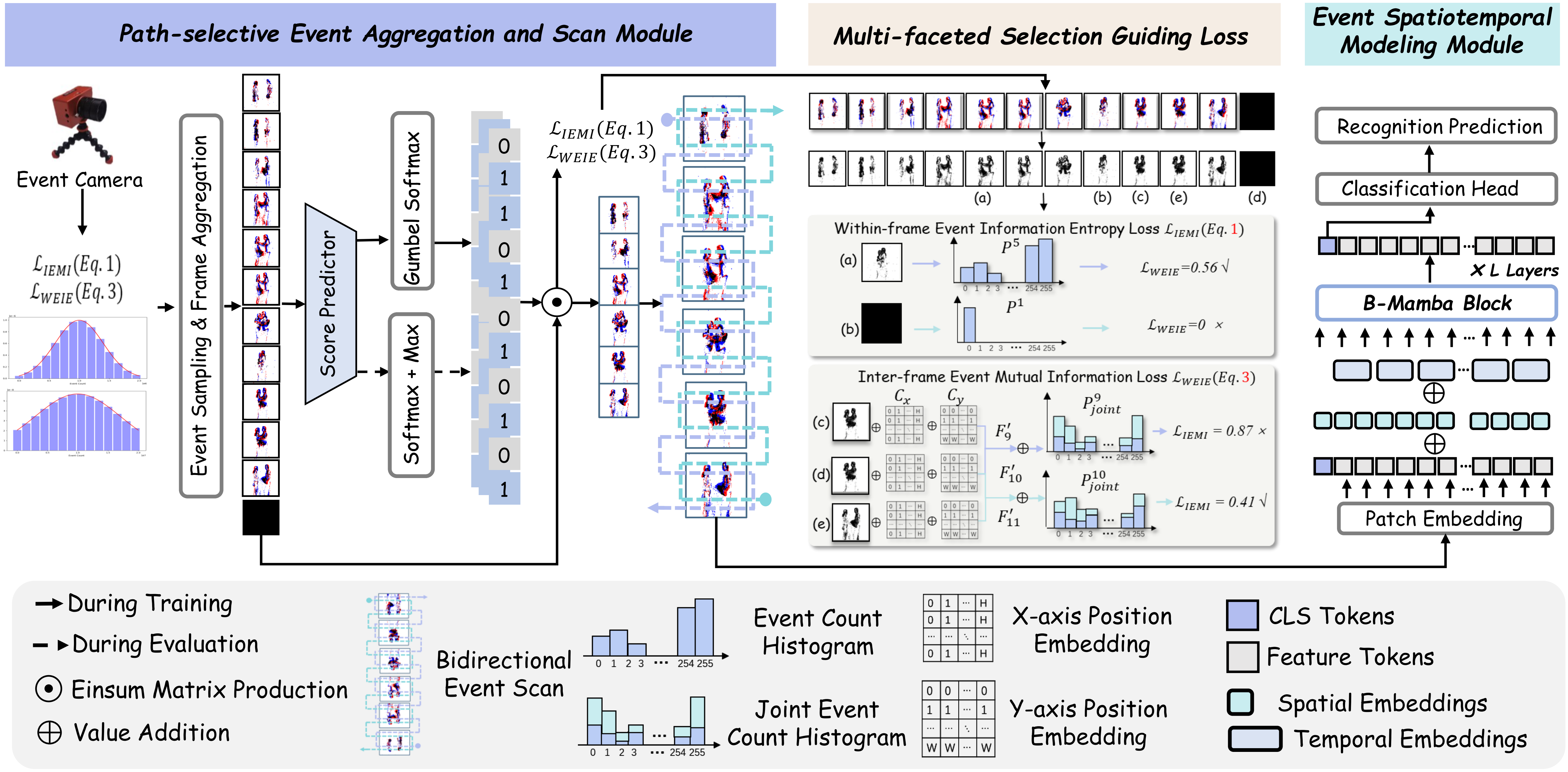
Suitable for Event Modeling

→



Event's spatiotemporal richness

Overall Framework of Our PASS



Path-selective Event Aggregation and Scan Module

Multi-faceted Selection Guiding Loss

Event Spatiotemporal Modeling Module

Event Camera

$\mathcal{L}_{IEMI}(Eq.1)$

$\mathcal{L}_{WEIE}(Eq.3)$

Event Sampling & Frame Aggregation

Score Predictor

Gumbel Softmax

Softmax + Max

0 1 1 0 1 0 1 0 1 0 1

$\mathcal{L}_{IEMI}(Eq.1)$

$\mathcal{L}_{WEIE}(Eq.3)$

(a) (b) (c) (e) (d)

