[ASP-DAC 2023] Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Yimeng Zhang*
Michigan State University, USA

Akshay Karkal Kamath*
Georgia Institute of Technology, USA

Qiucheng Wu*
UC, Santa Barbara, USA

Zhiwen Fan*
University of Texas at Austin, USA

Wuyang Chen
University of Texas at Austin, USA

Zhangyang Wang
University of Texas at Austin, USA

Shiyu Chang
UC, Santa Barbara, USA

Sijia Liu
Michigan State University, USA

Cong Hao
Georgia Institute of Technology, USA

MICHIGAN STATE UNIVERSITY

ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

OPTML

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Background

- The efficiency of “compressed” models are evaluated **without considering the practical hardware platform**, such as low-power FPGAs.

Xilinx Alveo U50@75W

Xilinx ZCU104@5W

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Background

- The efficiency of "compressed" models are evaluated **without considering the practical hardware platform**, such as low-power FPGAs.

- existing accelerators are evaluated on the ImageNet dataset with small input image sizes and **do not scale to real-world High-Definition (HD) video frames.**

MICHIGAN STATE UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Background

- The efficiency of “compressed” models are evaluated **without considering the practical hardware platform**, such as low-power FPGAs.

- existing accelerators are evaluated on the ImageNet dataset with small input image sizes and **do not scale to real-world High-Definition (HD) video frames.**

- **Multi-Object Tracking (MOT)** is the focus.

MICHIGAN STATE UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design
for Ultra-Light Video Intelligence on Edge Devices

Method

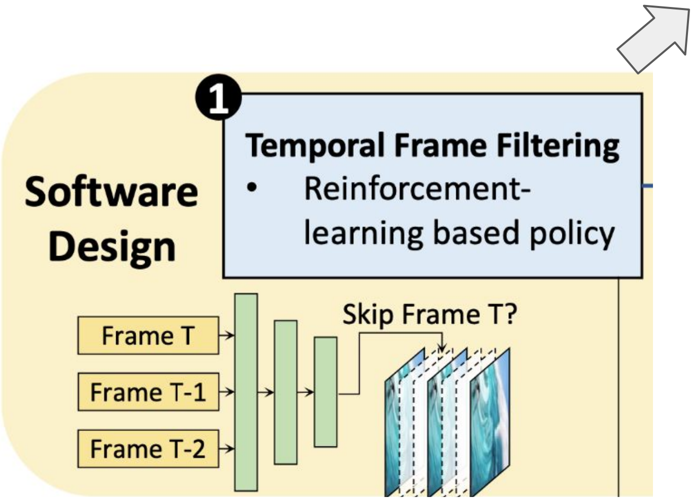
MICHIGAN STATE
UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

