UCI Georgia Tech Meta

MLSys | 2023

Sixth Conference on Machine Learning and Systems

XRBench: An Extended Reality (XR) Machine Learning Benchmark Suite for the Metaverse

Hyoukjun kwon (presenter), Krishnakumar Nair, *Jamin Seo, *Jason Yik, Debabrata Mohapatra, Dongyuan Zhan, Jinook Song, Peter Capak, Peizhao Zhang, Peter Vajda, Colby Banbury, Mark Mazumder, Liangzhen Lai, Ashish Sirasao, Tushar Krishna, Harshit Khaitan, Vikas Chandra, Vijay Janapa Reddi

XRBench
OPEN ML BENCHMARK FOR AR/VR

Project Homepage: https://xrbench.ai

Project Github: https://github.com/XRBench

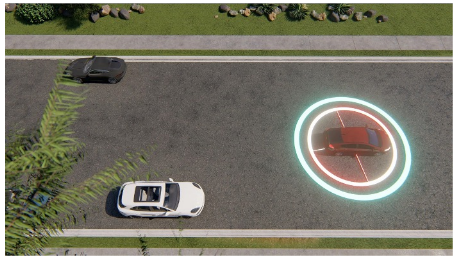
* Equal Contribution

Outline

- New ML Workload: Realtime **MTMM** (**M**ulti-**T**ask **M**ulti-**M**odel)
- XRBench: Realtime MTMM Benchmark Suite in XR (AR/VR)
- New Scoring Metric for Real-time MTMM
- Case Studies
- Conclusion

ML Workload Taxonomy

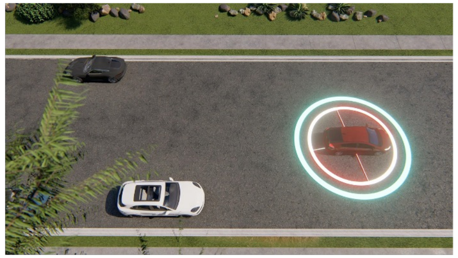
Model Concurrency?

Inter-model Dependency?	No	Yes
No	**Example: MLPerf-inference** MLPerf	**Example: Multi-tenancy in data centers** Recommendation Chatbot Video Analysis ...
Yes	**Example: Smart Speaker** If detected Keyword Detection → Speech Recognition	**Example: AR/VR** AR/VR  Autonomous Driving ... **Multi-task Multi-model (MTMM) ML Workloads**

Characteristics of Real-time MTMM ML Workloads

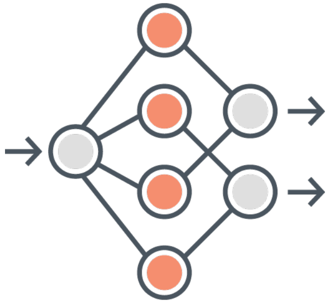
Real-time
MMMT
Applications

AR/VR



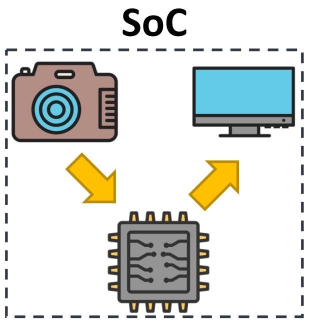
Autonomous Driving

...



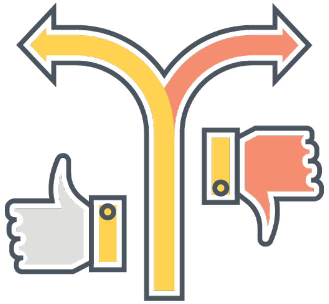
Concurrent and Cascaded Models

Real-time Processing



SoC-level Pipeline

Multi-Modal Inputs and Models



User-input-driven Dynamism

Context-driven Workloads

To guide ML system design for this new class of ML workloads,
we need a well-defined benchmark driven by practical use case with all the characteristics

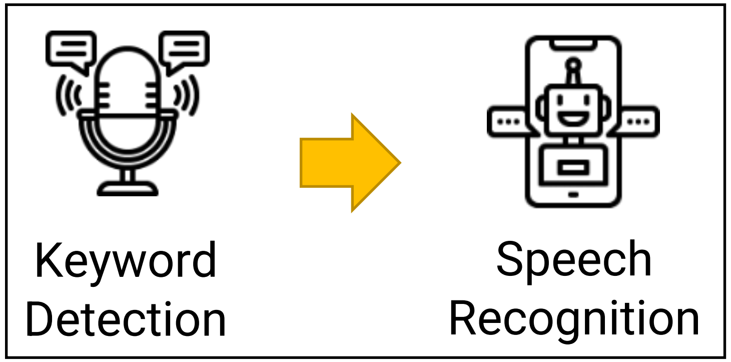
Outline

- New ML Workload: Realtime **MTMM** (**M**ulti-**T**ask **M**ulti-**M**odel)
- ➡ XRBench: Realtime MTMM Benchmark Suite in XR (AR/VR)
- New Scoring Metric for Real-time MTMM
- Case Studies
- Conclusion

XRBench v0.1: Unit Models

- **Three key task classes and unit models in XRBench**
  - **1) User-device Interaction**

