

The 30th Annual International Conference On Mobile Computing And Networking (MobiCom 2024)

Mobile Foundation Model as Firmware The Way Towards a Unified Mobile AI Landscape

Jinliang Yuan*, Chen Yang*, Dongqi Cai*

Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia

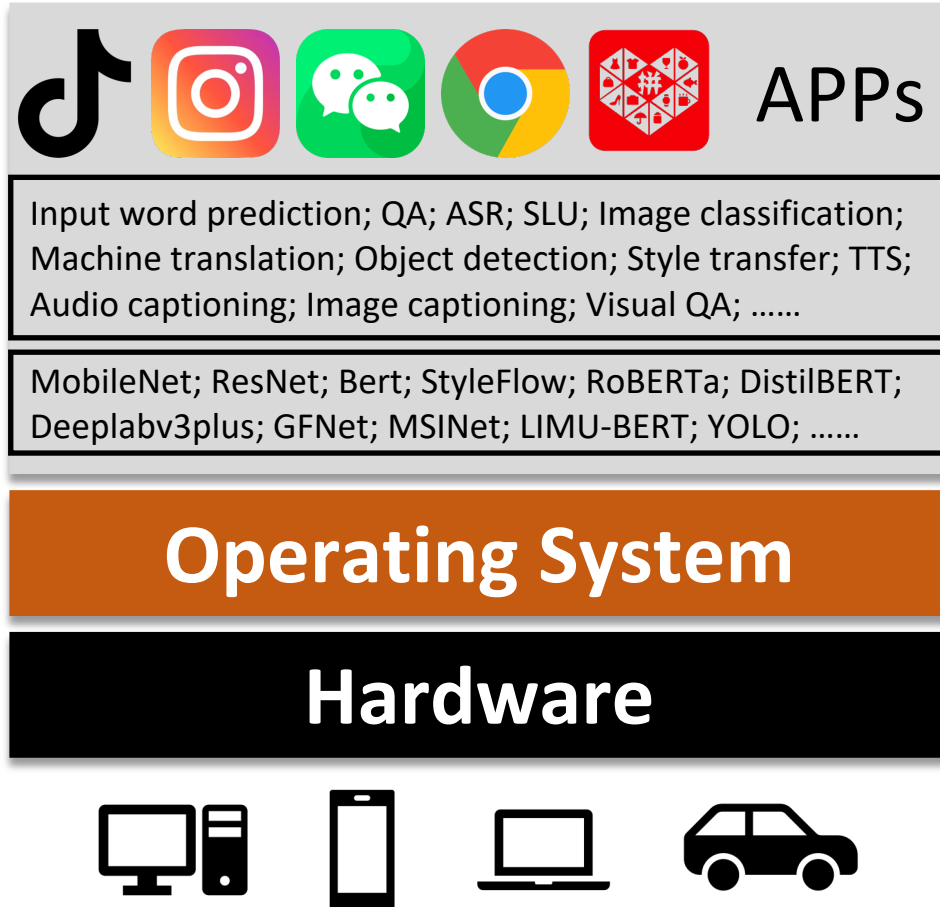
Shangguang Wang, Mengwei Xu

Beijing University of Posts and Telecommunications

Presenter: Hao Wen (Tsinghua University)