Within-frame Event Information Entropy Loss $\mathcal{L}_{IEMI}(Eq.1)$

(a) P^5 $\mathcal{L}_{WEIE}$=0.56 √

(b) P^1 $\mathcal{L}_{WEIE}$=0 ×

Inter-frame Event Mutual Information Loss $\mathcal{L}_{WEIE}(Eq.3)$

C_x C_y

(c) (d) (e)

F'_9 F'_{10} F'_{11}

P^9_{joint} $\mathcal{L}_{IEMI}$ = 0.87 ×

P^{10}_{joint} $\mathcal{L}_{IEMI}$ = 0.41 √

Recognition Prediction

Classification Head

× L Layers

B-Mamba Block

Patch Embedding

During Training

During Evaluation

Einsum Matrix Production

Value Addition

Bidirectional Event Scan

Event Count Histogram

Joint Event Count Histogram

X-axis Position Embedding

Y-axis Position Embedding

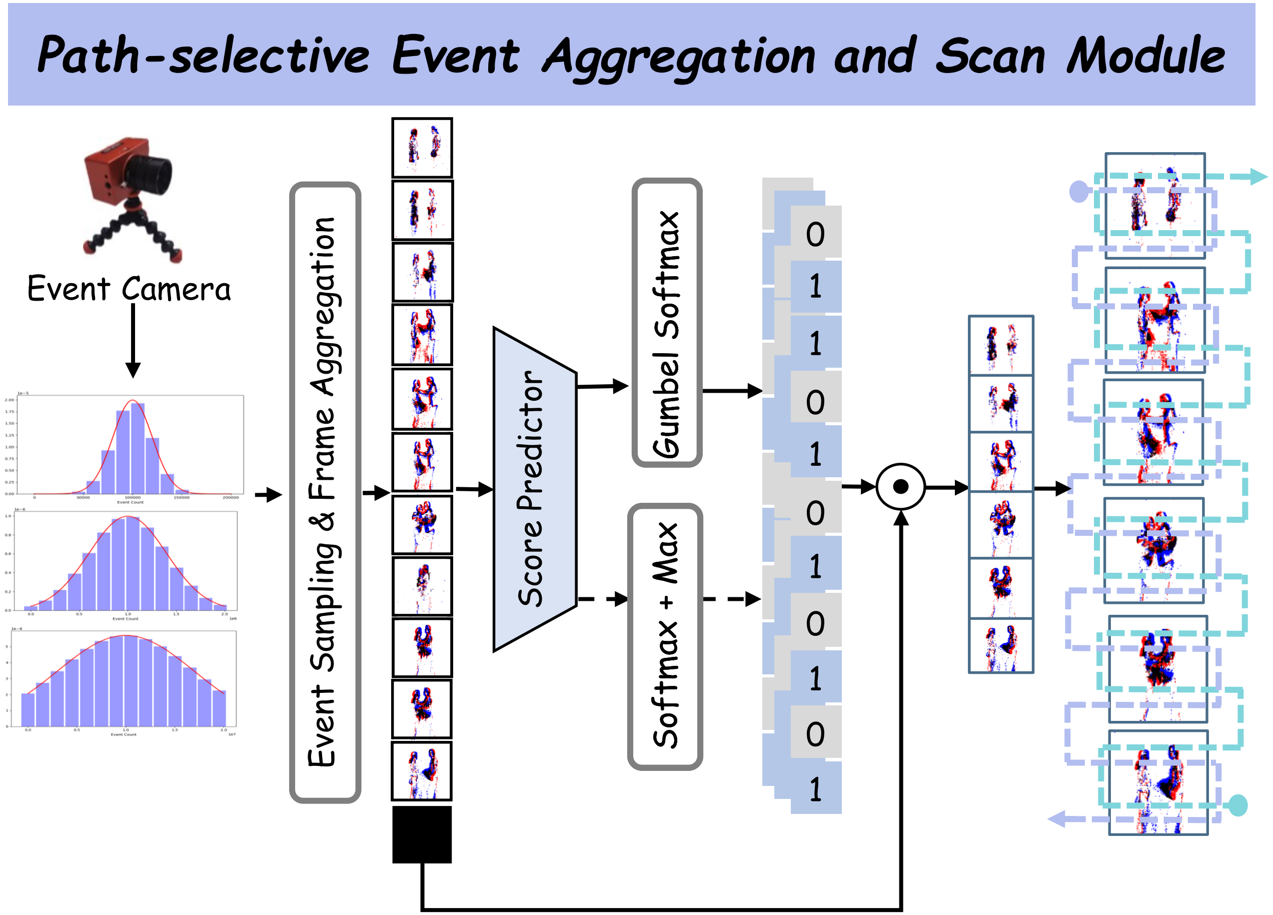
CLS Tokens

Feature Tokens

Spatial Embeddings

Temporal Embeddings

Methodology: Path-adaptive Event Aggregation and Scan (PEAS) Module


Path-selective Event Aggregation and Scan Module
Event Camera
Event Sampling & Frame Aggregation
Score Predictor
Gumbel Softmax
Softmax + Max
0
1
1
0
1
0
1
0
1
0
1

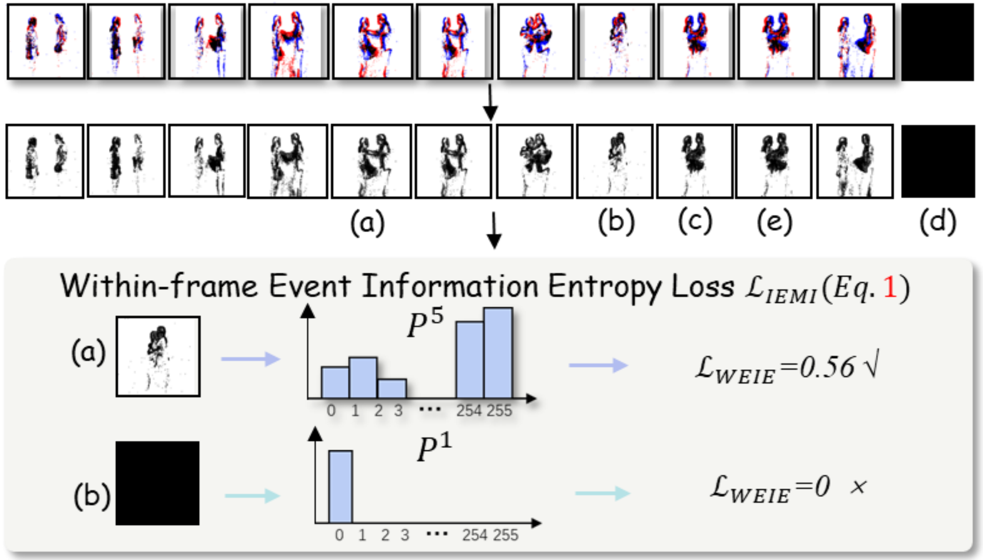
Converts asynchronous events into fixed-dimension sequence features:

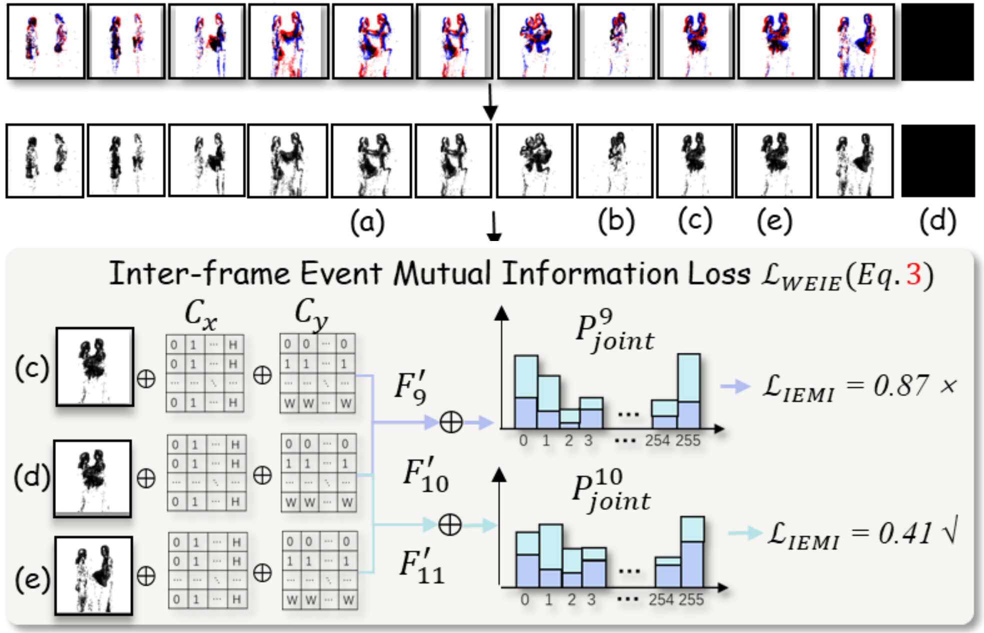
1. **Event Sampling/Aggregation**
2. **Adaptive Frame Selection**
3. **Bidirectional Scan**