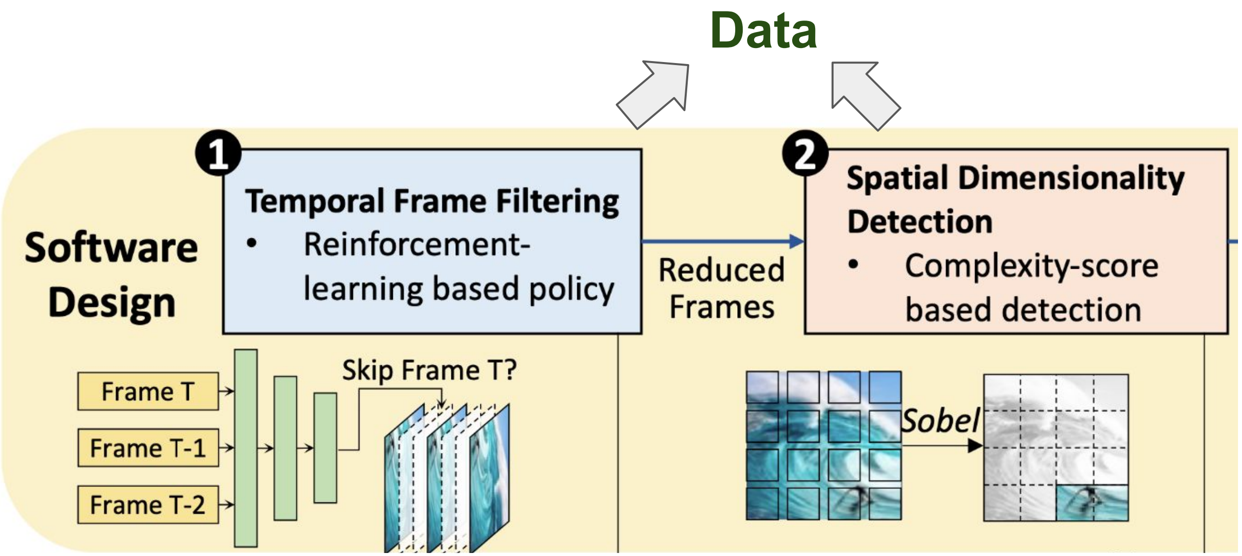
Method

Data



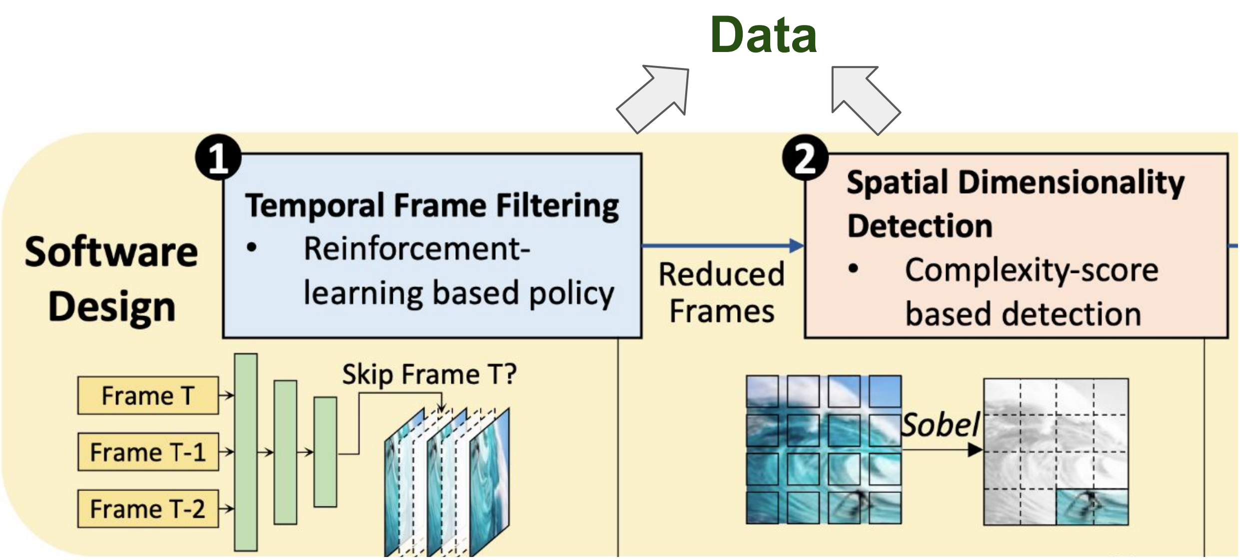
Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

