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* (*Co-first)

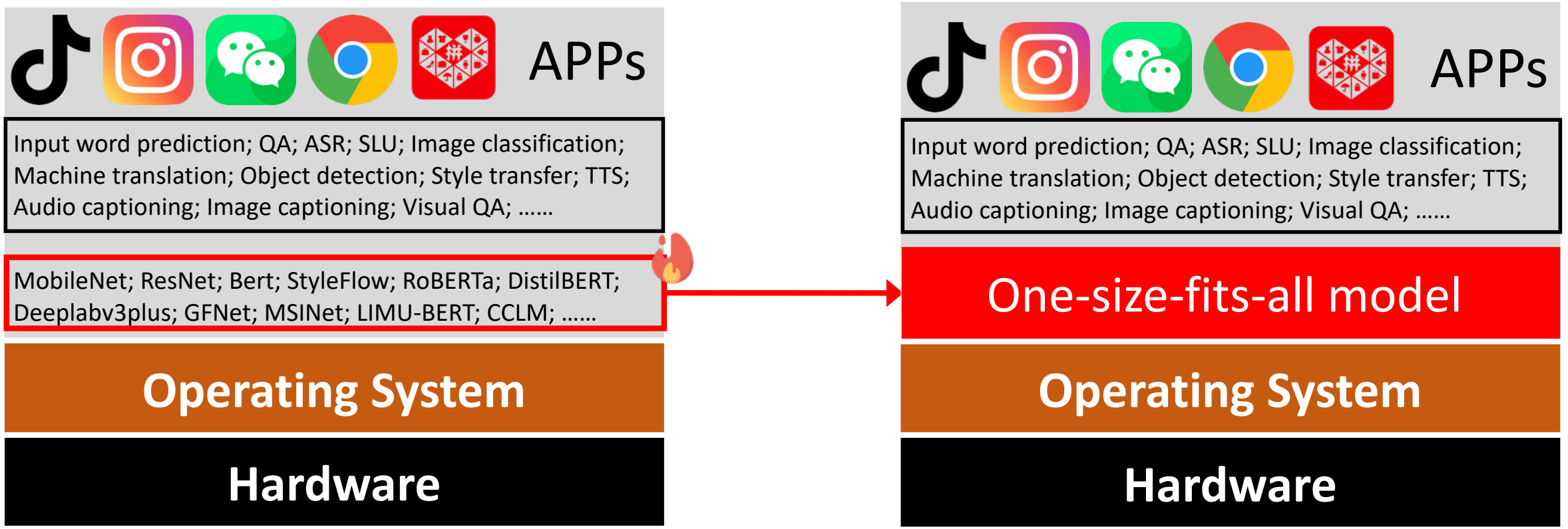
Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia

Shangguang Wang, Mengwei Xu

Beijing University of Posts and Telecommunications

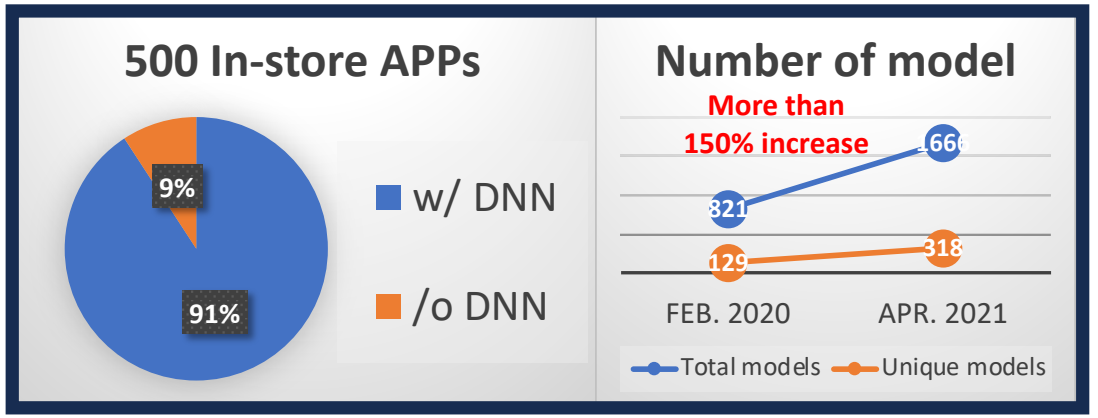
Presenter: Hao Wen (Tsinghua University)