Methodology: Multi-faceted Selection Guiding (MSG) Loss

Multi-faceted Selection Guiding Loss

Reduces randomness / redundancy in PEAS selection:

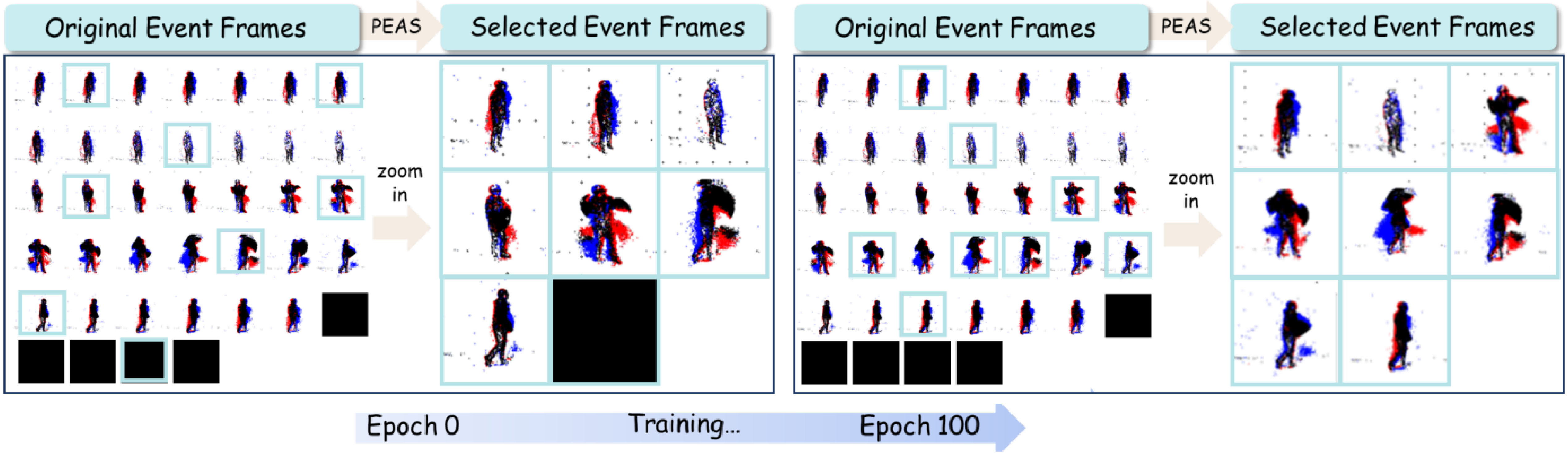

(a)
(b)
(c)
(e)
(d)
Within-frame Event Information Entropy Loss $\mathcal{L}_{IEMI}(Eq.1)$
(a)
P^5
0 1 2 3 ··· 254 255
$\mathcal{L}_{WEIE}$=0.56 √
(b)
P^1
0 1 2 3 ··· 254 255
$\mathcal{L}_{WEIE}$=0 ×


(a)
(b)
(c)
(e)
(d)
Inter-frame Event Mutual Information Loss $\mathcal{L}_{WEIE}(Eq.3)$
C_x
C_y
(c)
(d)
(e)
F'_9
F'_{10}
F'_{11}
P^9_{joint}
P^{10}_{joint}
0 1 2 3 ··· 254 255
$\mathcal{L}_{IEMI}$ = 0.87 ×
$\mathcal{L}_{IEMI}$ = 0.41 √

$\mathcal{L}_{WEIE}$ (Within-Frame Entropy):
Maximizes information per selected frame .