Dependency

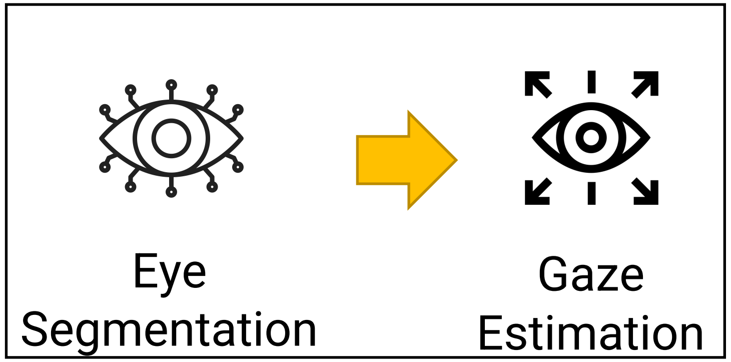

Keyword Detection

Speech Recognition

Speech-based Interaction

Hand Tracking

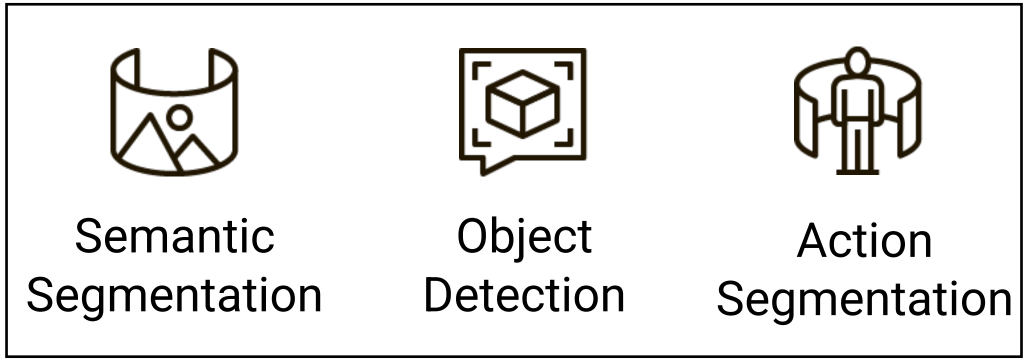
Hand-based Interaction


Eye Segmentation

Gaze Estimation

Eye-based Interaction

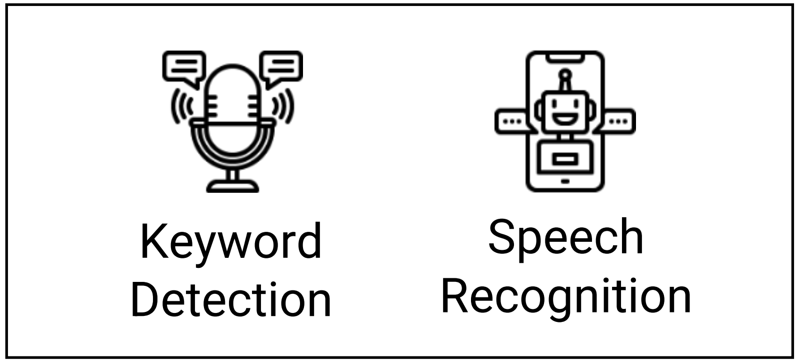
  - **2) User Context Understanding**


Semantic Segmentation

Object Detection

Action Segmentation

Vision-based Context Understanding

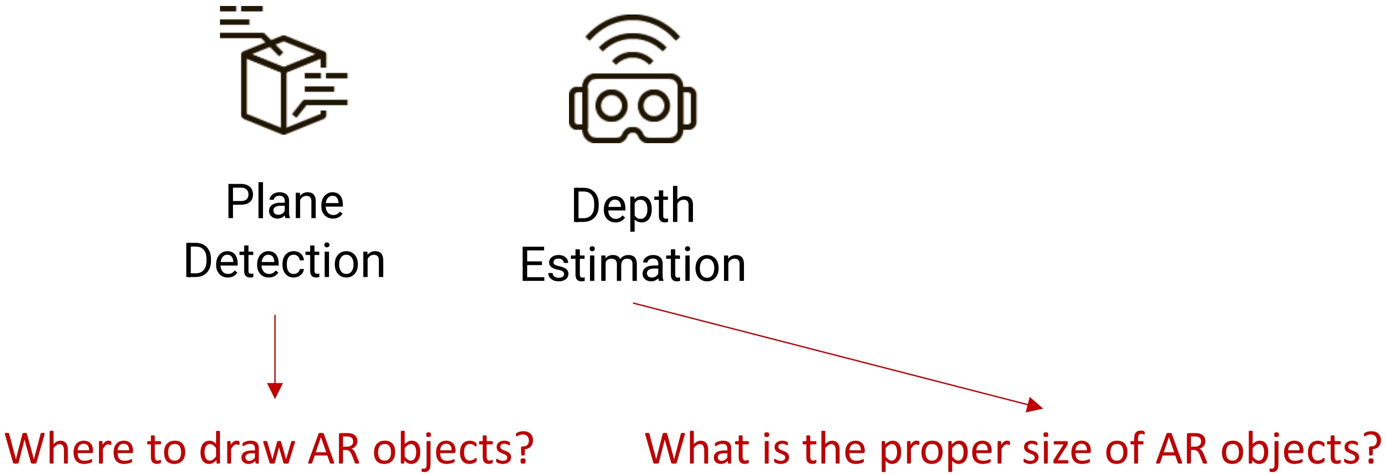

Keyword Detection

Speech Recognition

Audio-based Context Understanding

XRBench v0.1: Unit Models

- **Three key task classes and unit models in XRBench**
  - **3) World-locking: Identify how to draw AR objects on real world scenes**



Plane Detection

Depth Estimation

Where to draw AR objects?

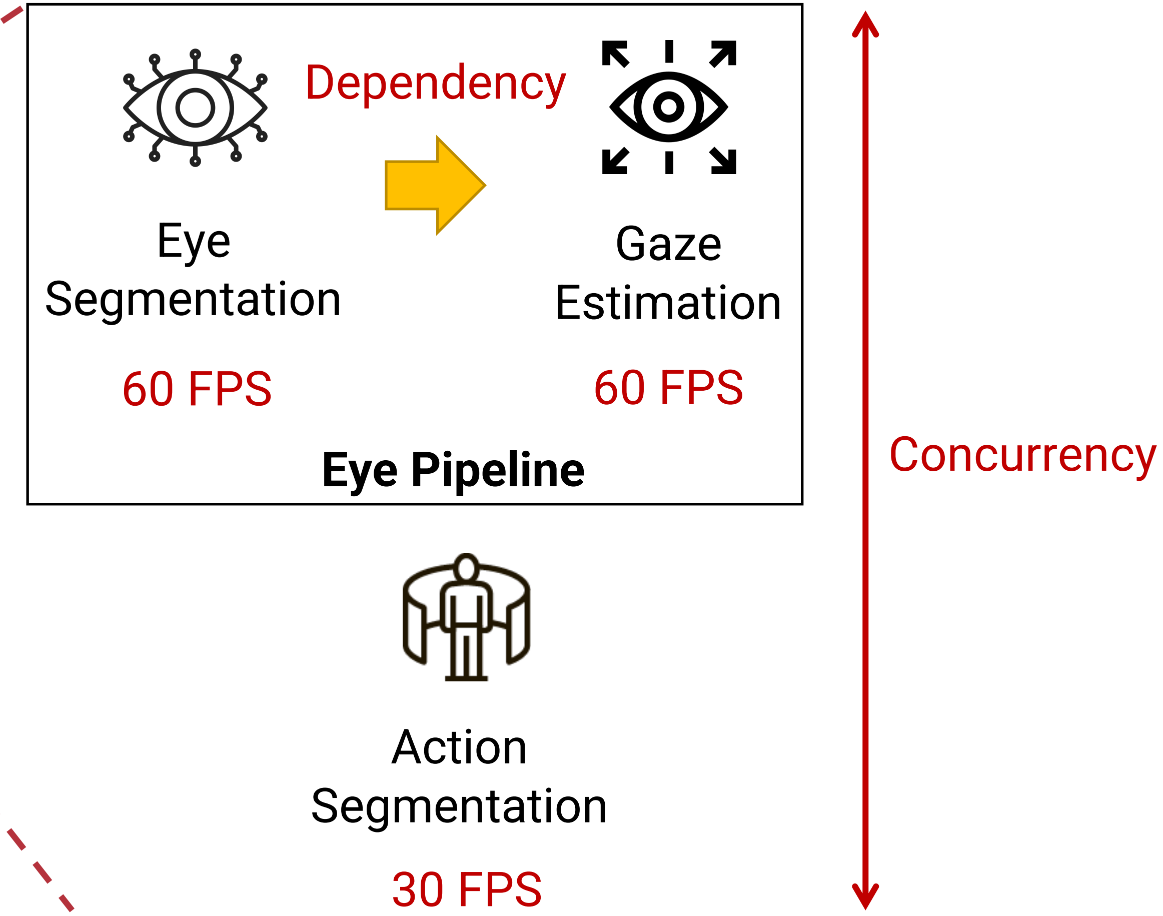
What is the proper size of AR objects?