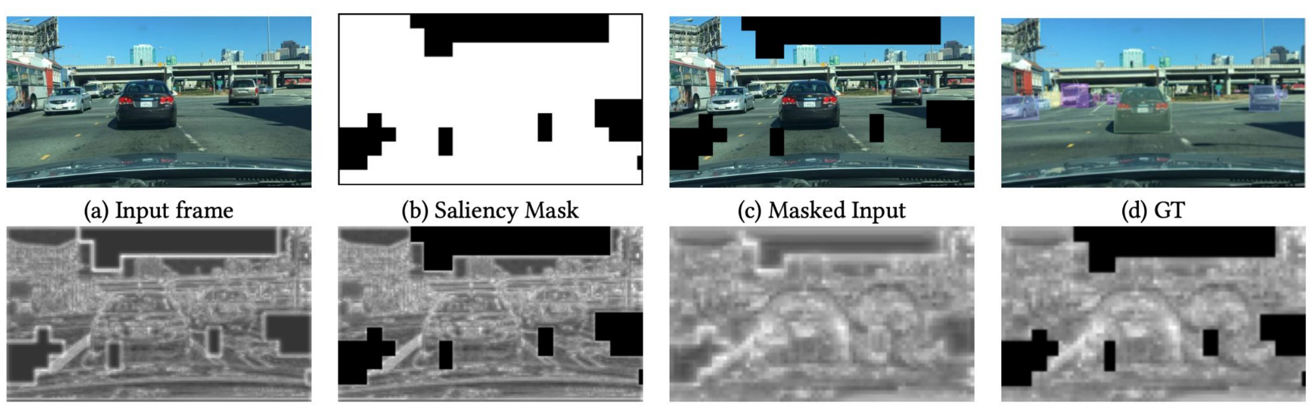
Method



Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Method





(a) Input frame (b) Saliency Mask (c) Masked Input (d) GT

(e) FM(2,0) (f) FM(2,0) w/ Mask (g) FM(3,0) (h) FM(3,0) w/ Mask

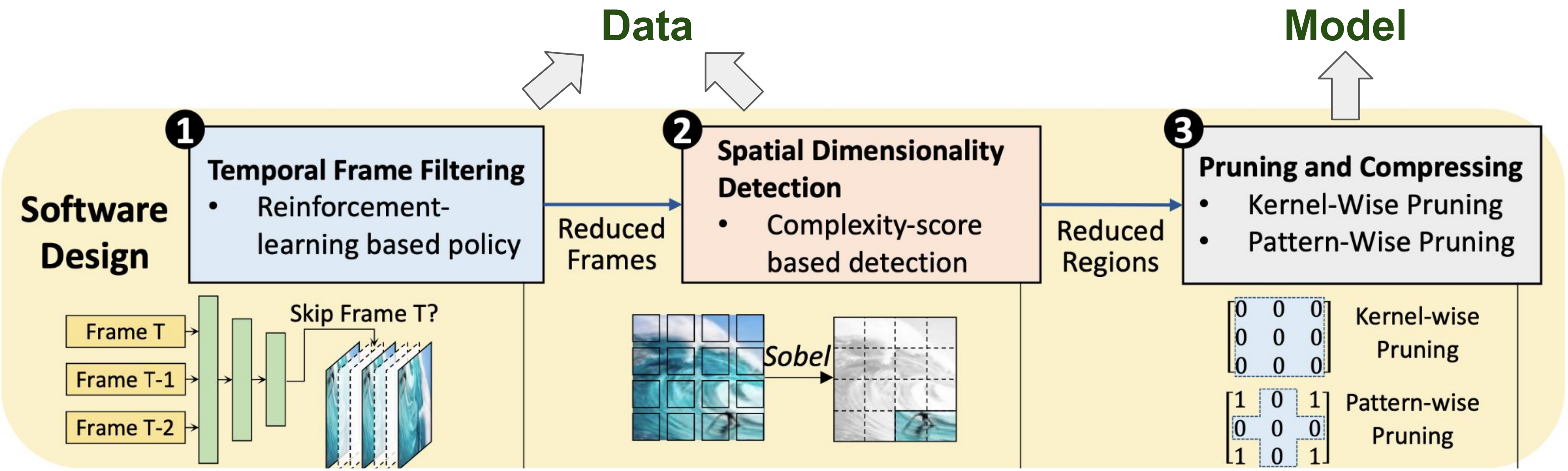
Patch Drop for Input Frame and intermediate features