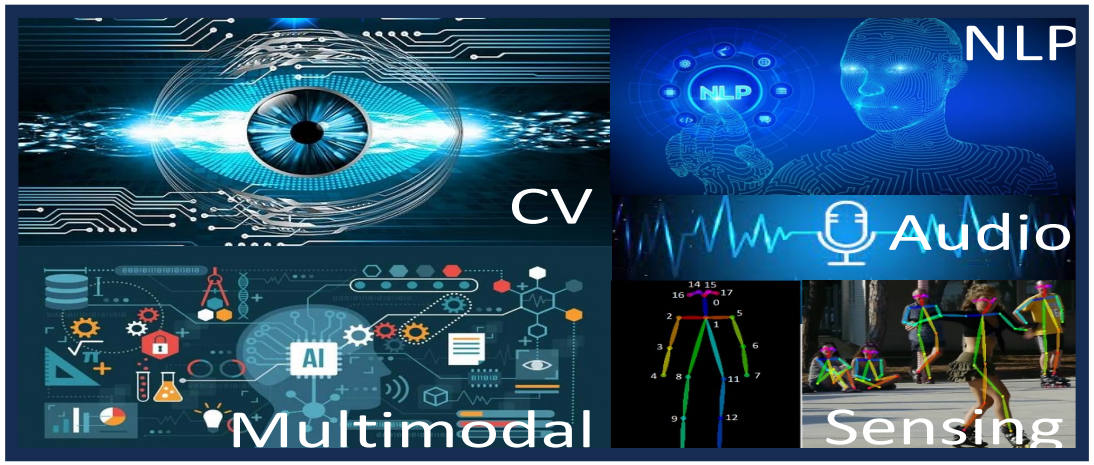
Mobile Foundation Model

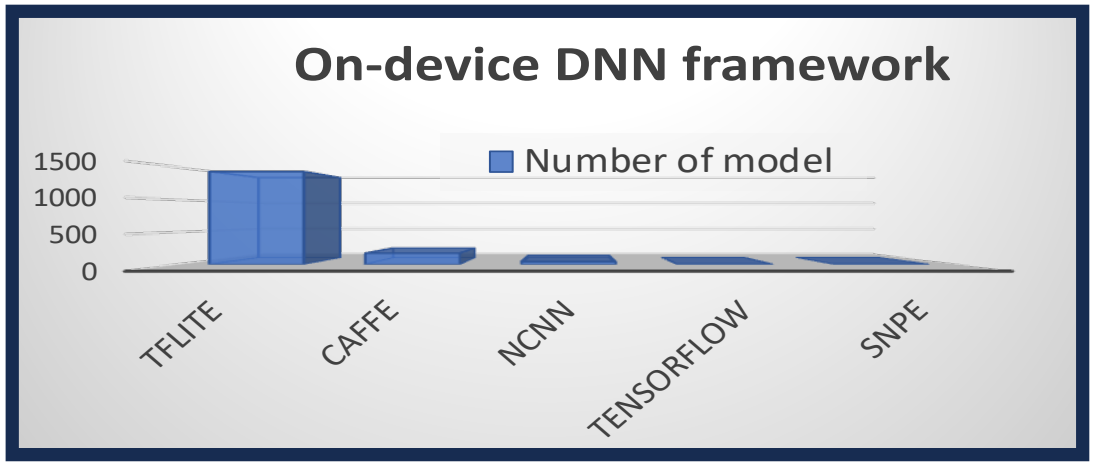


Challenge #1: Fragmented Mobile DNNs

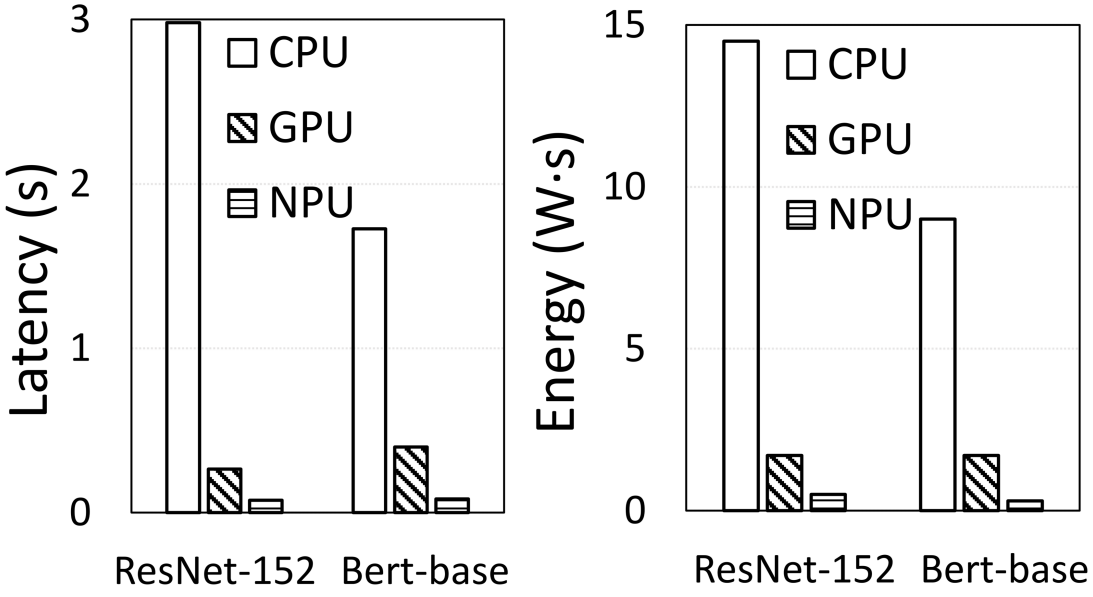
Mobile devices	Tasks	Task-specific models
	Question answering	RoBERTa
	Object detection	Libra-rcnn
	Keyword spotting	Cnn-trad-fpool3
	Text-to-speech	Transformer