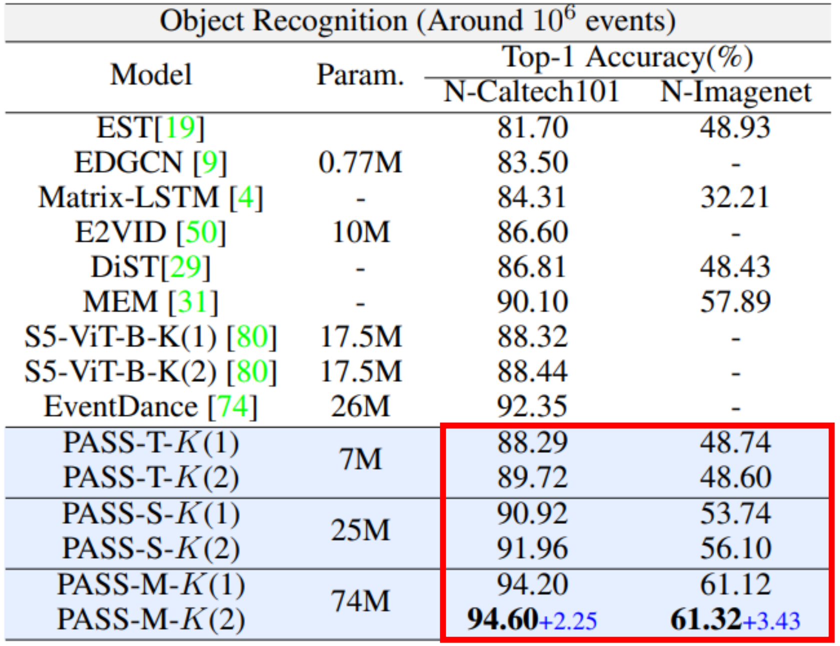
$\mathcal{L}_{IEMI}$ (Inter-Frame Mutual Information):
Minimizes redundancy between consecutive frames .

Qualitative Result: Multi-faceted Selection Guiding (MSG) Loss



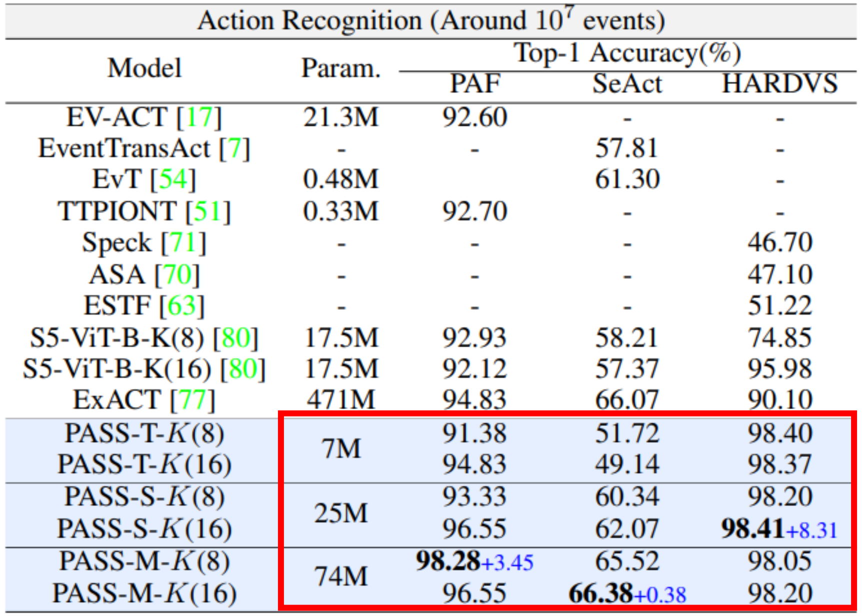
Quantitative Result: Event-based Recognition Experiment

Table 2: Comparison with previous methods for event-based object recognition.