Motivation: Mobile AI Fragmentation



Diversified on-device DNN types, parameters, and configurations

Motivation: Mobile AI Fragmentation

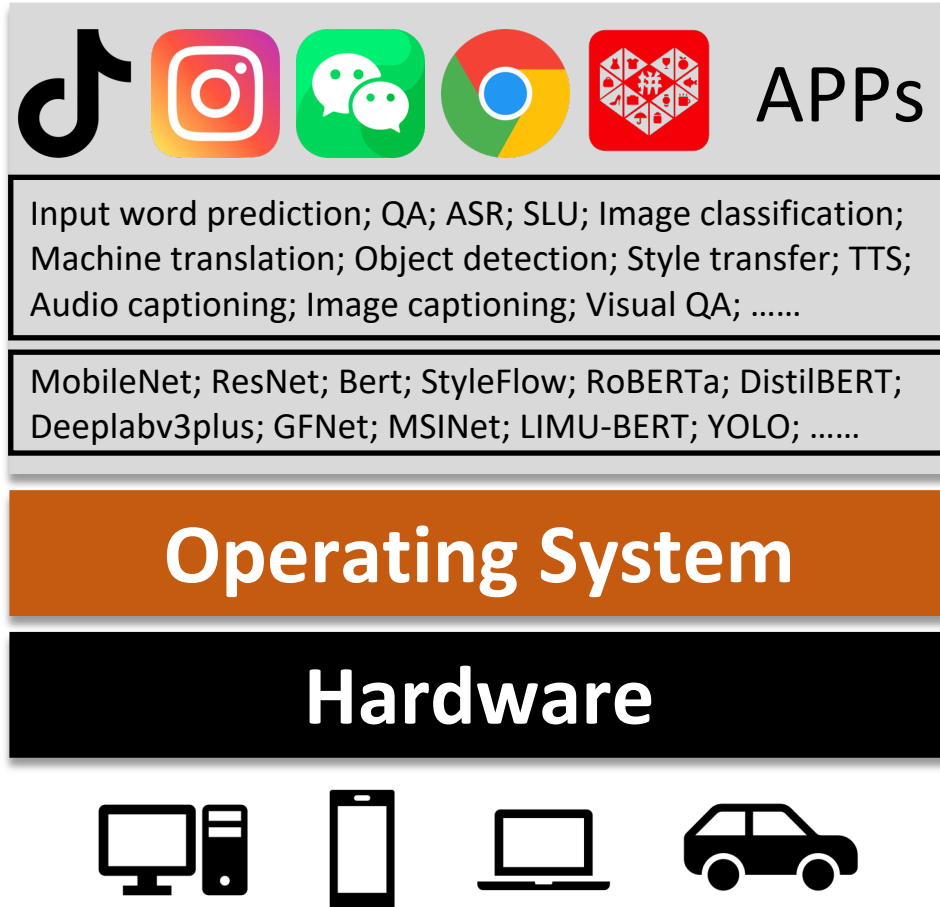


Diversified on-device DNN types, parameters, and configurations

1) H/W accelerator design trades off efficiency for generality



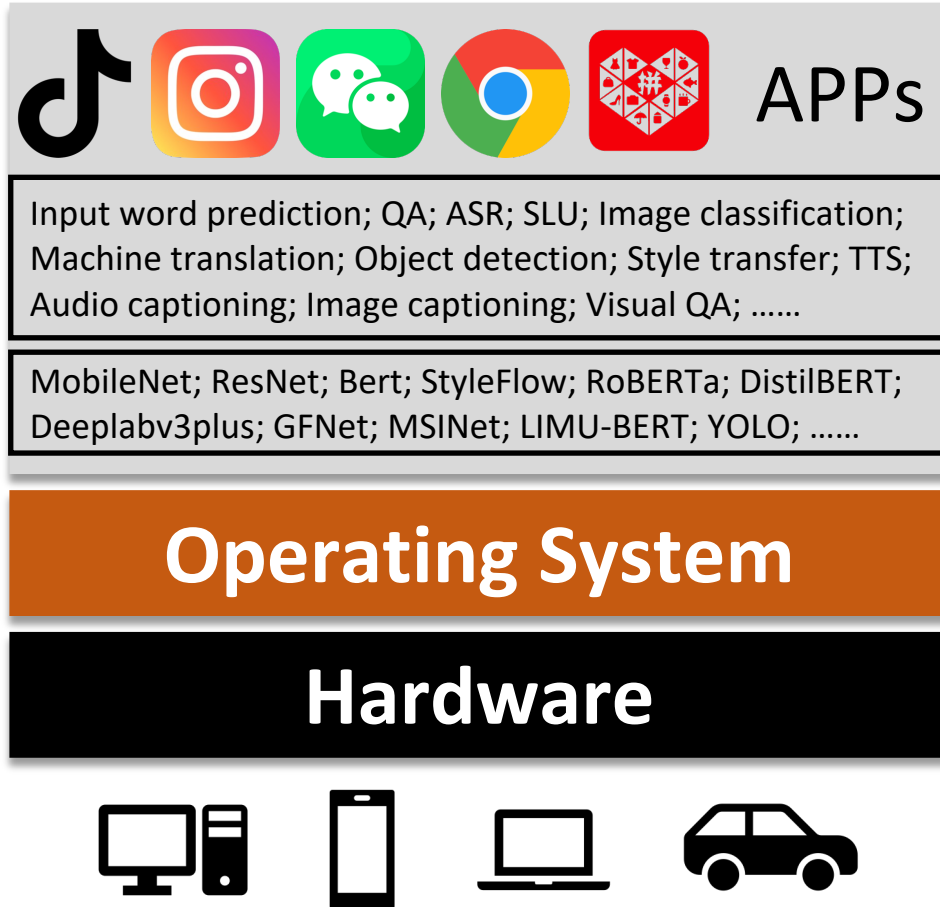
Motivation: Mobile AI Fragmentation



Diversified on-device DNN types, parameters, and configurations

- 1) H/W accelerator design trades off efficiency for generality
- 2) S/W optimizations becomes ad-hoc