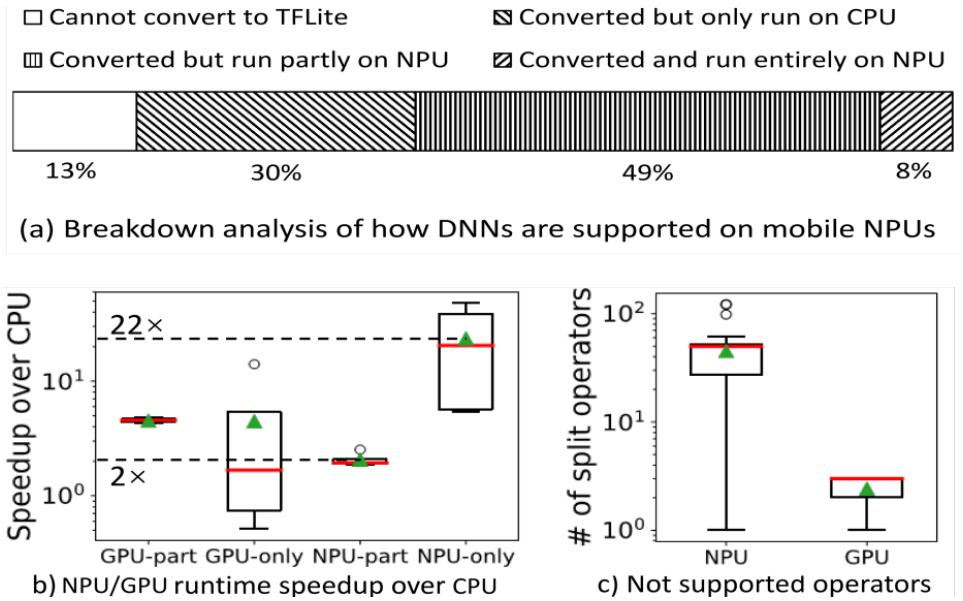




Challenge #2: A Dilemma of Mobile NPU

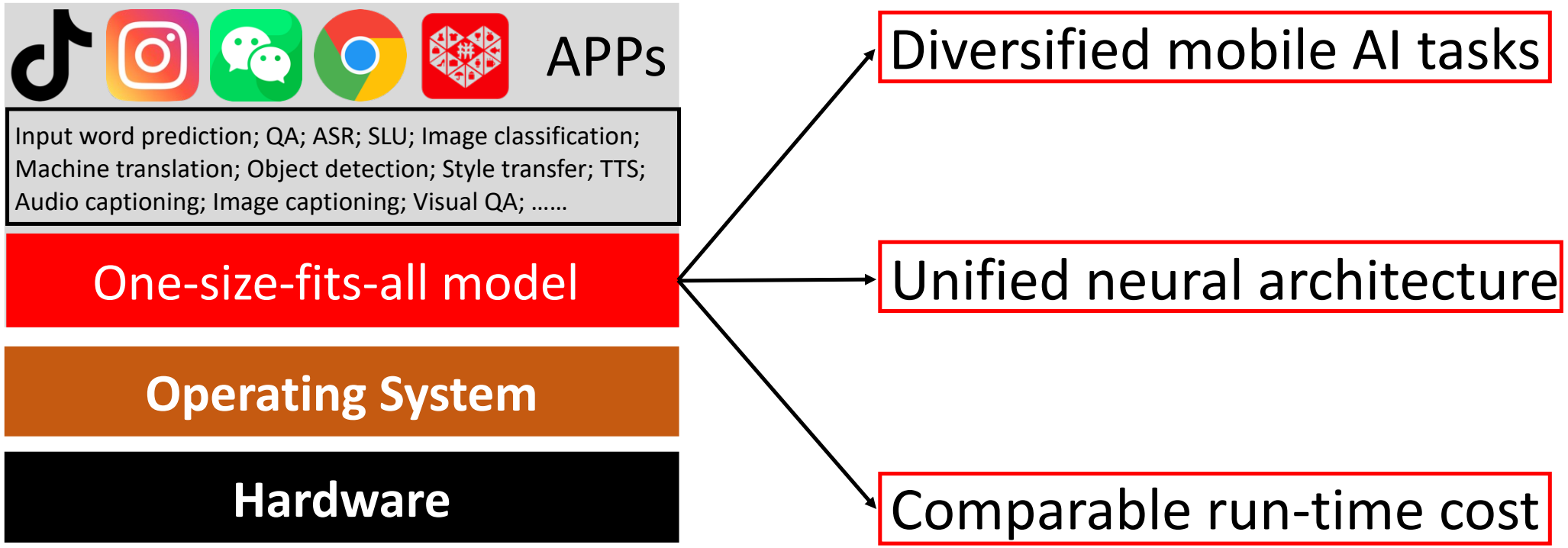


NPU highly improve the latency and energy compared with CPU and GPU.