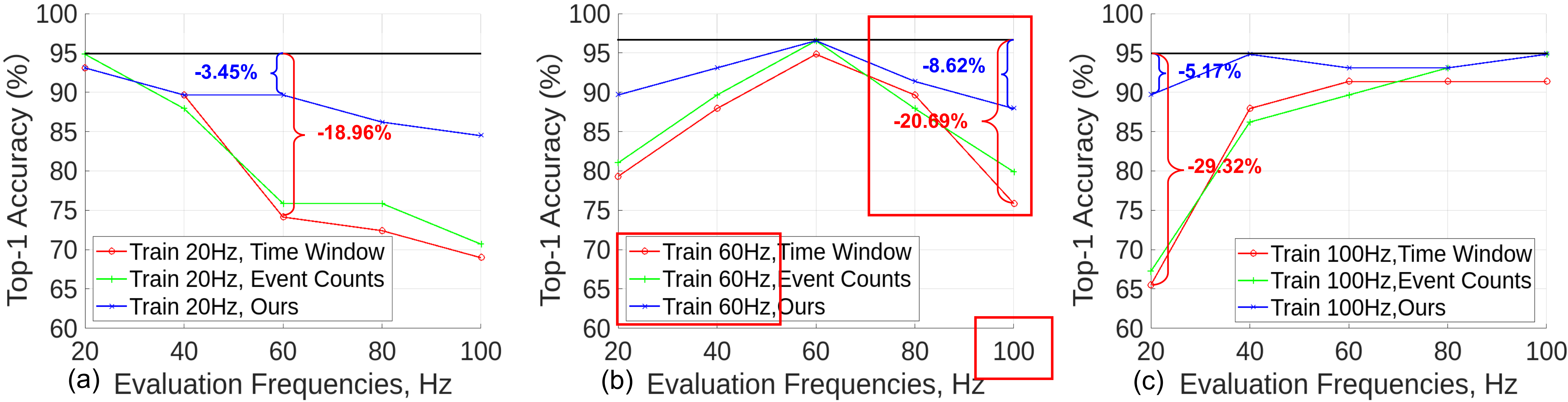
Object Recognition (Around 10^6 events)			
Model	Param.	Top-1 Accuracy(%)	
		N-Caltech101	N-Imagenet
EST[19]		81.70	48.93
EDGCN [9]	0.77M	83.50	-
Matrix-LSTM [4]	-	84.31	32.21
E2VID [50]	10M	86.60	-
DiST[29]	-	86.81	48.43
MEM [31]	-	90.10	57.89
S5-ViT-B-K(1) [80]	17.5M	88.32	-
S5-ViT-B-K(2) [80]	17.5M	88.44	-
EventDance [74]	26M	92.35	-
PASS-T-K(1)	7M	88.29	48.74
PASS-T-K(2)		89.72	48.60
PASS-S-K(1)	25M	90.92	53.74
PASS-S-K(2)		91.96	56.10
PASS-M-K(1)	74M	94.20	61.12
PASS-M-K(2)		**94.60**+2.25	**61.32**+3.43

Table 3: Comparison with previous methods for event-based action recognition.



Action Recognition (Around 10^7 events)				
Model	Param.	Top-1 Accuracy(%)		
		PAF	SeAct	HARDVS
EV-ACT [17]	21.3M	92.60	-	-
EventTransAct [7]	-	-	57.81	-
EvT [54]	0.48M		61.30	-
TTPIONT [51]	0.33M	92.70	-	-
Speck [71]	-	-	-	46.70
ASA [70]	-	-	-	47.10
ESTF [63]	-	-	-	51.22
S5-ViT-B-K(8) [80]	17.5M	92.93	58.21	74.85
S5-ViT-B-K(16) [80]	17.5M	92.12	57.37	95.98
ExACT [77]	471M	94.83	66.07	90.10
PASS-T-K(8)	7M	91.38	51.72	98.40
PASS-T-K(16)		94.83	49.14	98.37
PASS-S-K(8)	25M	93.33	60.34	98.20
PASS-S-K(16)		96.55	62.07	**98.41**+8.31
PASS-M-K(8)	74M	**98.28**+3.45	65.52	98.05
PASS-M-K(16)		96.55	**66.38**+0.38	98.20

Quantitative Result: Generalization results across Varying Inference Frequencies.



Experiment: Synthetic Datasets and Corresponding Event-based Recognition Results

- **ArDVS100**: 100 action transitions with diverse meta-actions. **[Synthetic]**
- **Tem-ArDVS100**: Same meta-actions as ArDVS100 but in different combinations, for fine-grained temporal recognition. **[Synthetic]**
- **Real-ArDVS10**: 10 real-world recorded action transitions, to test real-world generalization. **[Real-world]**