MICHIGAN STATE UNIVERSITY

OPTML

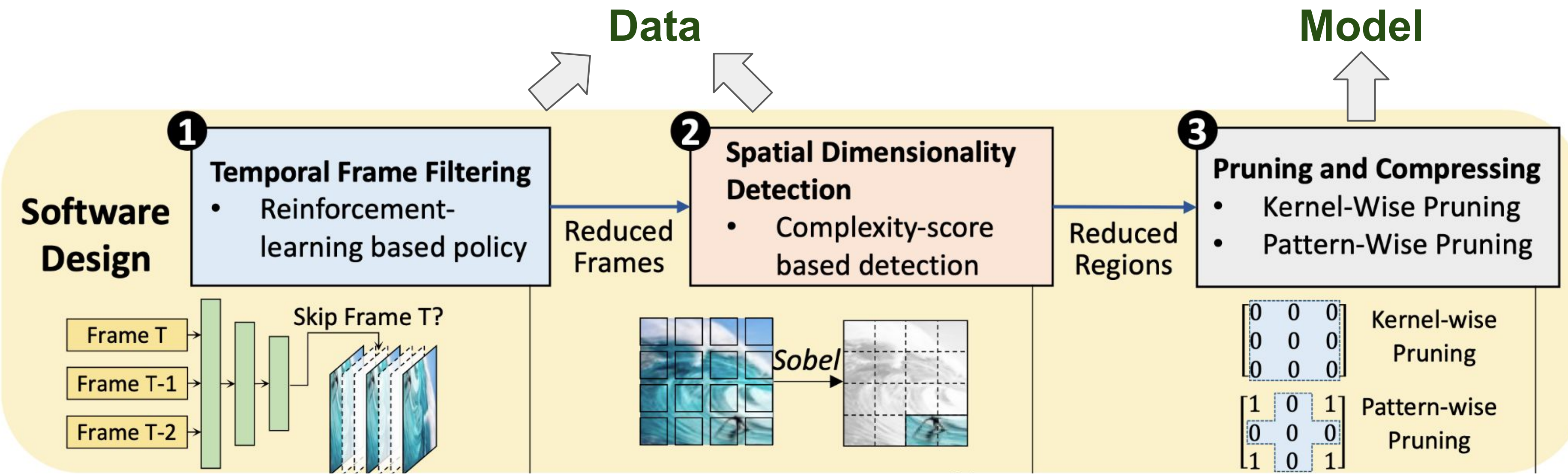
Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