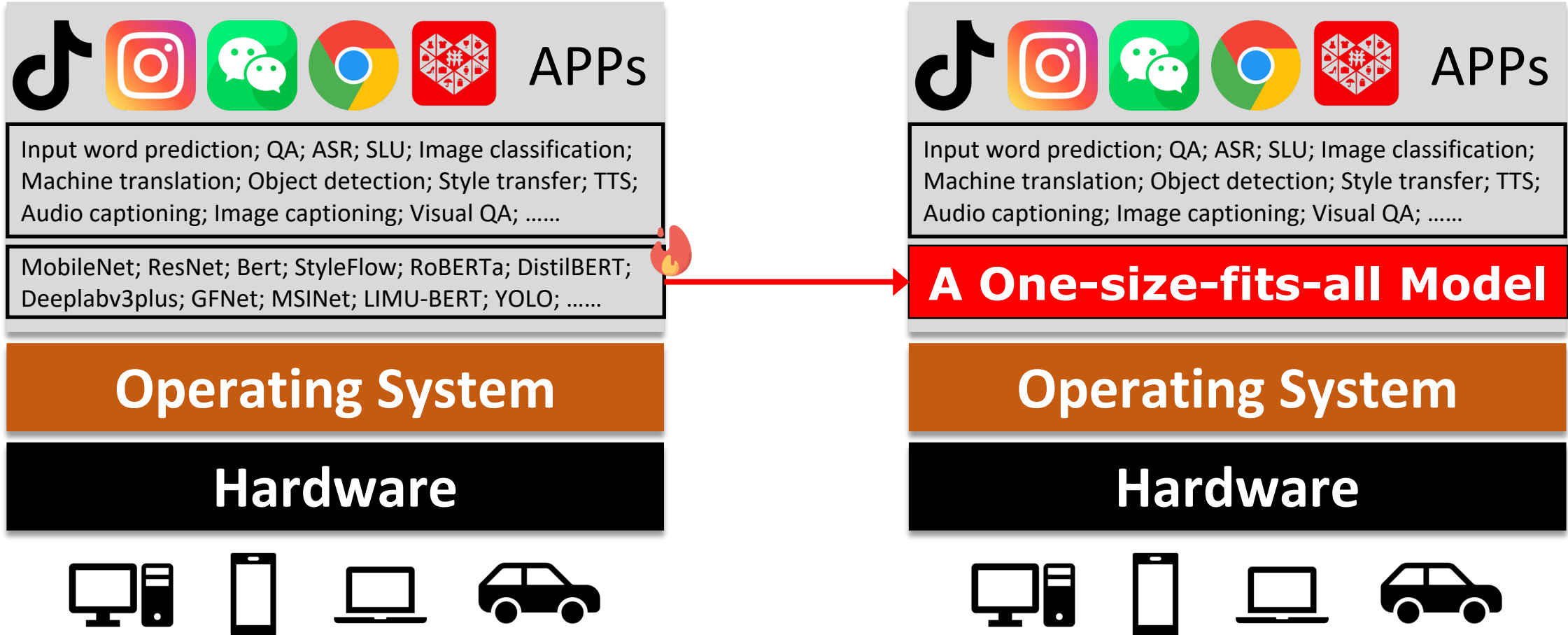
Motivation: Mobile AI Fragmentation



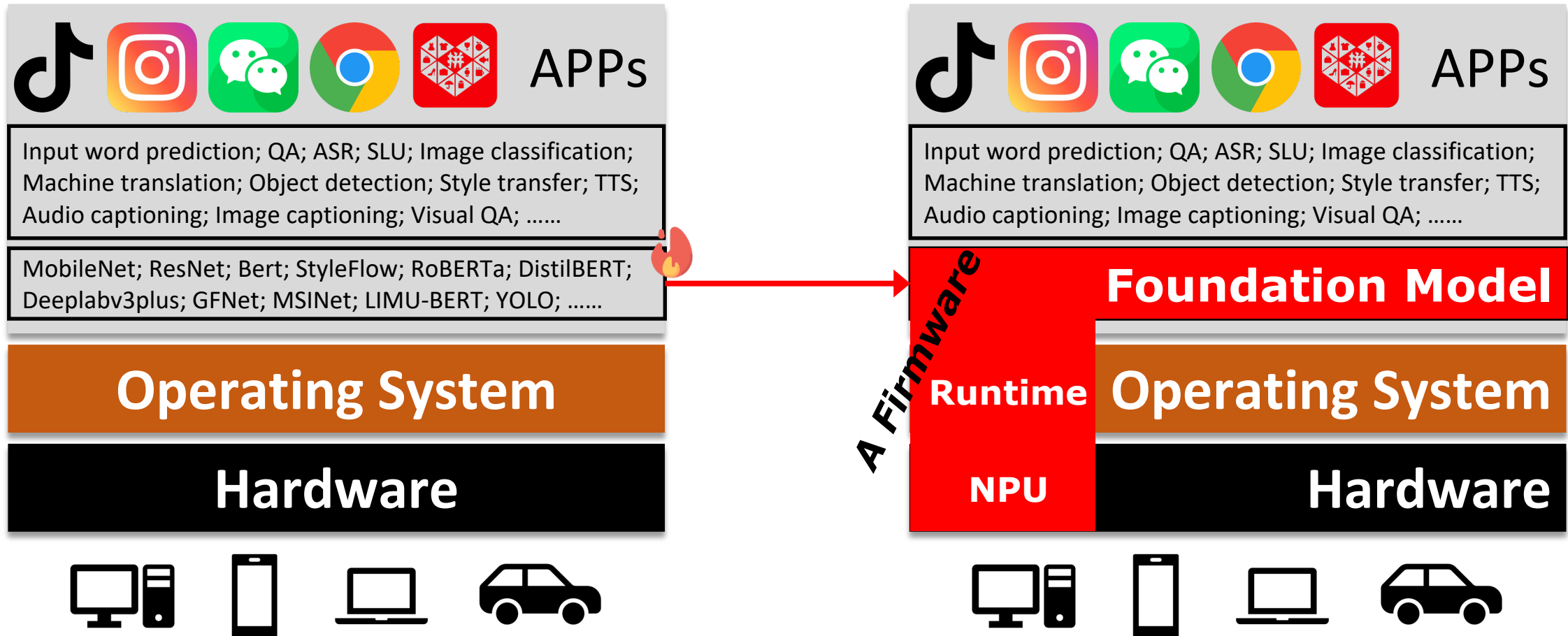
Diversified on-device DNN types, parameters, and configurations

- 1) H/W accelerator design trades off efficiency for generality
- 2) S/W optimizations becomes ad-hoc
- 3) OS-agnostic, so no opportunity for caching/scheduling/batching

A Vision: Mobile Foundation Model



A Vision: Mobile Foundation Model



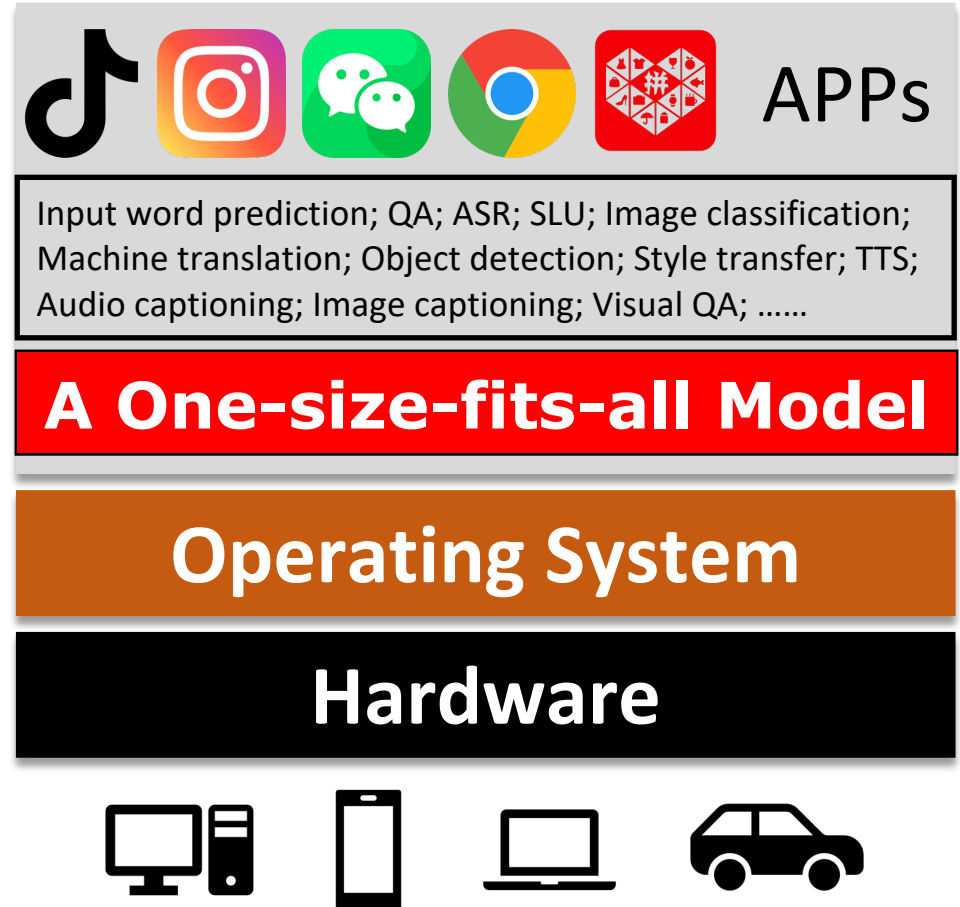
Our Contributions

a. How to **design** such a model?