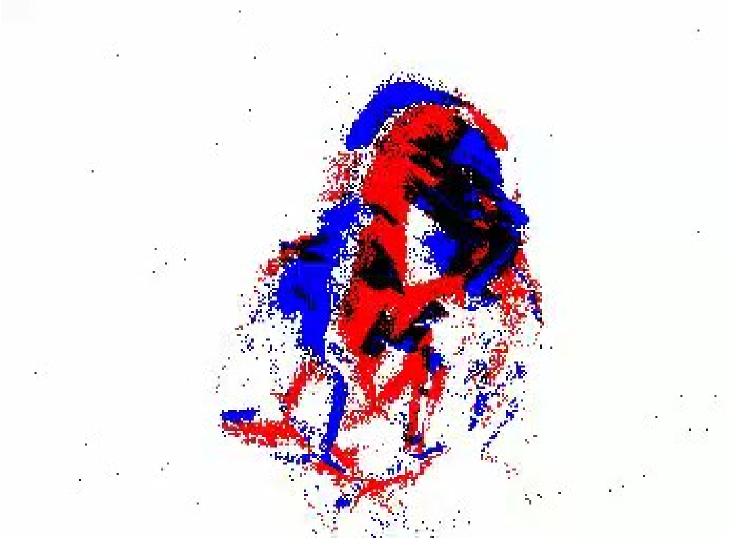
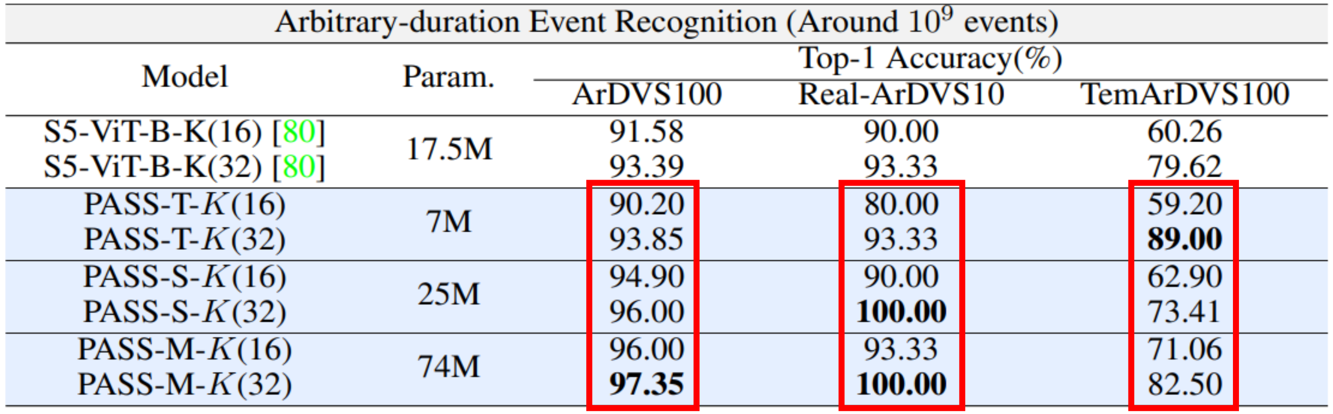


Illustration sample for ArDVS100

Table 4: Results of event-based action recognition with around 10^6 events).



Arbitrary-duration Event Recognition (Around 10^9 events)				
Model	Param.	Top-1 Accuracy(%)		
		ArDVS100	Real-ArDVS10	TemArDVS100
S5-ViT-B-K(16) [80]	17.5M	91.58	90.00	60.26
S5-ViT-B-K(32) [80]		93.39	93.33	79.62
PASS-T-K(16)	7M	90.20	80.00	59.20
PASS-T-K(32)		93.85	93.33	**89.00**
PASS-S-K(16)	25M	94.90	90.00	62.90
PASS-S-K(32)		96.00	**100.00**	73.41
PASS-M-K(16)	74M	96.00	93.33	71.06
PASS-M-K(32)		**97.35**	**100.00**	82.50

Ablation Study

Table 5: Ablation study on PEAS module & $\mathcal{L}_{MSG}$.

Settings	PAF ($K(16)$)	ArDVS100 ($K(16)$)
	Top1(%)	Top1(%)
No Sampling	92.90%	92.31%
Random Sampling	92.98%	92.23%
PEAS	93.33%	92.84%
PEAS + $\mathcal{L}_{MSG}$	**94.83%**	**93.85%**