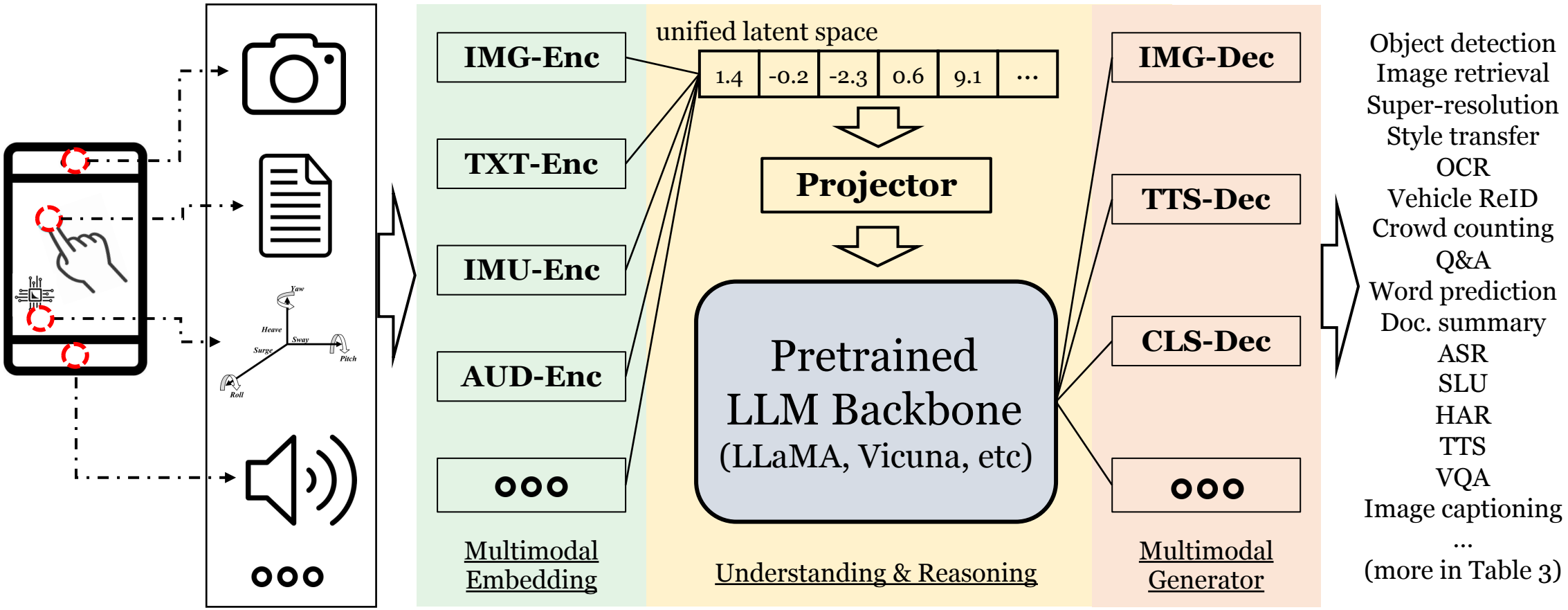


The supported DNN operators are serious limited.

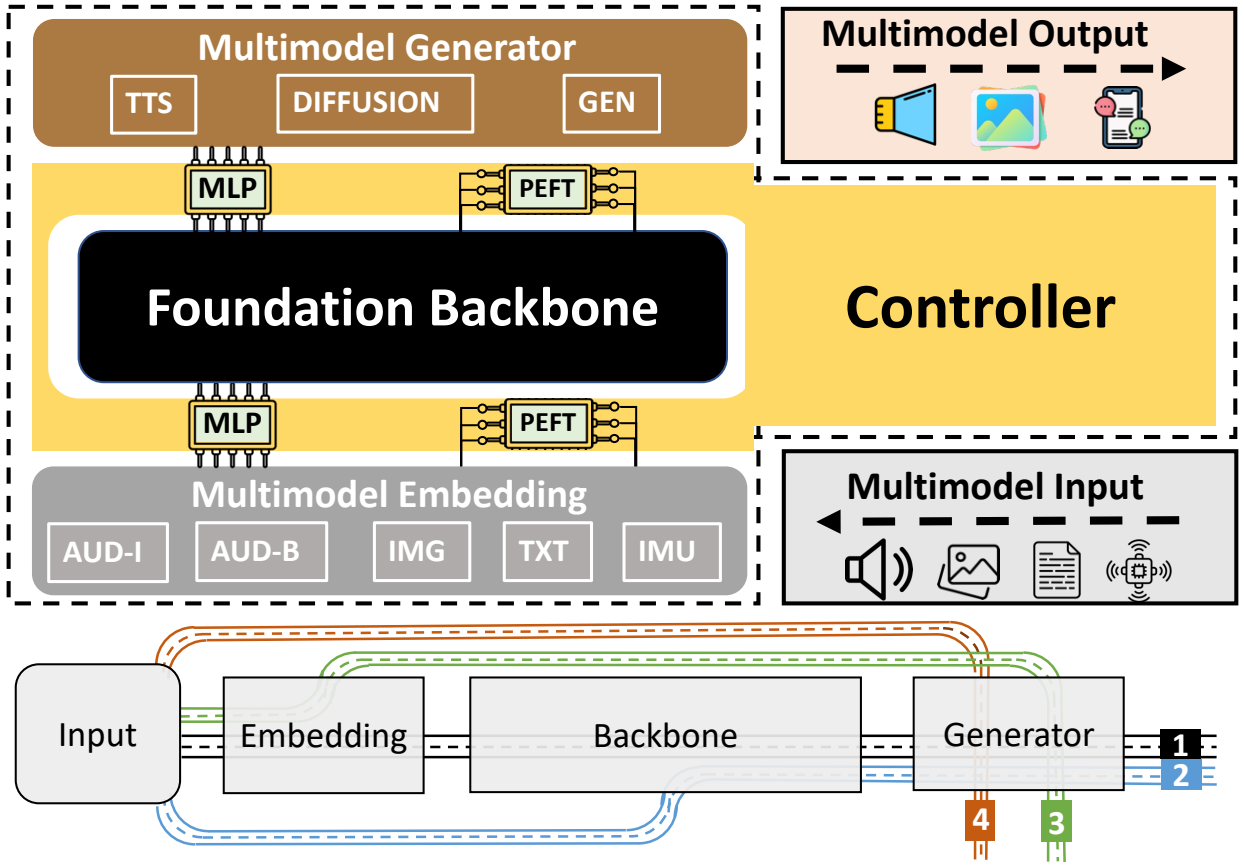
Transformer-based Foundation Models



System Overview



Concrete System Design



- **Prototyping with Off-the-Shelf LLMs**
  - The LLaMA backbone takes 86.1% in terms of parameter size
- **During Inference**
  - Multi-path Task Execution
- **During Training**
  - Only PEFT/MLP parameters are updated