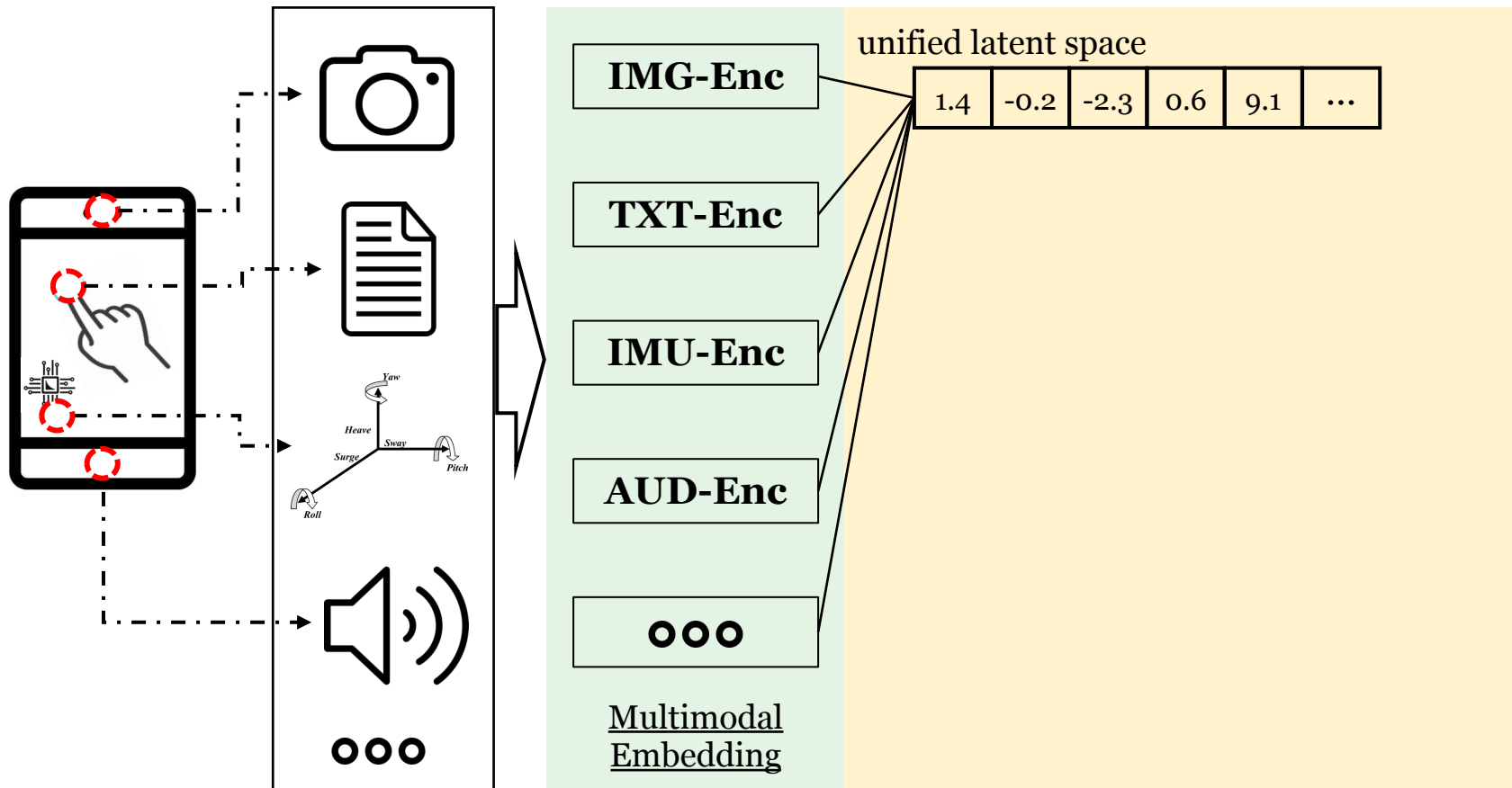
M4 – an any-to-any modality mobile foundation model

b. How to **evaluate** such a model?

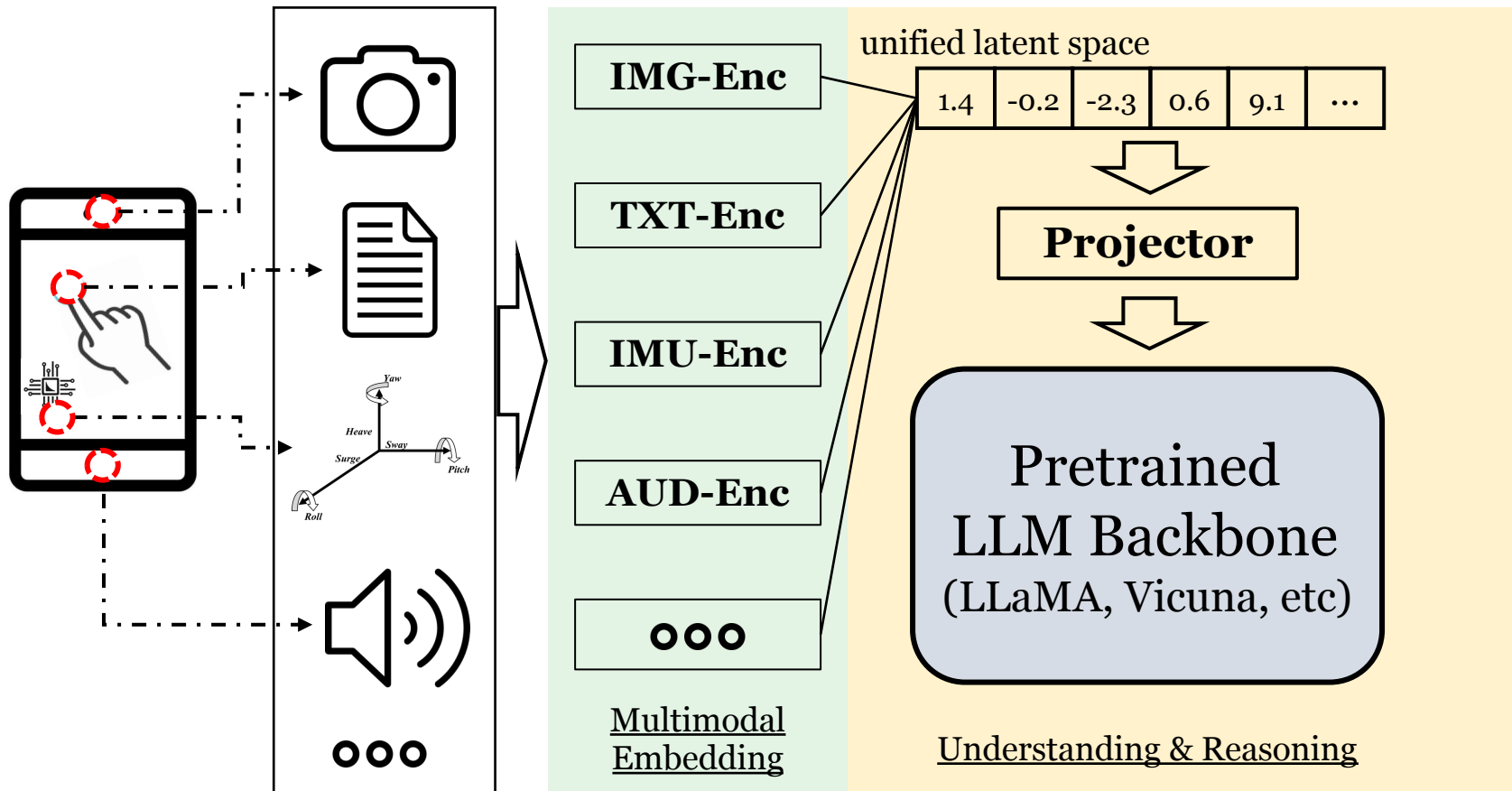
eAIBench – a comprehensive mobile AI benchmark