Note: This covers a subset of AR/VR workloads. More to be updated in the future version!

Usage Scenarios: How to combine unit models?

- **Example: Social Interaction B Scenario in XRBench**

e.g., action-based AR emoji drawing during in-person conversation



Dependency

Eye Segmentation

60 FPS

Gaze Estimation

60 FPS

Eye Pipeline

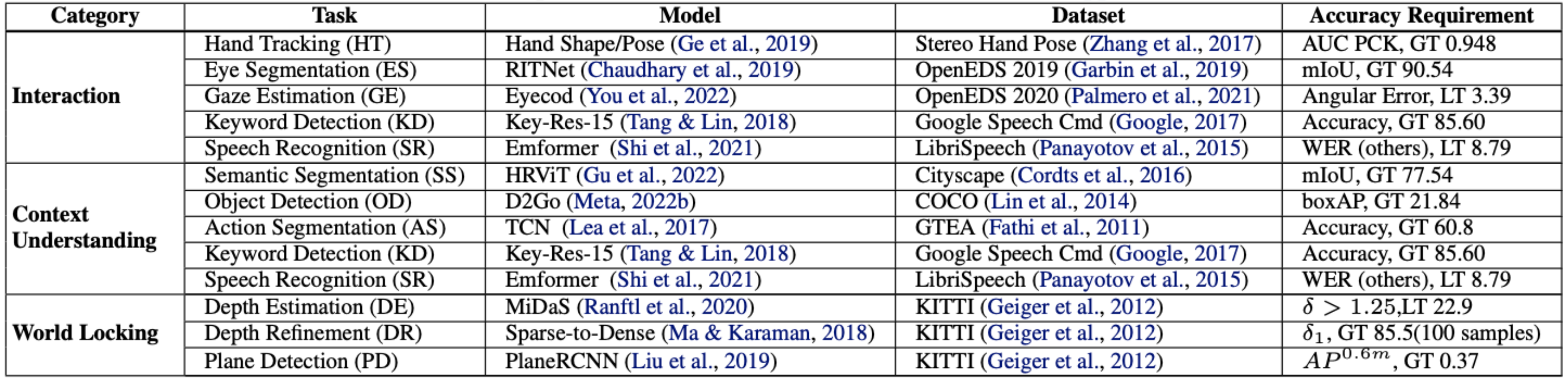
Action Segmentation

30 FPS

Concurrency

XRBench v0.1: Overview

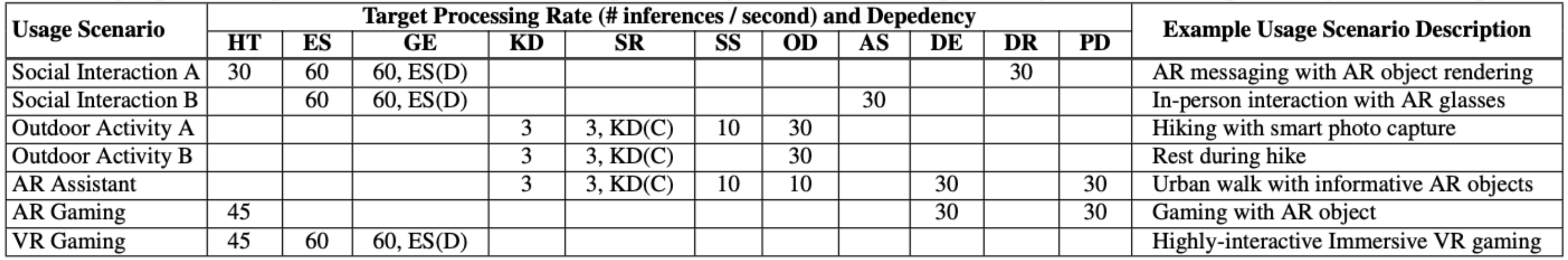
- **11 Unit Models**



Category	Task	Model	Dataset	Accuracy Requirement
Interaction	Hand Tracking (HT)	Hand Shape/Pose (Ge et al., 2019)	Stereo Hand Pose (Zhang et al., 2017)	AUC PCK, GT 0.948
	Eye Segmentation (ES)	RITNet (Chaudhary et al., 2019)	OpenEDS 2019 (Garbin et al., 2019)	mIoU, GT 90.54
	Gaze Estimation (GE)	Eyecod (You et al., 2022)	OpenEDS 2020 (Palmero et al., 2021)	Angular Error, LT 3.39
	Keyword Detection (KD)	Key-Res-15 (Tang & Lin, 2018)	Google Speech Cmd (Google, 2017)	Accuracy, GT 85.60
	Speech Recognition (SR)	Emformer (Shi et al., 2021)	LibriSpeech (Panayotov et al., 2015)	WER (others), LT 8.79
Context Understanding	Semantic Segmentation (SS)	HRViT (Gu et al., 2022)	Cityscape (Cordts et al., 2016)	mIoU, GT 77.54
	Object Detection (OD)	D2Go (Meta, 2022b)	COCO (Lin et al., 2014)	boxAP, GT 21.84
	Action Segmentation (AS)	TCN (Lea et al., 2017)	GTEA (Fathi et al., 2011)	Accuracy, GT 60.8
	Keyword Detection (KD)	Key-Res-15 (Tang & Lin, 2018)	Google Speech Cmd (Google, 2017)	Accuracy, GT 85.60
	Speech Recognition (SR)	Emformer (Shi et al., 2021)	LibriSpeech (Panayotov et al., 2015)	WER (others), LT 8.79
World Locking	Depth Estimation (DE)	MiDaS (Ranftl et al., 2020)	KITTI (Geiger et al., 2012)	$\delta > 1.25$,LT 22.9