Evaluation: Setup

- **Task:** 38 mobile AI tasks,
- **Dataset:** 50 classical datasets
- **Modality:** 5 popular modalities
- **SoCs:** 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 T3	Intelligent personal assistant (Siri)	SQuAD v2.0	RoBERTa	37	F1: 78.60
			TyDi QA	AraELECTRA	136	F1: 86.62
	Machine translation T4	Translator (Google Translate)	wmt22 en-de	Transformer	110	BLEU: 0.34
	Emoji prediction T5	Input method (Gboard)	tweet_eval	RoBERTa	125	Acc: 0.33
	Emotion prediction T6	Conversational analytics (Clarabridge)	go_emotion	RoBERTa	125	Acc: 0.47
	Sentiment analysis T7	Conversational analytics (Clarabridge)	tweet_eval	RoBERTa	125	Acc: 0.77
	Text classification T8 T9	Spam SMS filtering (Truecaller)	ag_news	BERT	110	Acc: 0.77
			SST2	DistilBERT	67	Acc: 0.51
	Grammatical error correction T10	Writing assistance (Grammarly)	JFLEG	FLAN-t5	801	BLEU: 0.68
	Text summary T11	Reading assistant (ChatPDF)	CNN Daily Mail	BART	400	BLEU:0.43
	Code document generation T12	Code editor (Javadoc)	CodeSearchNet	CodeT5-base	220	BLEU:32.9
	Code generation T13	Code editor (Copilot)	Shellcode_IA32	CodeBERT	125	BLEU:91.7
CV	Object detection T14 T15	Augmented Reality (Google Lens)	COCO	Libra-rcnn	42	AP: 0.43
			LVIS	X-Paste	952	AP: 0.51
	Image retrieval T16	Image searcher (Google Photos)	Clothes Retrieval	Resnet50-arcface	31.7	Recall: 0.90
	Super-resolution T17	Video/Image super-resolution (VSCO)	REDS	Real-ESRGAN	16.7	SSIM: 0.83
	Styler transfer T18	Painting & Beatifying (Meitu)	COCO, Wikiart	StyleFlow	16.8	SSIM: 0.45
	Semantic segmentation T19 T20	Smart camera (Segmentix)	ADE20K-150	Deeplabv3plus	40	mIoU: 0.43
			PASCAL VOC 2012	Deeplabv3plus	40	mIoU: 0.90
	Optical character recognitio		l SST2	CLIP	438	Acc: 0.71
	Image classification T22 T23		0	GFNet	54	Acc: 0.89
			t	Resnet-152	93	Acc: 0.91
	Traffic sign classification T24			MicronNet	0.43	Acc: 0.98
	Vehicle re-identification T25			MSINet	2.3	Rank: 0.96
	Gender recognition T26			MiVOLO-D1	27.4	Acc: 0.96
	Location recognition T27		211	CLIP	438	Acc: 0.46
	Pose estimation T28			ViTPose	44	Acc: 0.69
	Video classification T29	Video player (YouTube)	kinetics400	SlowFast	66	Acc: 0.89
	Counting T30	Smart camera (Fitness Tracking)	Crowd Counting	SASNet	37	MAE: 437
	Image matting T31	Virtual backgrounds (Zoom)	RefMatte-RW100	MDETR	170	MSE: 0.06
Audio	Automatic speech recognition T32	Private assistant (Siri)	LibriSpeech	CTC+attention	120	WER: 3.16%
	Spoken language understanding T33 T34	Private assistant (Siri)	FSC	Transformer	39	WER: 0.37%
			SLURP	CRDNN	26.5	Acc: 0.82
	Emotion recognition T35	Emoji recommendation (WeChat)	IEMOCAP	ECAPA-TDNN	16	Acc: 0.64
	Audio classification T36	Music discovery (Shazam)	ESC-50	ACDNet	4.74	Acc: 0.87
	Keyword spotting T37	Private assistant (Siri)	Speech command	Cnn-trad-fpool3	0.93	Acc: 0.87
Sensing	Human activity recognition T38 T39 T40	AI fitness coach (Keep)	Using Smartphones	MLP	0.28	Acc: 0.90
			HHAR	LIMU-BERT	0.18	Acc: 0.84
			MotionSense	LIMU-BERT	0.18	Acc: 0.91
Multimodal	Text-to-speech T41	Voice broadcast (WeChat reading)	LJSpeech	Transformer	54	MCD: 3.26
	Audio captioning T42 T43	Hearing-impaired accessibility (Ava)	Clotho	Transformer	8.91	BLEU: 0.52
			AudioSet	Transformer	8.91	BLEU: 0.64
	Image captioning T44 T45	Visual-impaired accessibility (Supersence)	MSCOCO'14	LSTM	188	BLEU: 0.69
			Flickr8k	LSTM	188	BLEU: 0.58
	Text-to-image retrieval T46 T47	Image search (Google Photos)	Flickr8k	CCLM	500	Recall: 0.39
			Flickr30k	CCLM	500	Recall: 0.69
	Audio/Text-to-image generation T48	Art creation (Verb Art)	VGGSound	Wav2clip	466	FID: 99.89
	Visual question answering T49 T50	Visual-impaired accessibility (Answerables)	VQA v2.0	MUTAN	218	Acc: 0.63
			VizWiz	MUTAN	218	Acc: 0.52