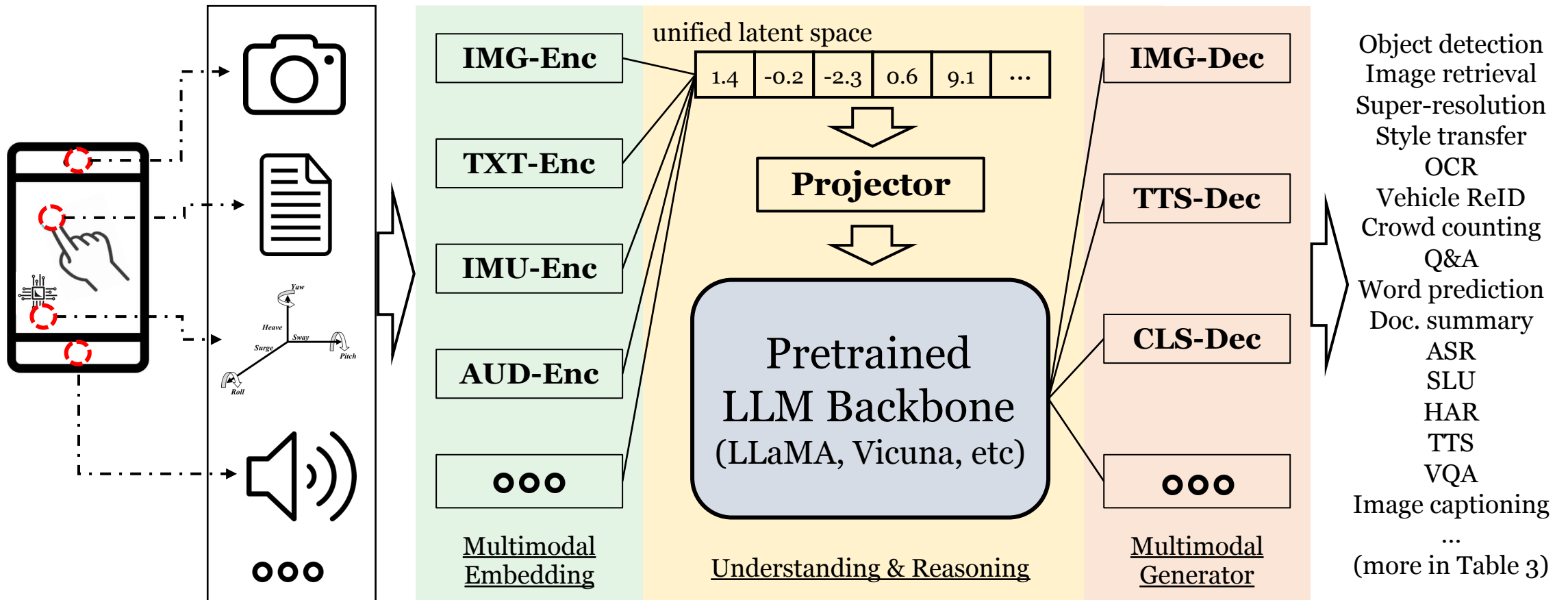
M4 Design



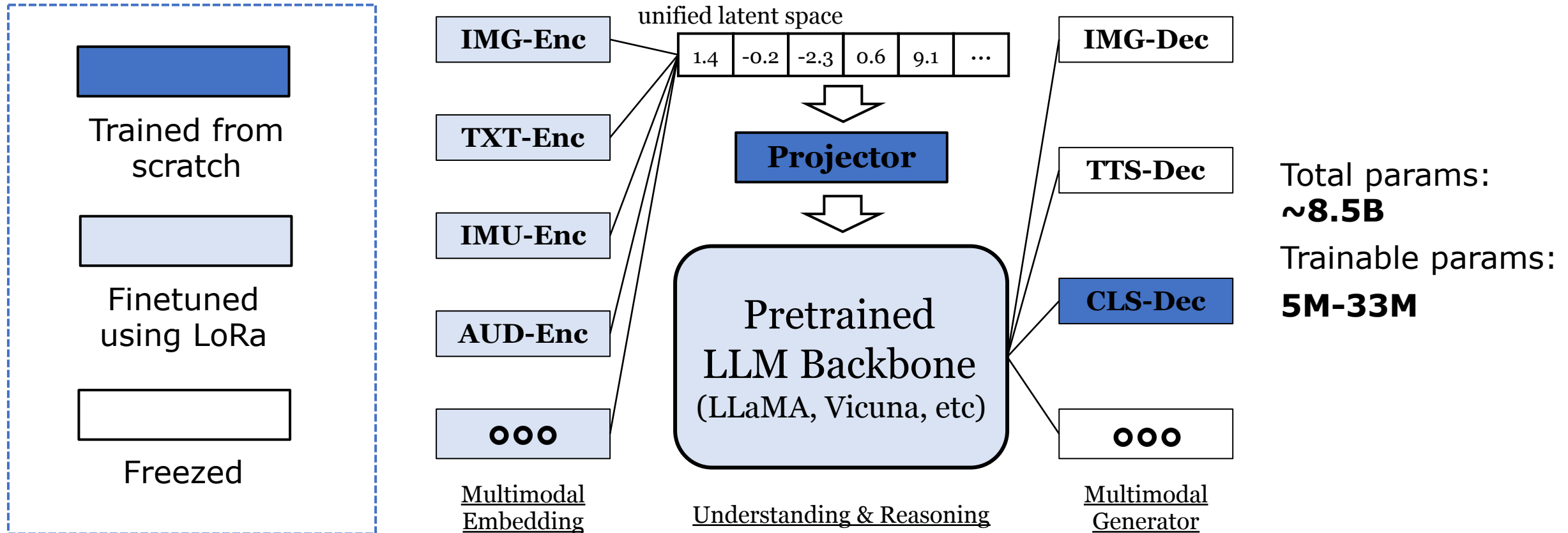
M4 Design



M4 Design



M4 Implementation



Evaluation: eAIBench

Tasks	Mobile Application	Dataset	Task-specific Models		Accuracy
			Name	Size (M)	
Input word prediction ^{T1}	Input method (Gboard)	PTB	RNN	1.4	Acc: 0.17

eAIBench: a collection of real-world mobile AI tasks and a representative baseline DNN

- **38** mobile AI tasks
- **50** classical ML datasets
- **4** modalities (text, image, audio, IMU)
- **Many hardware setups:** Octa-core CPU, Mali-G710 MP7 GPU, and edge TPU, etc.

