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

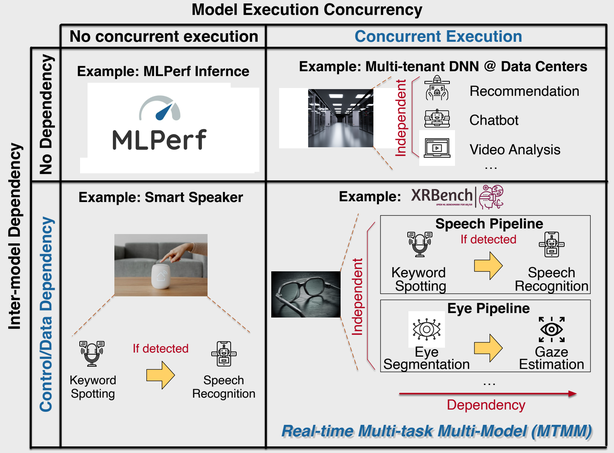
Hyoukjun kwon[1,2], Krishnakumar Nair[2], Jamin Seo[3*], Jason Yik[4*], Debabrata Mohapatra[2], Dongyuan Zhan[2], Jinook Song[2], Peter Capak[2], Peizhao Zhang[2], Peter Vajda[2], Colby Banbury[4], Mark Mazumder[4], Liangzhen Lai[2], Ashish Sirasao[2], Tushar Krishna[3], Harshit Khaitan[2], Vikas Chandra[2], Vijay Janapa Reddi[4]

EECS, University of California, Irvine[1], Meta[2], ECE, Georgia Institute of Technology[3], SEAS, Harvard University[4]

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

Real-time MTMM Workloads

ML Workload Taxonomy



Characteristics of Real-time MTMM ML Workloads

Characteristic	Example Implications to ML System Design
Concurrent and Cascaded Models	Concurrency and dependency-aware scheduling
Realtime Processing	Deadline-oriented optimizations (Just minimizing latency beyond the deadline does not contribute to better user experiences)
SoC Level Pipeline	Input sensor- and output device-aware scheduling
Multi-Modal Inputs and Models	Model heterogeneity-aware hardware design
User Input-driven Dynamism	Runtime software to support for user input-driven dynamic workloads
Context-driven Workloads	Highly diverse workload depending on user contexts

➡ To understand this new class of ML workloads, *real-time MTMM*, we need a well-defined **benchmark suite** based on realistic use cases!

XRBench v0.1

Unit Models

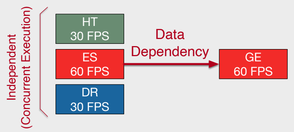
Category	Task	Model	Dataset	Model Perf. Requirement
Interaction	Hand Tracking (HT)	Hand Shape/Pose (Ge et al., 2019)	Stereo Hand Pose (Zhang et al., 2017)	AUC PCK, GT 0.948
	Eye Segmentation (ES)	RITNet (Chaudhary et al., 2019)	OpenEDS 2019 (Garbin et al., 2019)	mIoU, GT 90.54
	Gaze Estimation (GE)	Eyecod (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

; selected based on (i) ML industry recommendations (ii) model performance and efficiency

Usage Scenarios

Usage Scenario	HT	Eye Pipeline ES → GE (dep: D)		Speech Pipeline KD → SR (dep: C)		SS	OS	AS	DE	DR	PD	Example Usage Scenario Description
Social Interaction A	30	60	60							30		AR messaging with AR object rendering
Social Interaction B		60	60					30				In-person interaction with AR glasses
Outdoor Activity A				3	3	10	30					Hiking with smart photo capture
Outdoor Activity B				3	3		30					Rest during hike
AR Assistant				3	3	10	10		30		30	Urban walk with informative AR objects
AR Gaming	45								30		30	Gaming with AR object
VR Gaming	45	60	60									Highly-interactive Immersive VR gaming

Example Diagram of Usage Scenario (Social Interaction A)



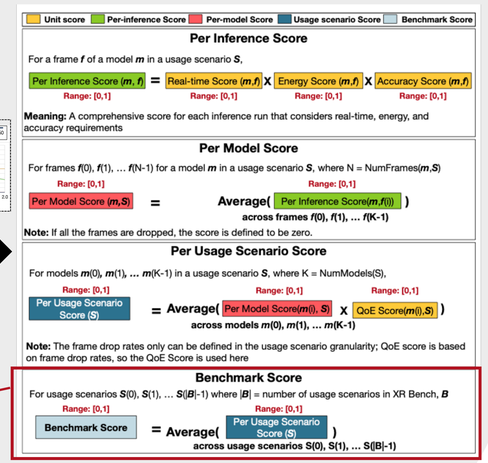
HT: hand tracking
ES: eye segmentation
GE: gaze estimation } Eye Pipeline
DR: Depth Refinement

→ Different FPS Requirements

Score Metric

4 **unit scores** (range [0,1]) are *hierarchically composed* into the **benchmark score.**

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



Single-metric
: Provides comprehensive insights and facilitates industry score submissions

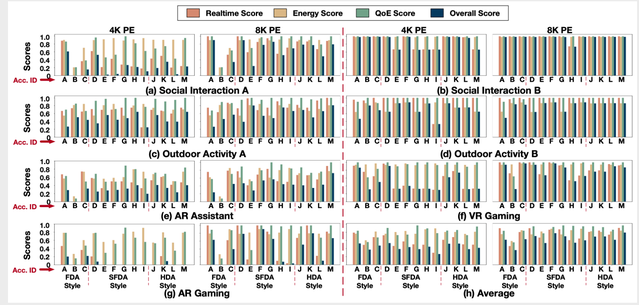
Case Studies

Main Evaluation Results

; Evaluated over 13 accelerator configurations [#PE = 4K, 8K)

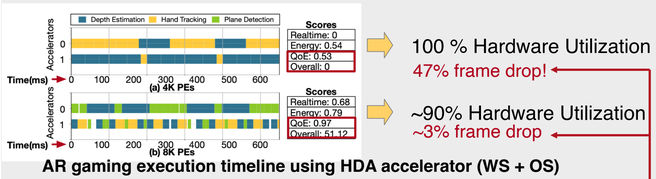
Accelerators Styles:
- *FDA (Fixed DA=Dataflow Accelerator)*
- *SFDA (Scaled-out Fixed DA)*
- *HDA (Heterogeneous DA)*

Accelerators Dataflow:
- *WS (Weight Stationary)*
- *OS (Output Stationary)*



- ML Systems for XR needs to be co-designed with usage scenarios
- Optimal accelerator style depends on the chip scale
- Multi-accelerator systems are friendly to XR systems

An Implication to ML System Design for XR



AR gaming execution timeline using HDA accelerator (WS + OS)

XR Bench score successfully captures this

Hardware utilization is an incorrect metric for XR ML system design!

XRBench
https://xrbench.ai