mAIBench

Evaluation: End-to-end Performance

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

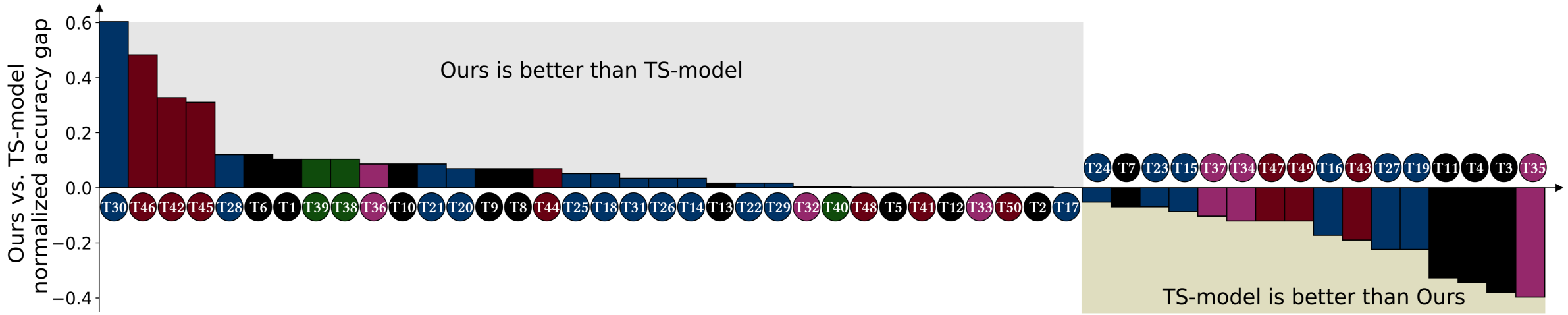
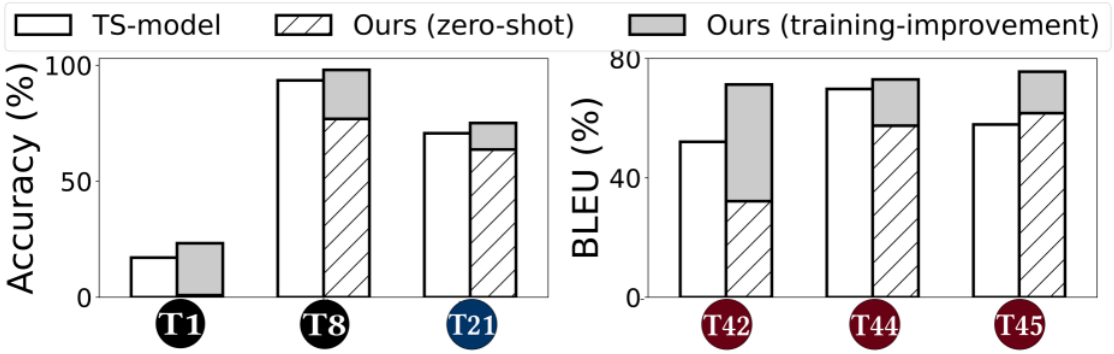


Figure 1. Normalized accuracy comparison of M4 and TS-models on 50 popular mobile tasks and datasets.

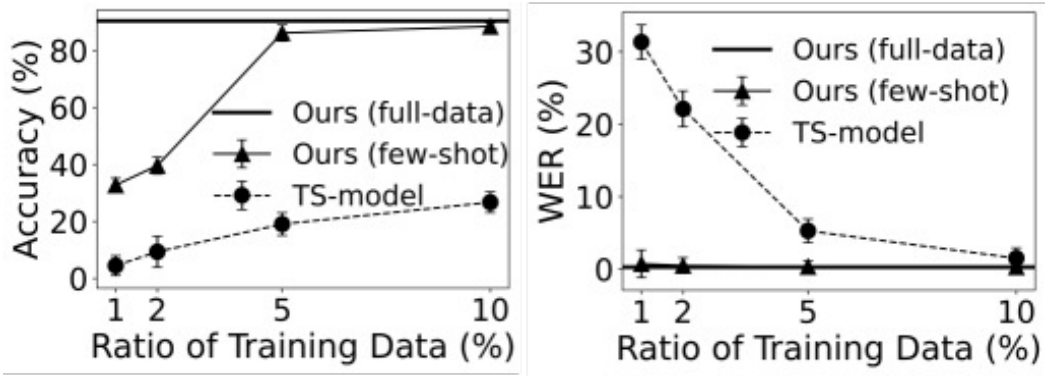
M4 can achieve comparable performance across **85%** of tasks, with over **50%** of these tasks showcasing considerable performance improvement.

Evaluation: Zero/Few-shot Ability

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

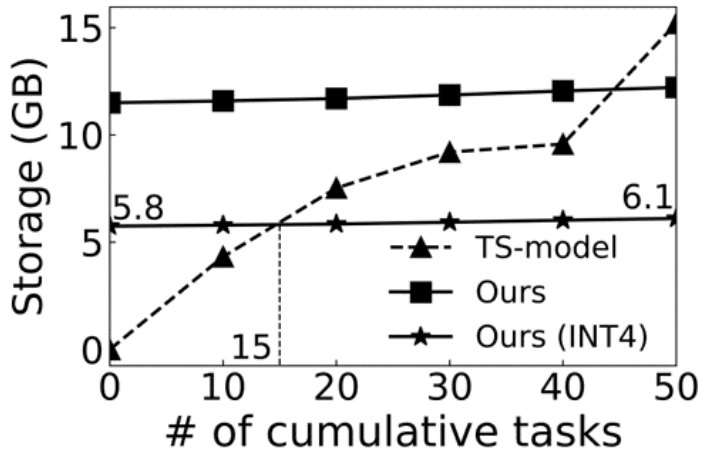


- **M4 has better few-shot ability than TS-models that are trained from scratch.**