Category	Tasks	Mobile Application	Dataset	Task-specific Models		Accuracy
				Name	Size (M)	
NLP	Input word prediction ^{T1}	Input method (Gboard)	PTB	RNN	1.4	Acc: 0.17
	Question answering ^{T2}	Intelligent personal assistant (Siri)	SQuAD v2.0	RoBERTa	37	F1: 78.60
	Machine translation ^{T3}	Translator (Google Translate)	TyDi QA	ArnoLECTRA	136	F1: 86.62
	Emoji prediction ^{T4}	Input method (Gboard)	tweet_eval	RoBERTa	125	Acc: 0.33
	Emotion prediction ^{T5}	Conversational analytics (Clarabridge)	go_emotion	RoBERTa	125	Acc: 0.47
	Sentiment analysis ^{T6}	Conversational analytics (Clarabridge)	tweet_eval	RoBERTa	125	Acc: 0.77
	Text classification ^{T7}	Spam SMS filtering (Truecaller)	ag_news	BERT	110	Acc: 0.77
	Grammatical error correction ^{T8}	Writing assistance (Grammarly)	SST2	DistilBERT	67	Acc: 0.51
	Text summary ^{T9}	Reading assistant (ChatPDF)	JFLEG	FLAN-t5	801	BLEU: 0.68
	Code document generation ^{T10}	Code editor (Javado)	CNN Daily Mail	BAIT	400	BLEU: 0.43
CV	Code generation ^{T11}	Code editor (Copilot)	Shellcode_IA32	CodeBERT	125	BLEU: 91.7
	Object detection ^{T12}	Augmented Reality (Google Lens)	COCO	Libra-rnn	42	AP: 0.43
	Image retrieval ^{T13}	Image searcher (Google Photos)	LVIS	X-Paste	952	AP: 0.51
	Super-resolution ^{T14}	Video/Image super-resolution (VSCO)	Clothes Retrieval	Resnet50-arcface	31.7	Recall: 0.90
	Styler transfer ^{T15}	Painting & Beautifying (Meitu)	REDS	Real-ESRGAN	16.7	SSIM: 0.83
	Semantic segmentation ^{T16}	Smart camera (Segmentix)	COCO Wikilart	StyleFlow	16.8	SSIM: 0.45
	Optical character recognition ^{T17}	Intelligent document automation software (Oculus)	ADE20K-150	DeepLabv3plus	40	mIoU: 0.43
	Image classification ^{T18}	Album management (Google Photos)	PASCAL VOC 2012	DeepLabv3plus	40	mIoU: 0.90
	Traffic sign classification ^{T19}	Intelligent transportation (Waze)	Rendered SST2	CLIP	438	Acc: 0.71
	Vehicle re-identification ^{T20}	Surveillance camera (AI Re-ID)	CIFAR100	GFNet	54	Acc: 0.89
Audio	Gender recognition ^{T21}	Smart camera (Face++)	GTSRB	MicronNet	0.43	Acc: 0.98
	Location recognition ^{T22}	Navigation search (Google Maps)	Veri776	MSINet	2.3	Rank: 0.96
	Pose estimation ^{T23}	AI fitness coach (Keep)	Adience	MiVOLO-D1	27.4	Acc: 0.96
	Video classification ^{T24}	Video player (YouTube)	Country211	CLIP	438	Acc: 0.46
	Counting ^{T25}	Smart camera (Fitness Tracking)	AP-10K	VITPose	44	Acc: 0.69
	Image matting ^{T26}	Virtual backgrounds (Zoom)	Kinetics400	SlowFast	66	Acc: 0.89
	Automatic speech recognition ^{T27}	Private assistant (Siri)	Smart camera (Fitness Tracking)	Crowd Counting	37	MAE: 437
	Spoken language understanding ^{T28}	Private assistant (Siri)	RefMatte-RW100	MDETR	170	MSE: 0.06
	Emotion recognition ^{T29}	Emoji reco	LibriSpeech	CTC+attention	120	WER: 3.16%
	Audio classification ^{T30}	Music disc	FSC	Transformer	39	WER: 0.37%
Multimodal	Keyword spotting ^{T31}	Private ass	SLRP	CRDNN	26.5	Acc: 0.82
	Human activity recognition ^{T32}	AI fitness				
	Text-to-speech ^{T33}	Voice broa				
	Audio captioning ^{T34}	Hearing-im				
	Image captioning ^{T35}	Visual-imp				
	Text-to-image retrieval ^{T36}	Image search (Google Lens)	Flickr30k	CCLM	500	Recall: 0.69
	Audio/Text-to-image generation ^{T37}	Art creation (Verb Art)	VGGSound	Wav2clip	466	FID: 99.89
	Visual question answering ^{T38}	Visual-impaired accessibility (Answerables)	VQA v2.0	MUTAN	218	Acc: 0.63
			VizWiz	MUTAN	218	Acc: 0.52

eAIBench

Evaluation: End-to-end Performance

- **M4 can well support most mobile AI tasks and datasets.**

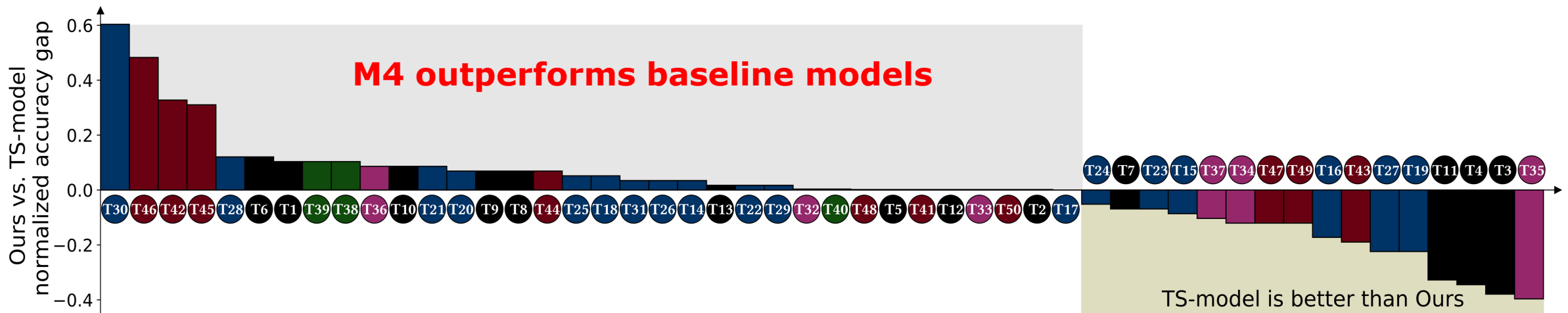
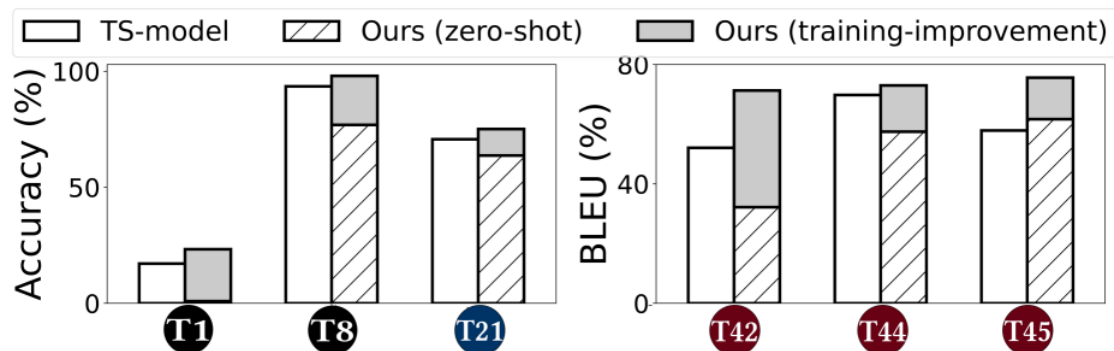


Figure 1. Normalized accuracy comparison of M4 and TS-models on 50 popular mobile tasks and datasets.