	Depth Refinement (DR)	Sparse-to-Dense (Ma & Karaman, 2018)	KITTI (Geiger et al., 2012)	δ_1, GT 85.5(100 samples)
	Plane Detection (PD)	PlaneRCNN (Liu et al., 2019)	KITTI (Geiger et al., 2012)	$AP^{0.6m}$, GT 0.37

- **Considerations for Model Selection**
  - Realistic workload: Recommendation from ML engineers/researchers in industry
  - Model efficiency: Consider battery / compute power-limited wearable devices
  - Model performance: Reported accuracy, mIoU, etc.

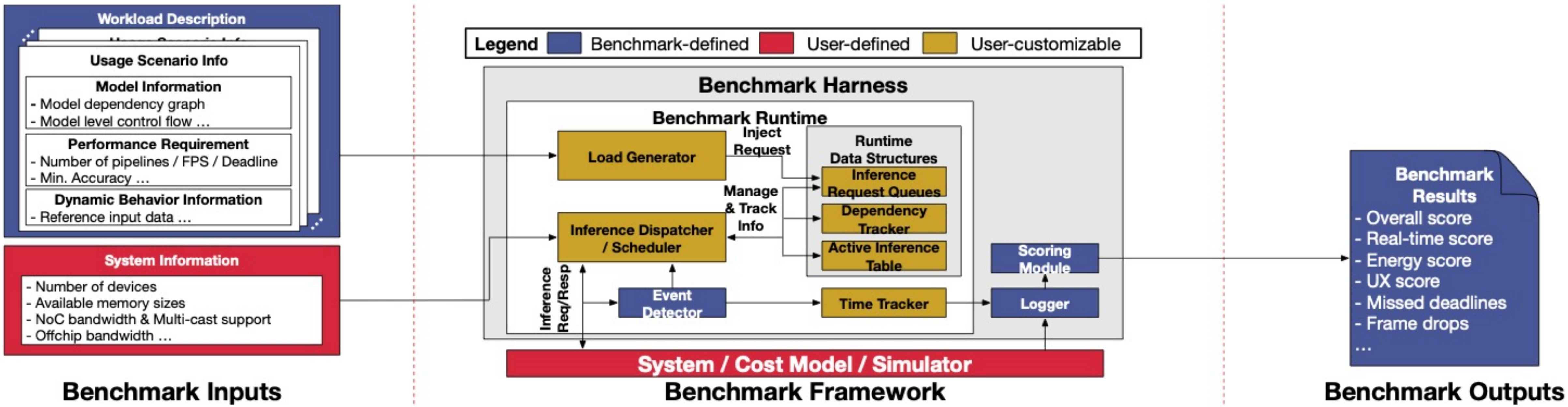
- **7 Usage Scenarios**



Usage Scenario	Target Processing Rate (# inferences / second) and Depedency											Example Usage Scenario Description
	HT	ES	GE	KD	SR	SS	OD	AS	DE	DR	PD	
Social Interaction A	30	60	60, ES(D)							30		AR messaging with AR object rendering
Social Interaction B		60	60, ES(D)					30				In-person interaction with AR glasses
Outdoor Activity A				3	3, KD(C)	10	30					Hiking with smart photo capture
Outdoor Activity B				3	3, KD(C)		30					Rest during hike
AR Assistant				3	3, KD(C)	10	10		30		30	Urban walk with informative AR objects
AR Gaming	45								30		30	Gaming with AR object
VR Gaming	45	60	60, ES(D)									Highly-interactive Immersive VR gaming

Please refer to our paper for details!

Benchmark Harness



- **Goal**
  - Provide a research platform for academia and industry researchers

- **Development Plan**
  - **Available Today:** DNN accelerator analytical model (MAESTRO*)-based benchmark harness
  - **Under development:** XRBench-Desktop and XRBench-Mobile
  - Please refer to our homepage for the latest info: https://xrbench.ai

How should we compare ML systems running XRBench?

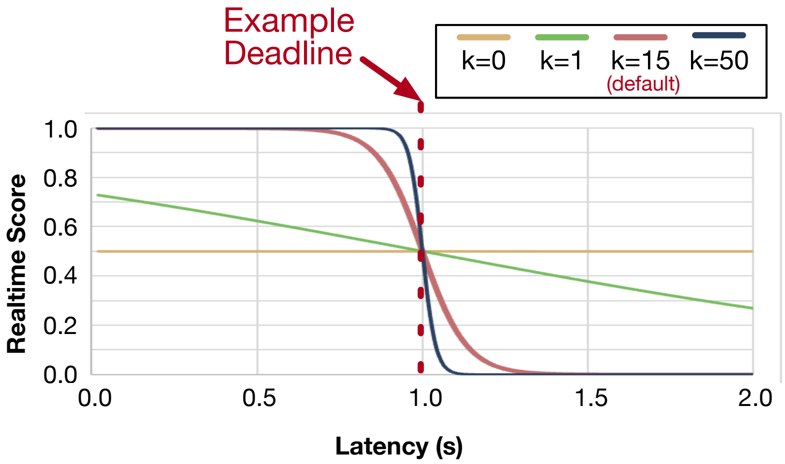
- H. Kwon et al., “Understanding Reuse, Performance, and Hardware Cost of DNN Dataflows: A Data-Centric Approach.” MICRO 2019.