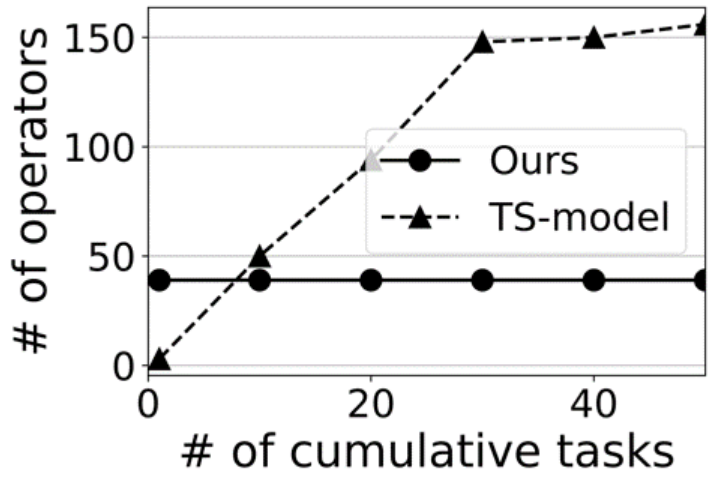


Evaluation: Runtime Cost

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

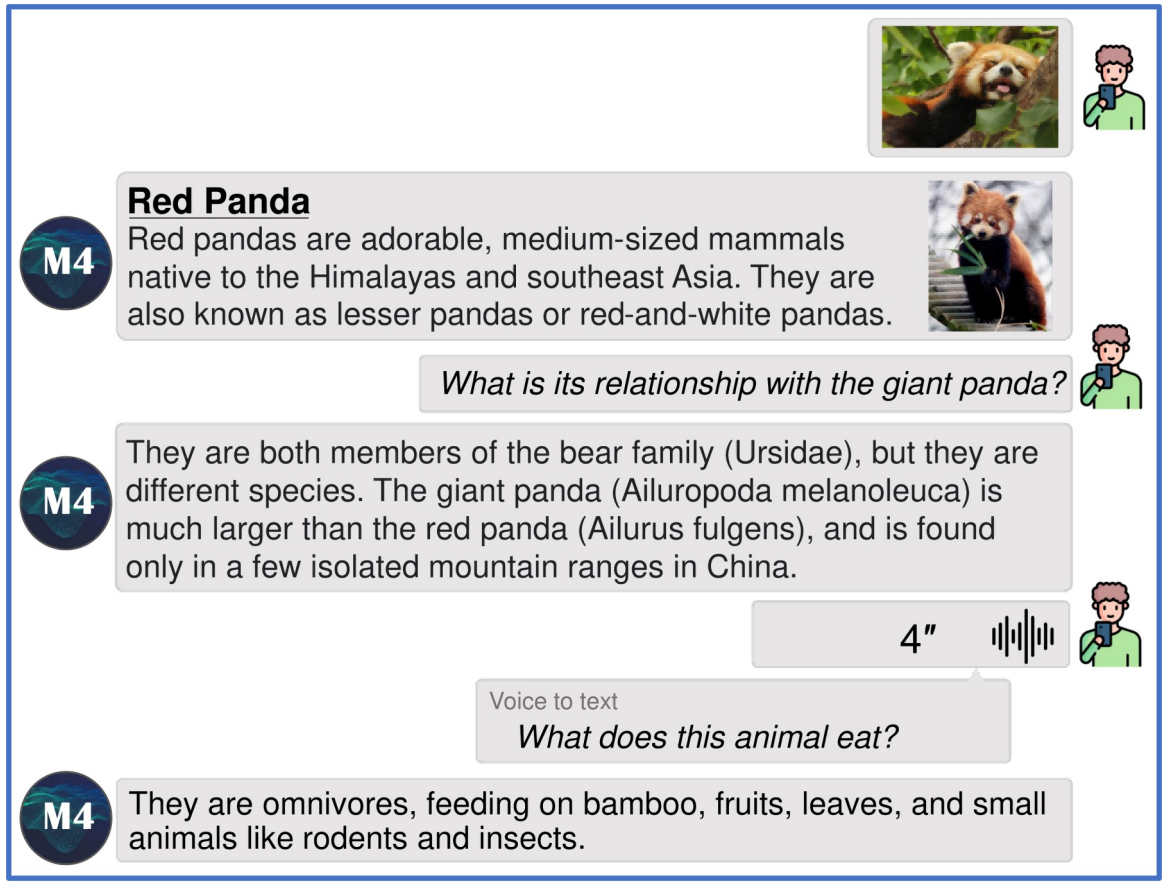


- M4 greatly simplify accelerator design.