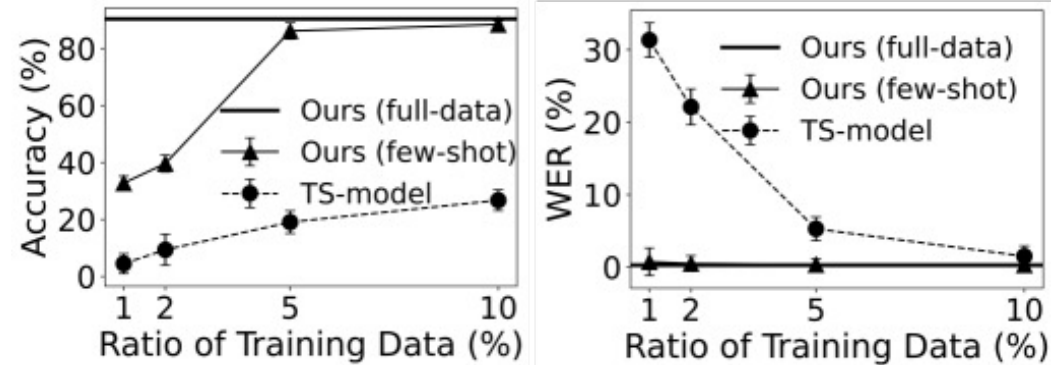
M4 can achieve comparable performance across **85%** of tasks, with over **50%** of these tasks showcasing considerable performance improvement.

Evaluation: Zero/Few-shot Ability

- **M4 also has a certain zero-shot ability, but fine-tuning makes it much more accurate.**

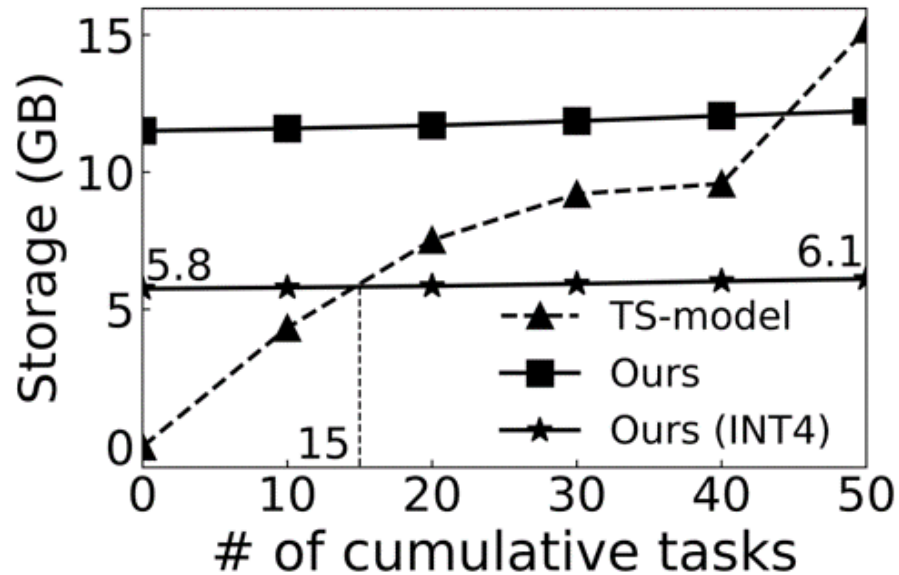


- **M4 has better few-shot ability than TS-models that are trained from scratch.**