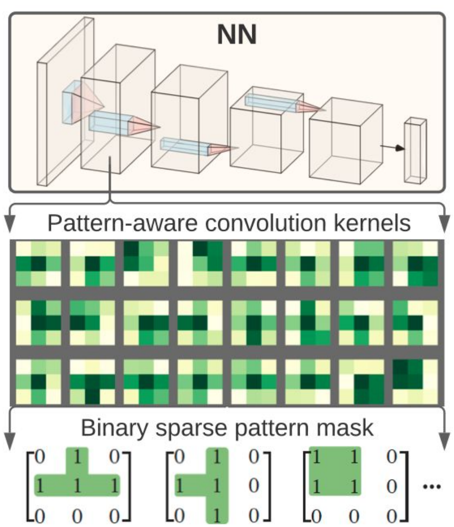
Method



Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Method





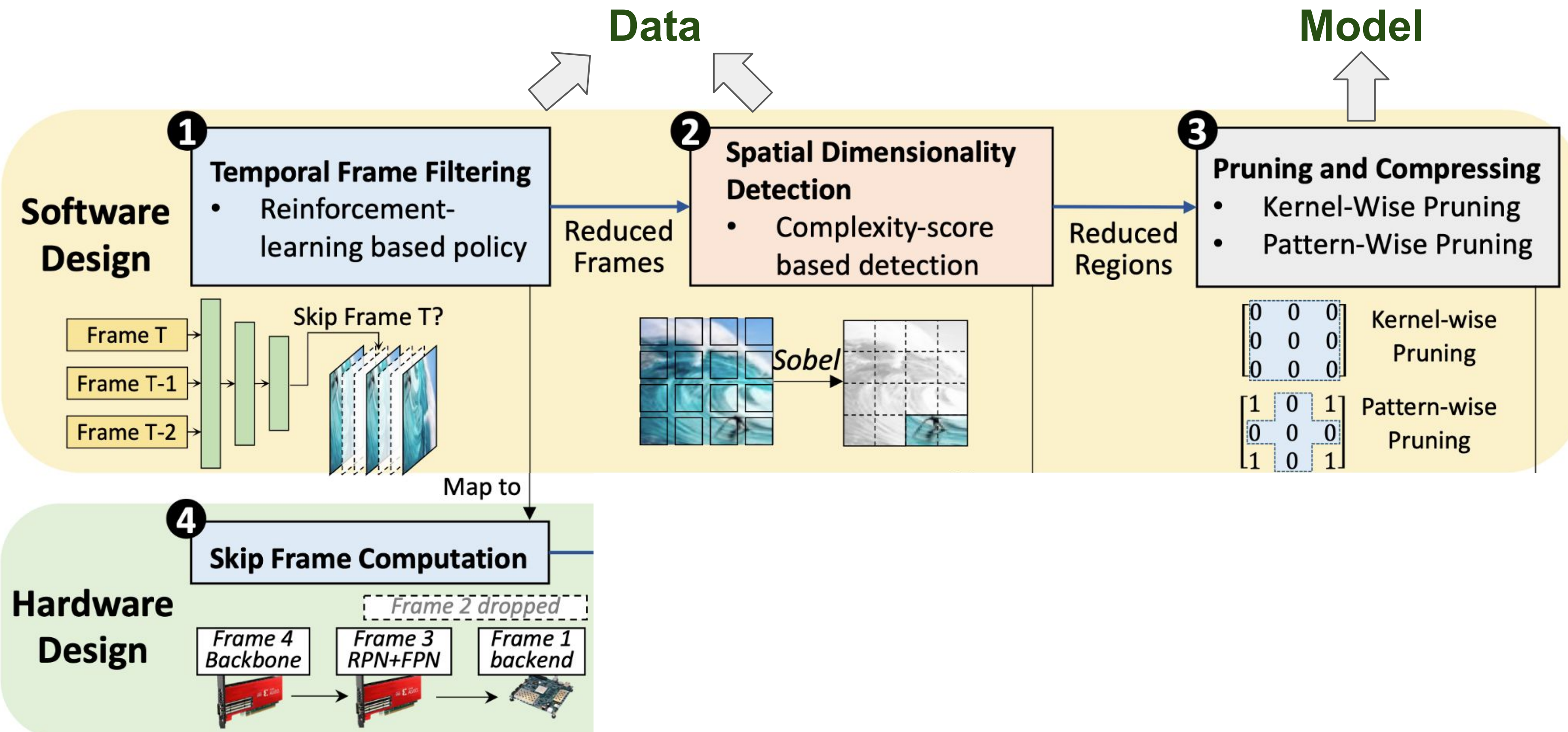
- **pre-define** irregular sparse patterns for 3 × 3 kernels
- leverage them to conduct irregular but **pattern-aware** weight pruning