Outline

- New ML Workload: Realtime **MTMM** (**M**ulti-**T**ask **M**ulti-**M**odel)
- XRBench: Realtime MTMM Benchmark Suite in XR (AR/VR)
- ➡ New Scoring Metric for Real-time MTMM
- Case Studies
- Conclusion

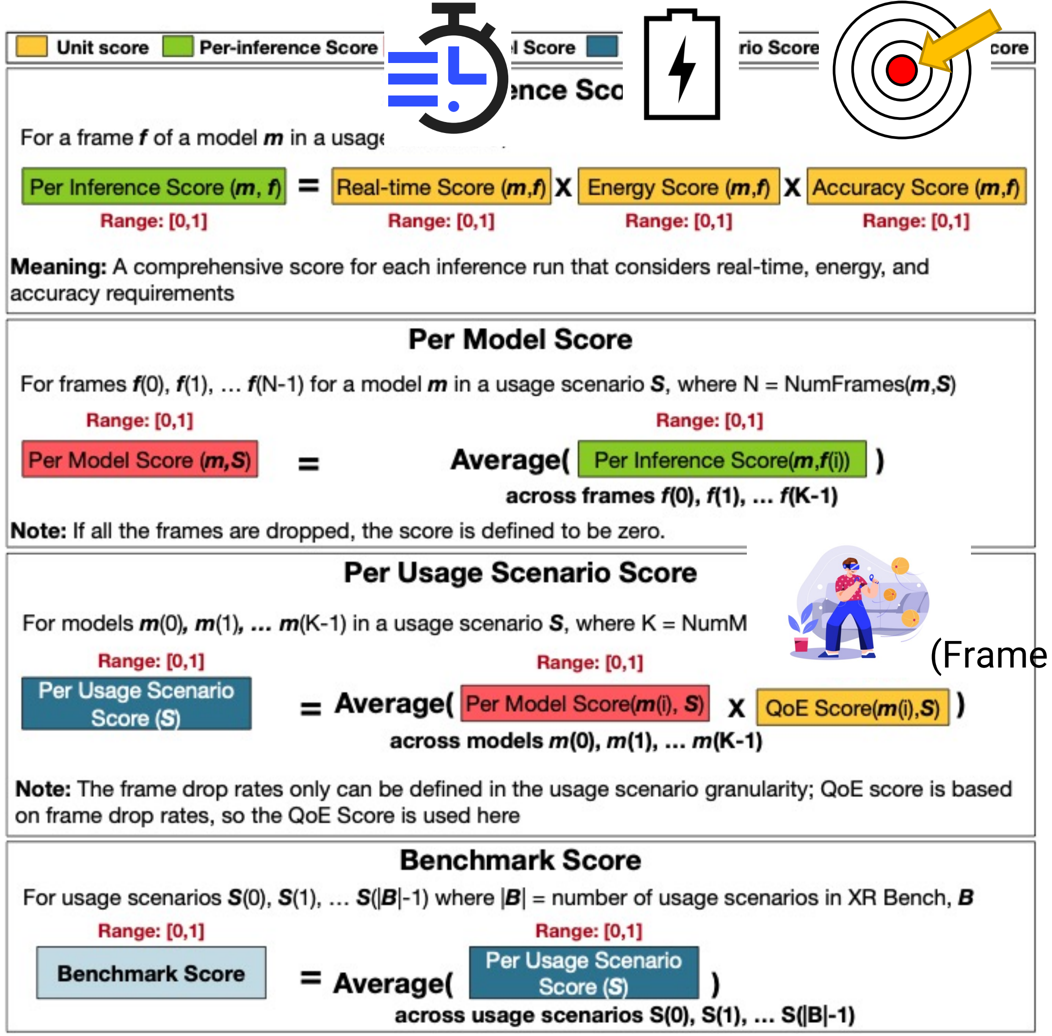
Score Metric: Unit Scores

Unit Score	What does it measure?
Real-time	Degree of deadline violations (Not absolute latency!)
Energy	Energy consumption
Accuracy	Relative model performance compared to reported numbers in original papers
Quality of Experience (QoE)	Frame drop rate



All formulated to be higher-is-better metrics in [0,1] range focusing on what matters to users

A Comprehensive Score Metric: XRBench Score



🟨 Unit score 🟩 Per-inference Score l Score 🟦 rio Score core

ence Sco

For a frame *f* of a model *m* in a usage

Per Inference Score (*m*, *f*) = Real-time Score (*m*,*f*) X Energy Score (*m*,*f*) X Accuracy Score (*m*,*f*)

Range: [0,1] Range: [0,1] Range: [0,1] Range: [0,1]

Meaning: A comprehensive score for each inference run that considers real-time, energy, and accuracy requirements

Per Model Score

For frames *f*(0), *f*(1), … *f*(N-1) for a model *m* in a usage scenario *S*, where N = NumFrames(*m*,*S*)

Range: [0,1] Range: [0,1]

Per Model Score (*m*,*S*) = **Average(** Per Inference Score(*m*,*f*(i)) **)**

across frames *f*(0), *f*(1), … *f*(K-1)

Note: If all the frames are dropped, the score is defined to be zero.

Per Usage Scenario Score

For models *m*(0), *m*(1), … *m*(K-1) in a usage scenario *S*, where K = NumM

Range: [0,1] Range: [0,1]

Per Usage Scenario Score (*S*) = **Average(** Per Model Score(*m*(i), *S*) **x** QoE Score(*m*(i),*S*) **)**

across models *m*(0), *m*(1), … *m*(K-1)

Note: The frame drop rates only can be defined in the usage scenario granularity; QoE score is based on frame drop rates, so the QoE Score is used here

Benchmark Score

For usage scenarios **S**(0), **S**(1), … **S**(|**B**|-1) where |**B**| = number of usage scenarios in XR Bench, ***B***

Range: [0,1] Range: [0,1]

Benchmark Score = **Average(** Per Usage Scenario Score (*S*) **)**

across usage scenarios S(0), S(1), … S(|B|-1)

(Frame drop rate)