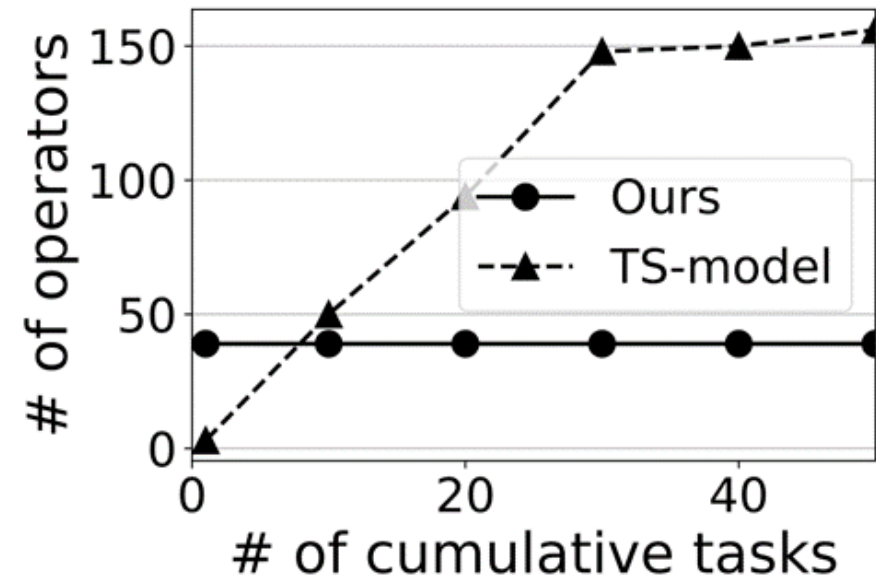


Evaluation: Runtime Cost

- **M4 is more storage-efficient when the model number scales out.**

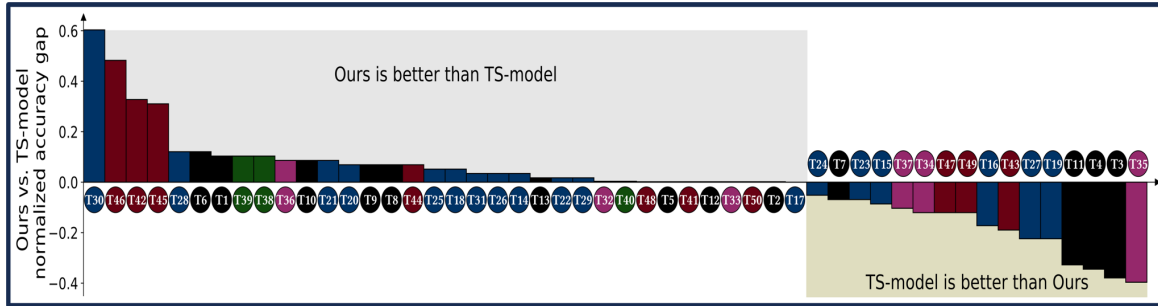


- **M4 greatly simplify accelerator design.**



M4 evolves with better backbones

Beat over **50%** of TS-model



Default Backbone

LLaMA 1 ∞
●
2023.02

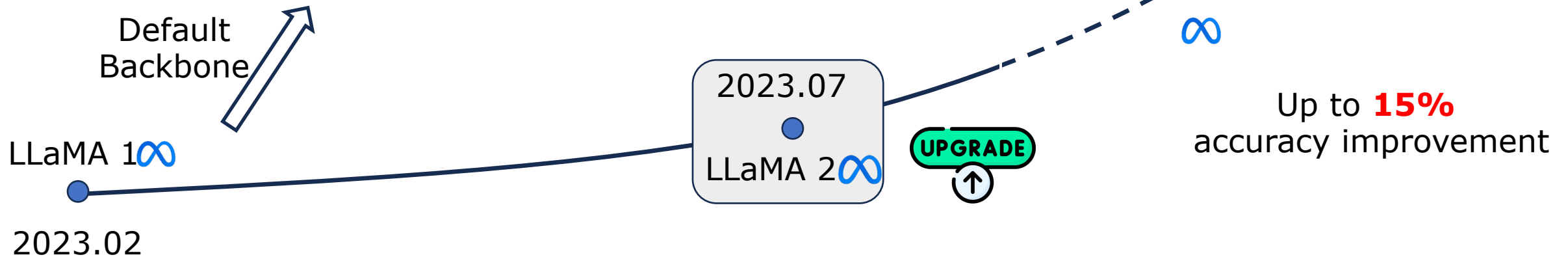
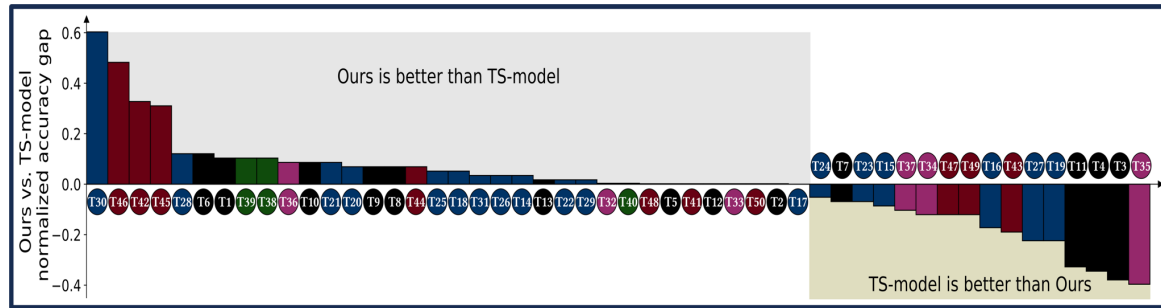
2023.07
LLaMA 2 ∞



M4 can be further enhanced with enhanced foundation models.

M4 evolves with better backbones

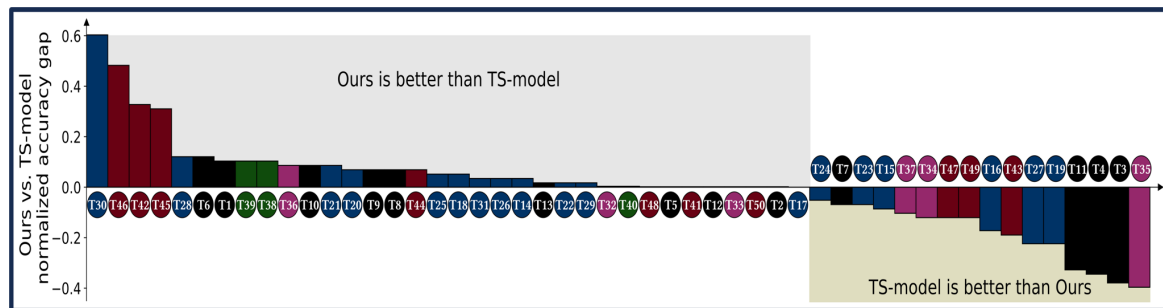
Beat over **50%** of TS-model



M4 can be further enhanced with enhanced foundation models.

M4 evolves with better backbones

Beat over **50%** of TS-model



Default Backbone
LLaMA 1 ∞
2023.02

2023.07
LLaMA 2 ∞

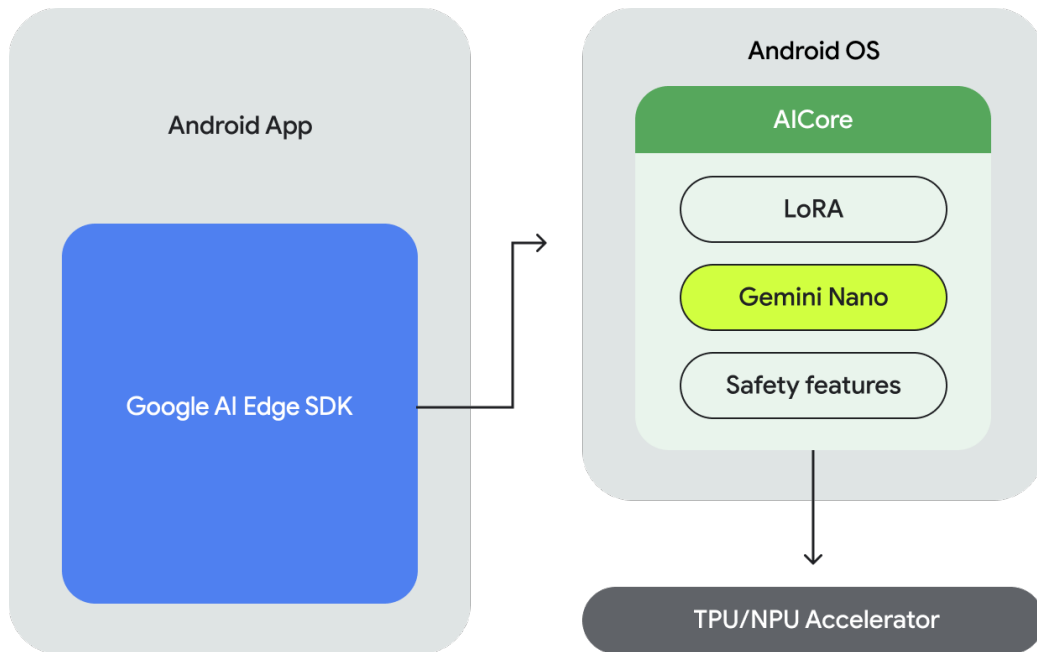
UPGRADE
↑

Up to **15%** accuracy improvement
More LLM backbones now available!

M4 can be further enhanced with enhanced foundation models.

No longer a vision now!

- Google has turned our vision into reality on Android!



- **Gemini Nano as a standalone system service**
- **Apps access the LLM through prompts or LoRa**
- **On-device LLM directly accelerated by TPU/NPU**

[1] <https://developer.android.com/ai/gemini-nano>

Mobile Foundation Model as Firmware



Jinliang Yuan*, Chen Yang*, Dongqi Cai*, ..., Shanguang Wang, Mengwei Xu

Contact: mwx@bupt.edu.cn

Summary of our contribution

- A vision of having one foundation model to replace all fragmented DNNs on mobile devices.
- A design and prototype of such a foundation model.
- A comprehensive benchmark that demonstrates its feasibility.
- **Takeaway: time to revolutionize mobile AI landscape!**

Code: <https://github.com/UbiquitousLearning/MobileFM>