MICHIGAN STATE UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Method

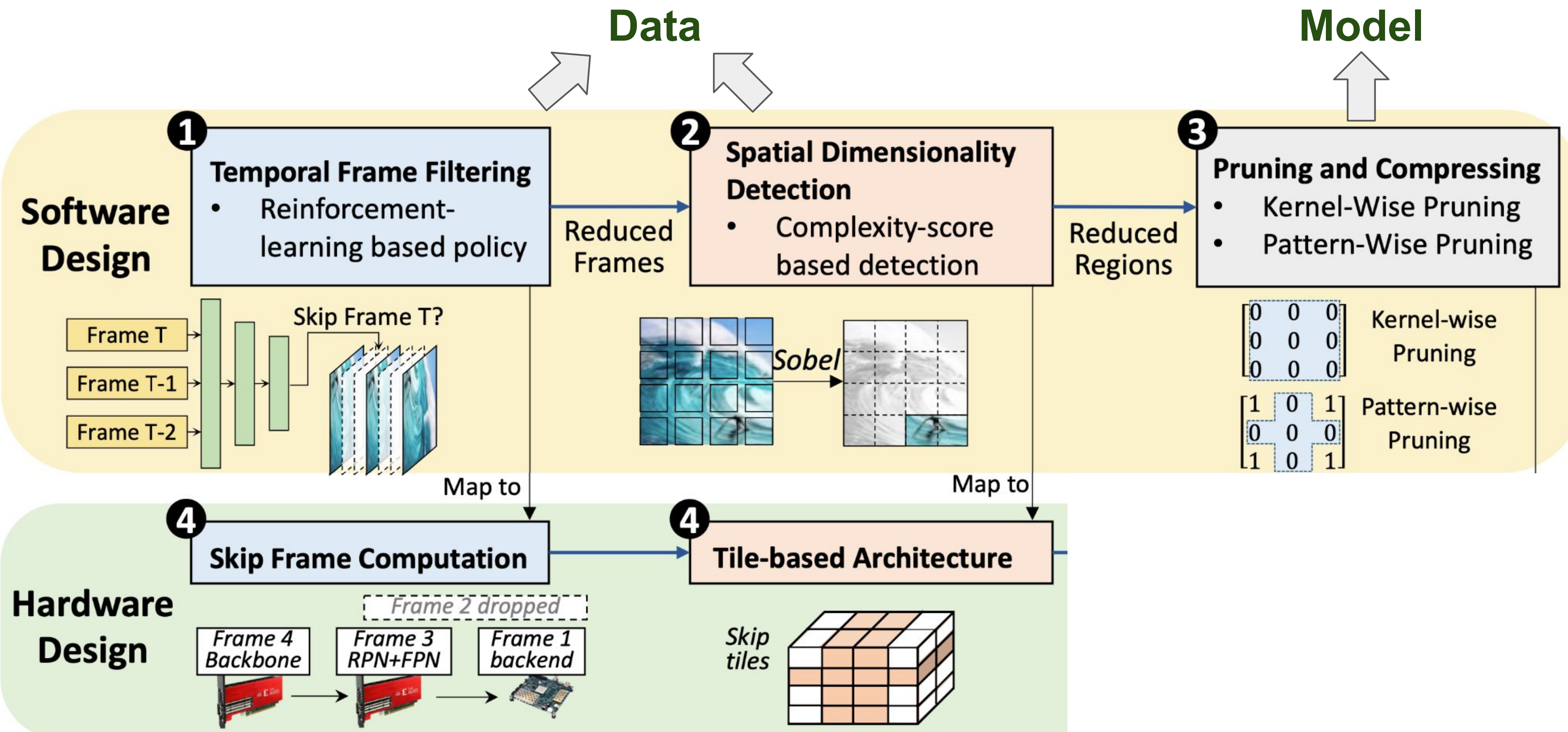


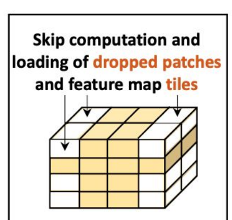
MICHIGAN STATE UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Method