Combine unit scores via product

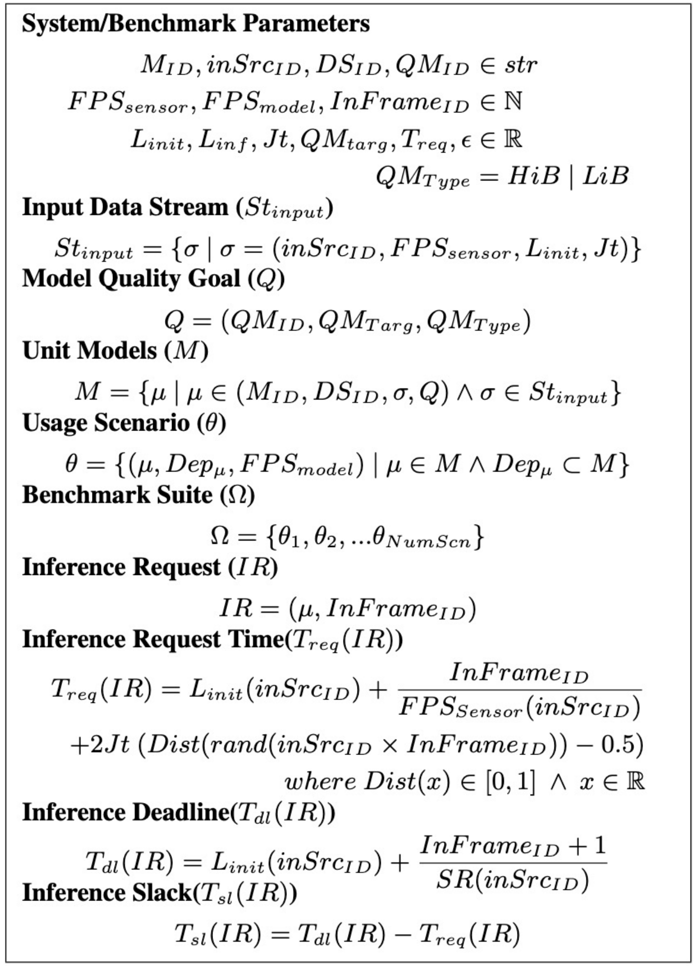
- **Hierarchical Formulation**
  - Score for each inference run -> … -> Score for the entire benchmark
- **Composable Formulation**
  - All scores in [0,1] range as higher-is-better metrics

Why is the single metric (XRBench Score) useful?

- Easier comparison across models
- Facilitate benchmark result submissions from industry

Break-down scores are reported to users
(Not mandatory to submit them)

Score Metrics: Formal Definitions



System/Benchmark Parameters

$$M_{ID}, inSrc_{ID}, DS_{ID}, QM_{ID} \in str$$

$$FPS_{sensor}, FPS_{model}, InFrame_{ID} \in \mathbb{N}$$

$$L_{init}, L_{inf}, Jt, QM_{targ}, T_{req}, \epsilon \in \mathbb{R}$$

$$QM_{Type} = HiB \mid LiB$$

Input Data Stream (St_{input})

$$St_{input} = \{\sigma \mid \sigma = (inSrc_{ID}, FPS_{sensor}, L_{init}, Jt)\}$$

Model Quality Goal (Q)

$$Q = (QM_{ID}, QM_{Targ}, QM_{Type})$$

Unit Models (M)

$$M = \{\mu \mid \mu \in (M_{ID}, DS_{ID}, \sigma, Q) \wedge \sigma \in St_{input}\}$$

Usage Scenario (θ)

$$\theta = \{(\mu, Dep_{\mu}, FPS_{model}) \mid \mu \in M \wedge Dep_{\mu} \subset M\}$$

Benchmark Suite (Ω)

$$\Omega = \{\theta_1, \theta_2, ...\theta_{NumScn}\}$$

Inference Request (IR)

$$IR = (\mu, InFrame_{ID})$$

Inference Request Time($T_{req}(IR)$)

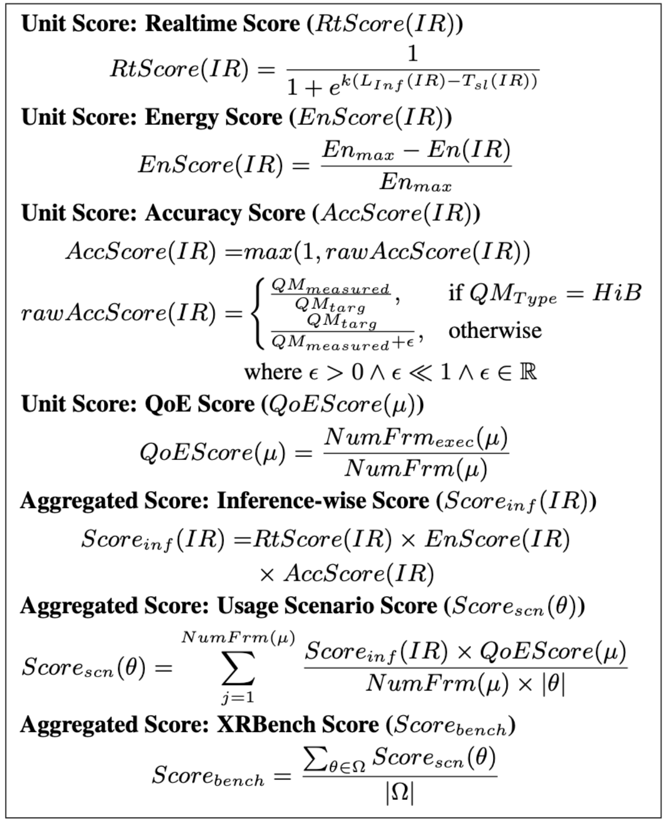
$$T_{req}(IR) = L_{init}(inSrc_{ID}) + \frac{InFrame_{ID}}{FPS_{Sensor}(inSrc_{ID})}$$
$$+2Jt\,(Dist(rand(inSrc_{ID} \times InFrame_{ID})) - 0.5)$$
$$where\ Dist(x) \in [0,1] \ \wedge \ x \in \mathbb{R}$$

Inference Deadline($T_{dl}(IR)$)

$$T_{dl}(IR) = L_{init}(inSrc_{ID}) + \frac{InFrame_{ID} + 1}{SR(inSrc_{ID})}$$

Inference Slack($T_{sl}(IR)$)

$$T_{sl}(IR) = T_{dl}(IR) - T_{req}(IR)$$



Unit Score: Realtime Score ($RtScore(IR)$)

$$RtScore(IR) = \frac{1}{1 + e^{k(L_{Inf}(IR) - T_{sl}(IR))}}$$

Unit Score: Energy Score ($EnScore(IR)$)

$$EnScore(IR) = \frac{En_{max} - En(IR)}{En_{max}}$$

Unit Score: Accuracy Score ($AccScore(IR)$)

$$AccScore(IR) = max(1, rawAccScore(IR))$$

$$rawAccScore(IR) = \begin{cases} \frac{QM_{measured}}{QM_{targ}}, & \text{if } QM_{Type} = HiB \\ \frac{QM_{targ}}{QM_{measured} + \epsilon}, & \text{otherwise} \end{cases}$$

$$\text{where } \epsilon > 0 \wedge \epsilon \ll 1 \wedge \epsilon \in \mathbb{R}$$