Demo: Multimodal Mobile Chat

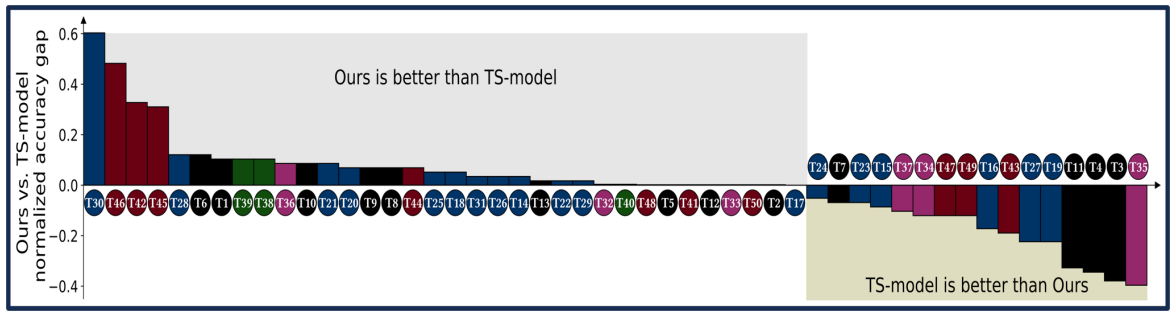
`Illustrative demo`



`Live demo` 🔥

Scaling Law of Mobile Foundation Model

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



Default Backbone

LLaMA 1

2023.02

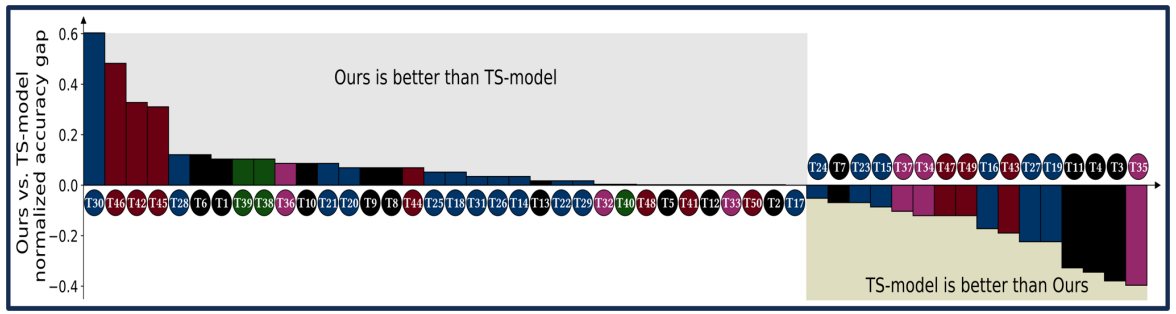
2023.07

LLaMA 2

M4 can be further enhanced with enhanced foundation models.

Scaling Law of Mobile Foundation Model

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



Default Backbone

LLaMA 1

2023.02

2023.07

LLaMA 2

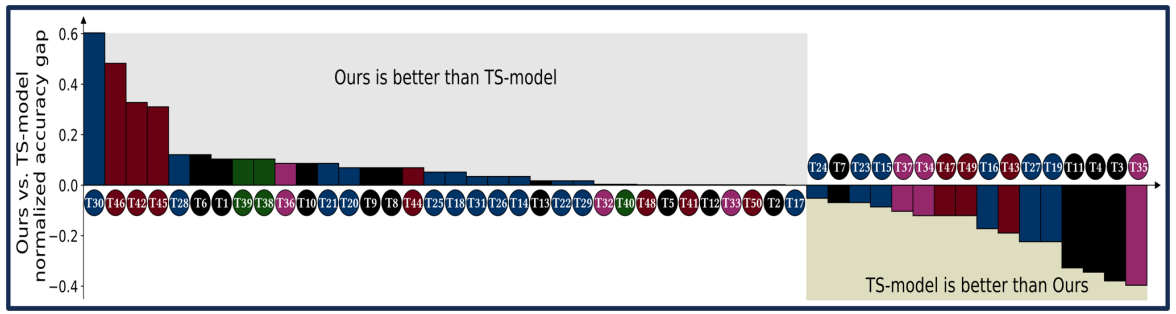
UPGRADE

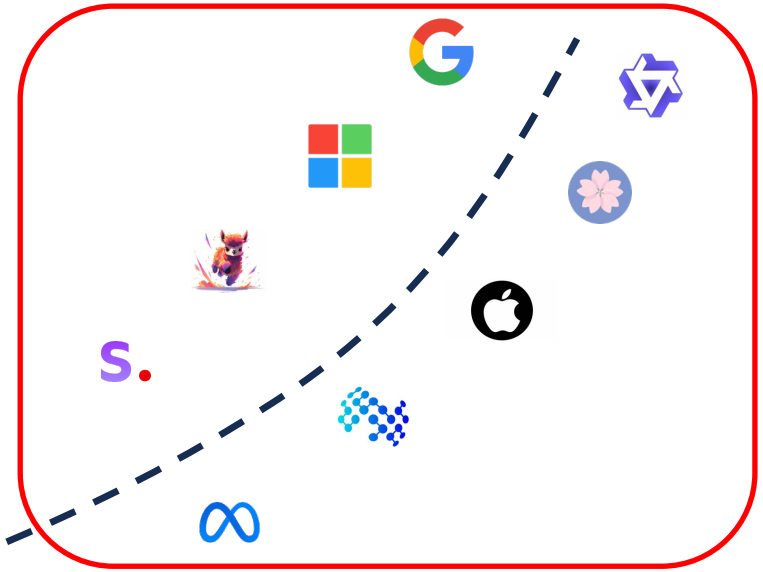
Up to **15%** accuracy improvement

M4 can be further enhanced with enhanced foundation models.

Scaling Law of Mobile Foundation Model

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.

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

Contact: mwx@bupt.edu.cn

Conclusion

- We envision a mobile hardware-OS co-managed multimodal foundation model to serve almost every on-device AI tasks.
- We design and prototype the first such model using off-the-shelf LLMs, all the recipes and weights are opensourced!
- Evaluated on a comprehensive benchmark consisting of 50 representative mobile AI tasks, M4 shows good accuracy, better scalability and reduced runtime cost through its shared weights and operator simplicity.

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

Appendix