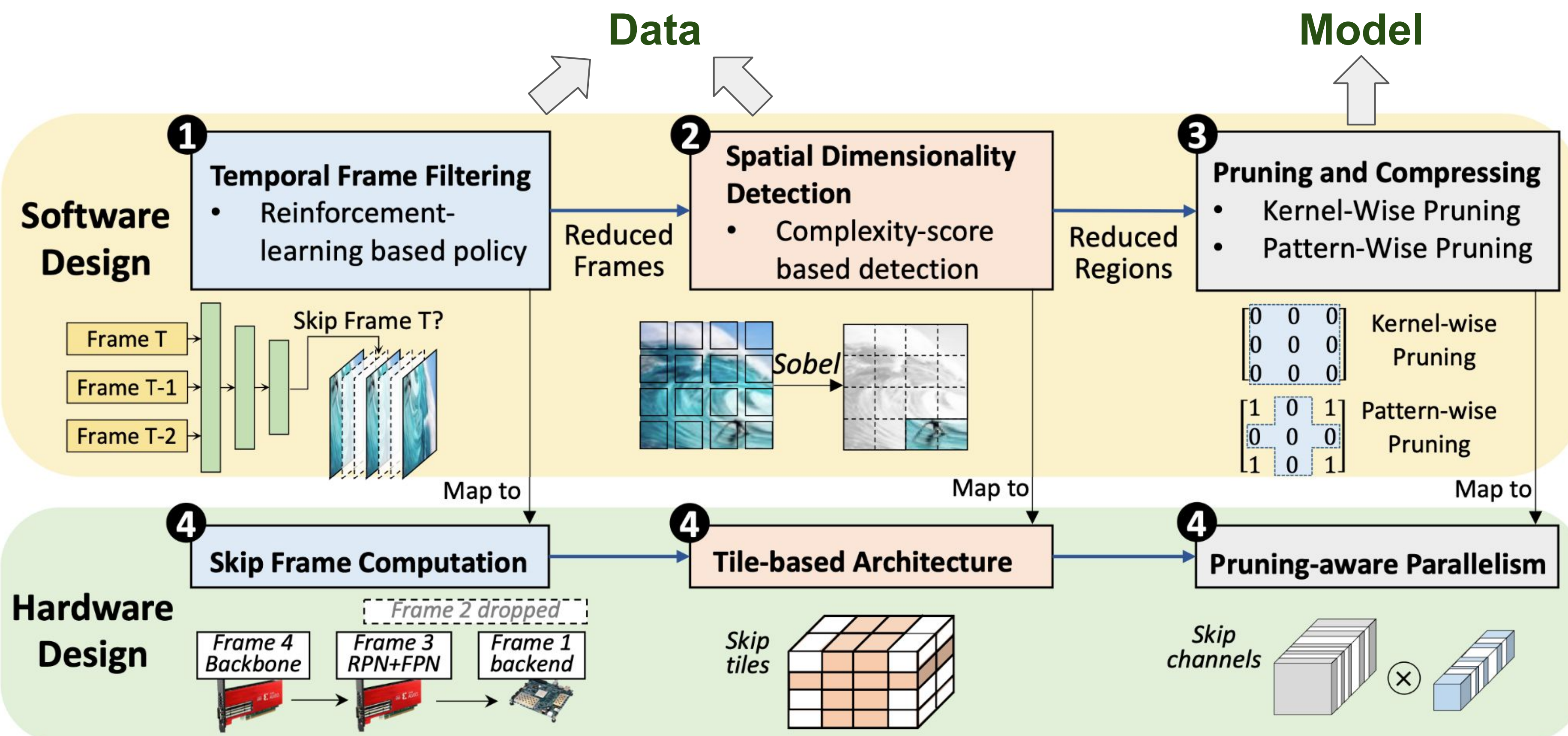


MICHIGAN STATE UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design
for Ultra-Light Video Intelligence on Edge Devices

Method



MICHIGAN STATE UNIVERSITY

OPTML

Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices

Performance

Methods	Data/model compression		Metrics					
	Data reduction	Pruning	IDF1 (↑)	MOTA (↑)	Latency (↓)	EFR (↑)	Power (↓)	Energy Efficiency (↓)
QDTrack (GPU baseline)	×	×	0.714	0.637	60.9	22.5	296 W	13.2 J/frame
QDTrack on FPGA	×	×	0.714	0.637	554.7	1.8	50.8 W	28.2 J/frame
Variant: Frame + patch drop	(40%, 20%)	×	0.71	0.628	443.8	2.3	50.8 W	22.0 J/frame
Tri-design (ours)	(40%, 20%)	90%	0.704	0.617	44.4	37.6	50.8 W	1.35 J/frame
Improv. over GPU baseline	—	—	-1.40%	-3.14%	**1.37×**	**1.67×**	**5.83×**	**9.78×**
Improv. over FPGA baseline	—	—	-1.40%	-3.14%	**12.5×**	**20.9×**	—	**20.9×**

Implementation Details

- 40% temporal **frame** dropping
- 20% spatial **patch** dropping
- 90% **model** pruning

Hardware Metrics given by on-board latency in the unit of millisecond

- effective frame rate (**EFR**) → FPS
- **power** in the unit of Watt

Accuracy

- ID F1 Score (**IDF1**)
- Multi-Object Tracking Accuracy (**MOTA**)

MICHIGAN STATE
UNIVERSITY

terima kasih

mulțumesc

ありがとう

谢谢 ngiyabonga suksema

Met dank
obrigada

Thank You

baie dankie
molte grazie

merci 감사합니다

Danke schön!

obrigado

謝謝

Благодарность شكرًا

dank u

mahalo

gracias

Спасибі Dziękuję

tusind tak

MICHIGAN STATE
UNIVERSITY

OPTML