Unit Score: QoE Score ($QoEScore(\mu)$)

$$QoEScore(\mu) = \frac{NumFrm_{exec}(\mu)}{NumFrm(\mu)}$$

Aggregated Score: Inference-wise Score ($Score_{inf}(IR)$)

$$Score_{inf}(IR) = RtScore(IR) \times EnScore(IR)$$
$$\times\ AccScore(IR)$$

Aggregated Score: Usage Scenario Score ($Score_{scn}(\theta)$)

$$Score_{scn}(\theta) = \sum_{j=1}^{NumFrm(\mu)} \frac{Score_{inf}(IR) \times QoEScore(\mu)}{NumFrm(\mu) \times |\theta|}$$

Aggregated Score: XRBench Score ($Score_{bench}$)

$$Score_{bench} = \frac{\sum_{\theta \in \Omega} Score_{scn}(\theta)}{|\Omega|}$$

...

Please refer to our paper for details!

Paper Link: https://arxiv.org/pdf/2211.08675.pdf

Outline

- New ML Workload: Realtime **MTMM** (**M**ulti-**T**ask **M**ulti-**M**odel)
- XRBench: Realtime MTMM Benchmark Suite in XR (AR/VR)
- New Scoring Metric for Real-time MTMM
- ➡ Case Studies
- Conclusion

Case Study

- **Key Questions we answered**
  - Why new benchmark score?
  - Why different usage scenarios?
  - What are the implications to ML hardware design?

We will focus on this in this talk

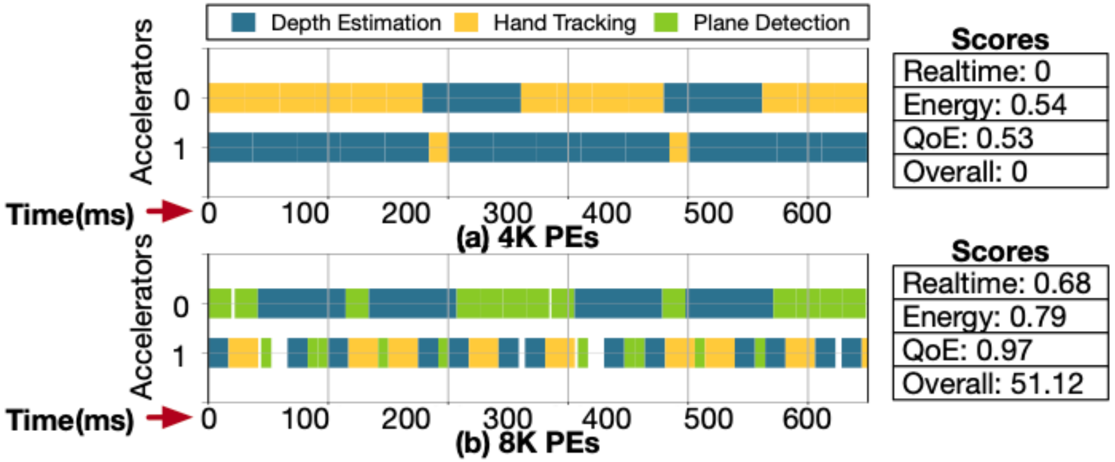
Please refer to our paper for other key insights!

Paper Link: https://arxiv.org/pdf/2211.08675.pdf

XRBench
OPEN ML BENCHMARK FOR AR/VR

Project Homepage: https://xrbench.ai

An Insight: HW Utilization as a Metric



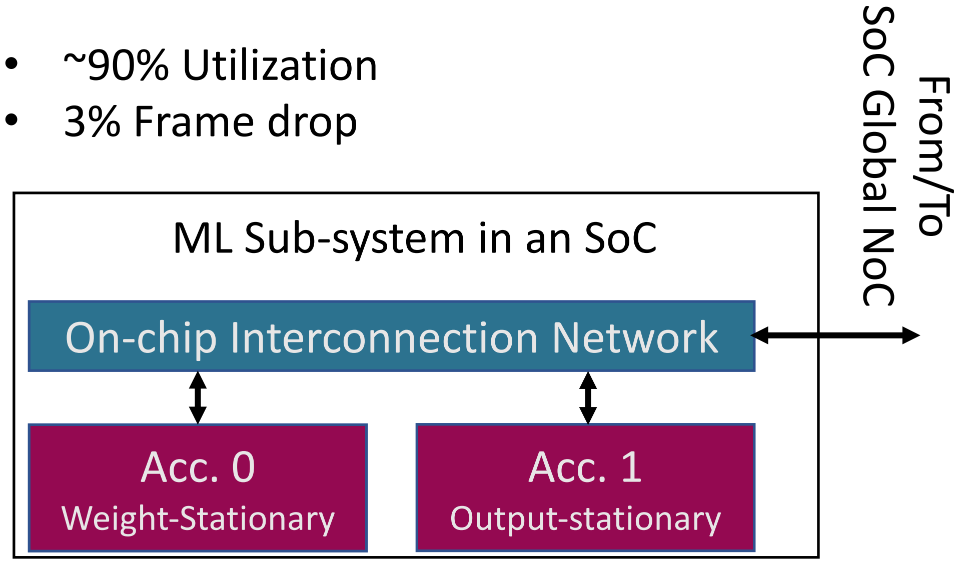
Scores
Realtime: 0
Energy: 0.54
QoE: 0.53
Overall: 0

Scores
Realtime: 0.68
Energy: 0.79
QoE: 0.97
Overall: 51.12

Figure 6. Execution timeline of AR gaming scenario on 4k and 8k PE versions of WS and OS HDA accelerator (accelerator J).