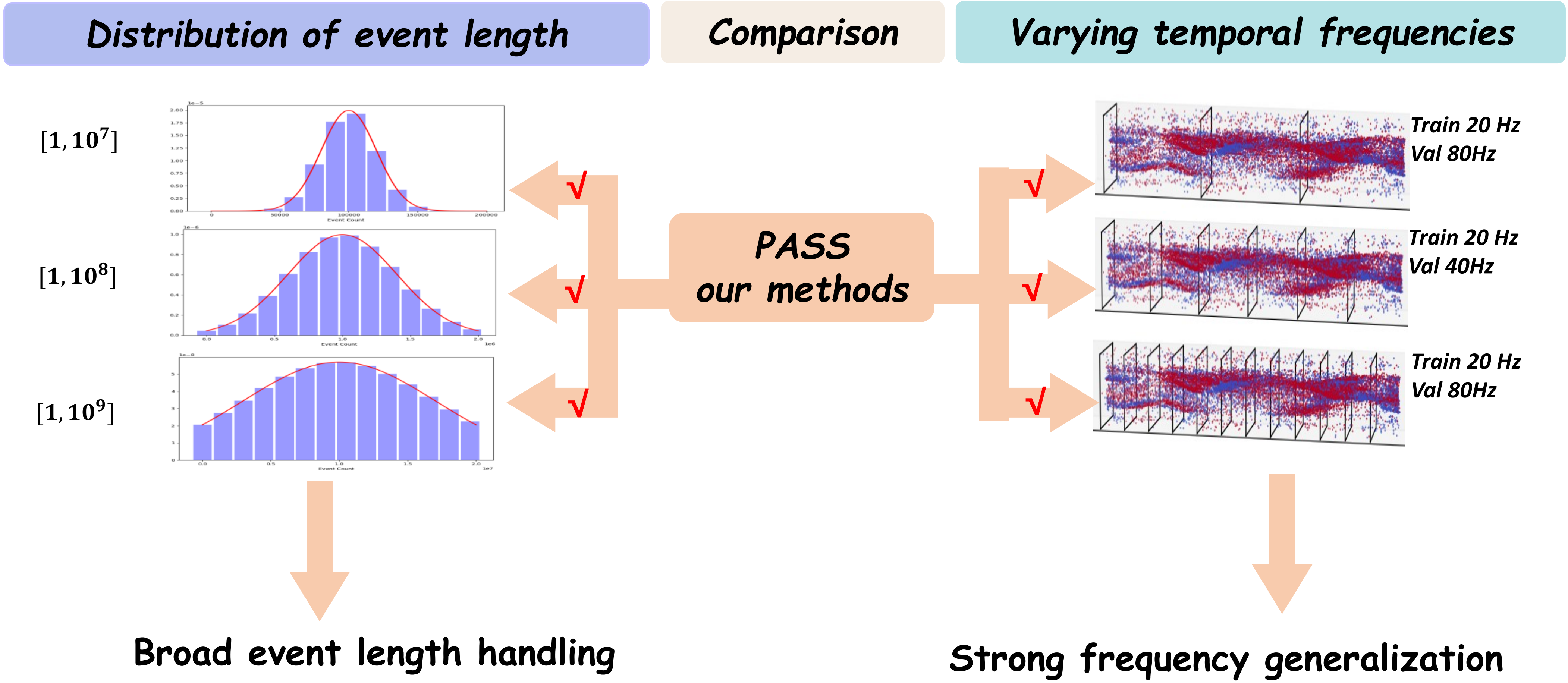
Table 6: Ablation study on $\mathcal{L}_{MSG}$.

$\mathcal{L}_{MSG}$			PAF($K(16)$)
$\mathcal{L}_{CLS}$	$\mathcal{L}_{IEMI}$	$\mathcal{L}_{WEIE}$	Top1(%)
✓	✗	✗	92.98%
✓	✓	✗	93.75%+0.77
✓	✓	✓	**94.83%**+1.85

Table 7: Ablation study on event representation.

Representation	N-Caltech101 ($K(1)$)		PAF ($K(16)$)	
	Top1(%)	Top5(%)	Top1(%)	Top5(%)
Frame(Gray) [75]	90.48%	97.53%	93.33%	100.00%
Frame(RGB) [75]	90.94%	97.82%	94.83%	100.00%
Voxel [11]	90.19%	97.02%	92.47%	100.00%
TBR [27]	90.24%	97.13%	91.72%	100.00%
EST [19]	90.54%	97.66%	93.04%	100.00%

Summary



Distribution of event length

Comparison

Varying temporal frequencies

$[1, 10^7]$

$[1, 10^8]$

$[1, 10^9]$

√ √ √

PASS
our methods

√ √ √

Train 20 Hz
Val 80Hz

Train 20 Hz
Val 40Hz

Train 20 Hz
Val 80Hz

Broad event length handling

Strong frequency generalization

NEURAL INFORMATION
PROCESSING SYSTEMS
2025

香港科技大学（广州）
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY
粤港澳大湾区数字经济研究院（福田）

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

PASS: Path-selective State Space Model for Event-based Recognition

Thank you for your listening!

Project Page: https://jiazhou-garland.github.io/PASS_Homepage/