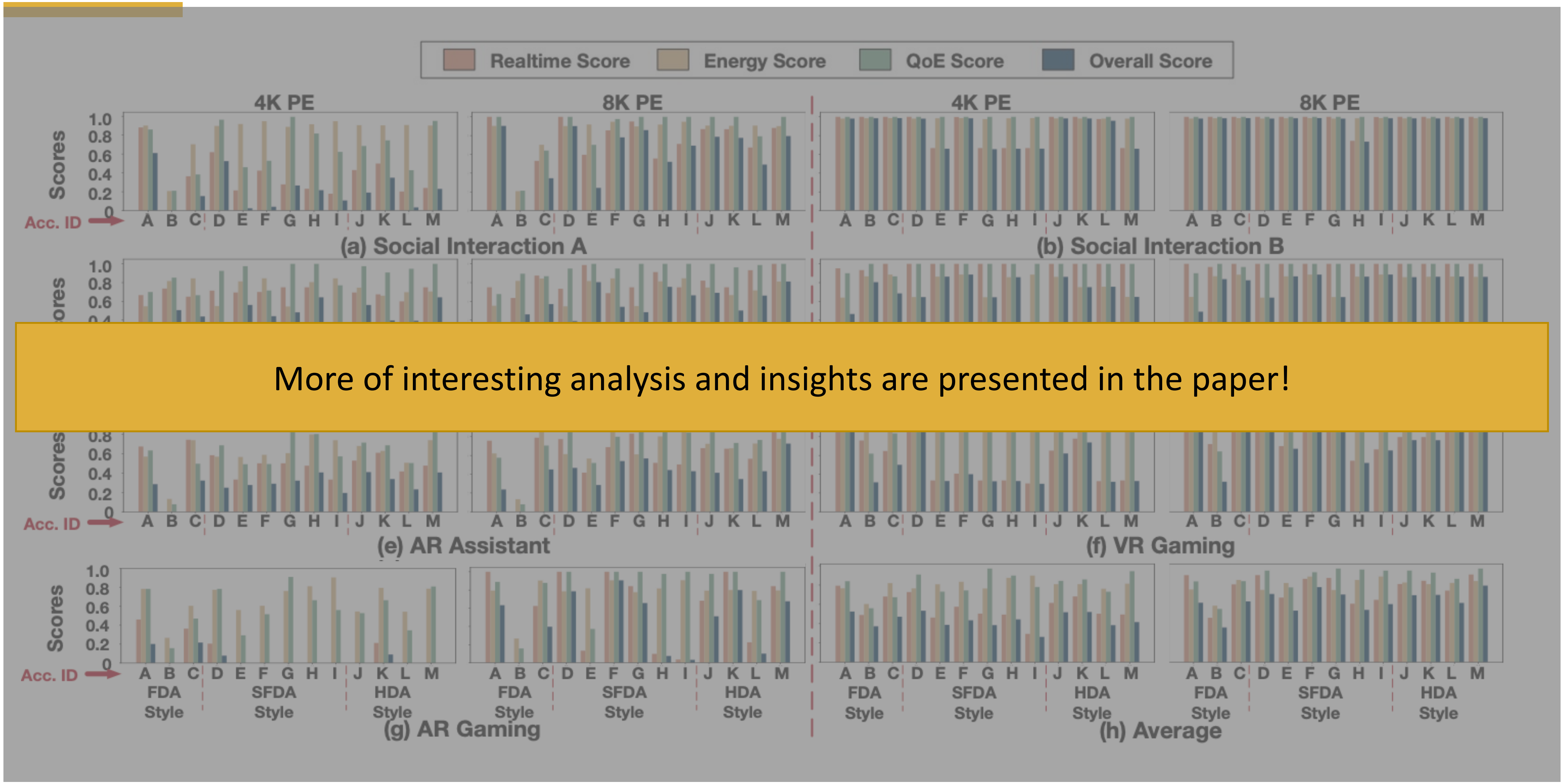
- ~100% Utilization
- 47% Frame drop

- ~90% Utilization
- 3% Frame drop



Utilization as an absolute metric is an incorrect approach for real-time MTMM ML Workloads!

Evaluation Results



More of interesting analysis and insights are presented in the paper!

- Assumes no optimizations affecting the model performance (accuracy); Fix accuracy score == 1

Outline

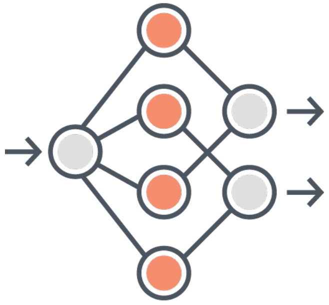
- New ML Workload: Realtime **MTMM** (**M**ulti-**T**ask **M**ulti-**M**odel)
- XRBench: Realtime MTMM Benchmark Suite in XR (AR/VR)
- New Scoring Metric for Real-time MTMM
- Case Studies
- ➡ Conclusion

Conclusion

- **Emerging Realtime MTMM ML Workloads (e.g., AR/VR)**
  - Unique characteristics leading to new challenges to ML system design, ML algorithm, etc.

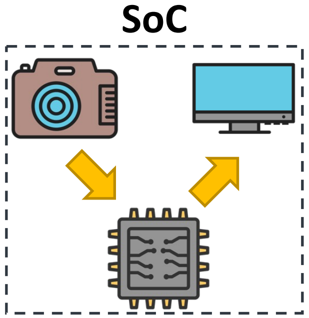
- **XRBench: An effort to publicize the research problem in MTMM ML workloads**
  - **Vision:** Keep XRBench as an open project to foster research in ML system design for real-time MTMM ML workloads

We worked to open the new research problem domain: ML System Design for RT-MTMM ML workloads
We look forward to working on this problem together!



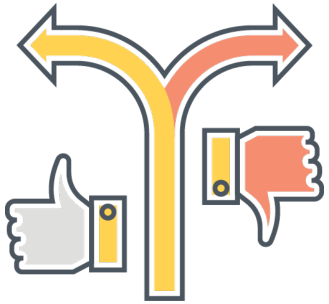
Concurrent and Cascaded Models

Real-time Processing



SoC-level Pipeline

Multi-Modal Inputs and Models



User-input-driven Dynamism

Context-driven Workloads

Acknowledgement

- **This presentation is based on the following collaboration work**
  - Hyoukjun Kwon, Krishnakumar Nair, Jamin Seo, Jason Yik, Debabrata Mohapatra, Dongyuan Zhan, Jinook Song, Peter Capak, Peizhao Zhang, Peter Vajda, Colby Banbury, Mark Mazumder, Liangzhen Lai, Ashish Sirasao, Tushar Krishna, Harshit Khaitan, Vikas Chandra, Vijay Janapa Reddi, ***"XRBench: An Extended Reality (XR) Machine Learning Benchmark Suite for the Metaverse."*** MLSys 2023 (Paper link: https://arxiv.org/pdf/2211.08675.pdf)

This project was possible thanks to everyone's contribution!

This will evolve into an open-project; we look forward to having you with us in the future!