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*

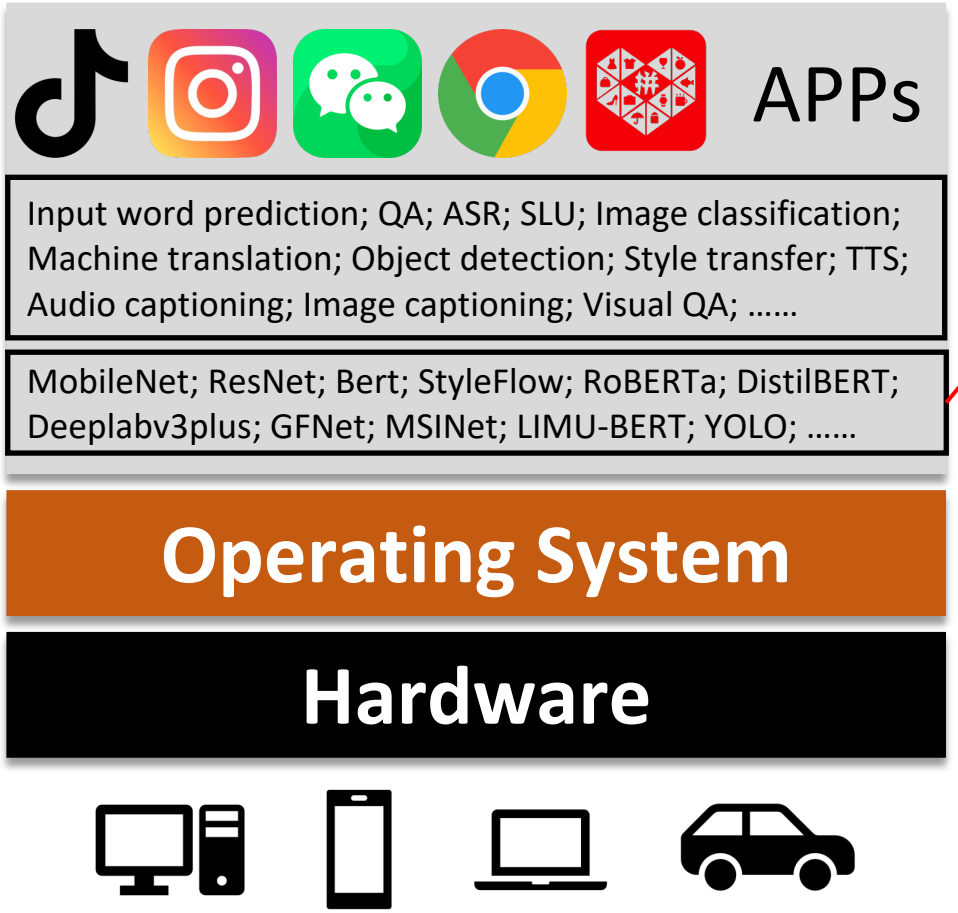
Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia

Shangguang Wang, Mengwei Xu

Beijing University of Posts and Telecommunications

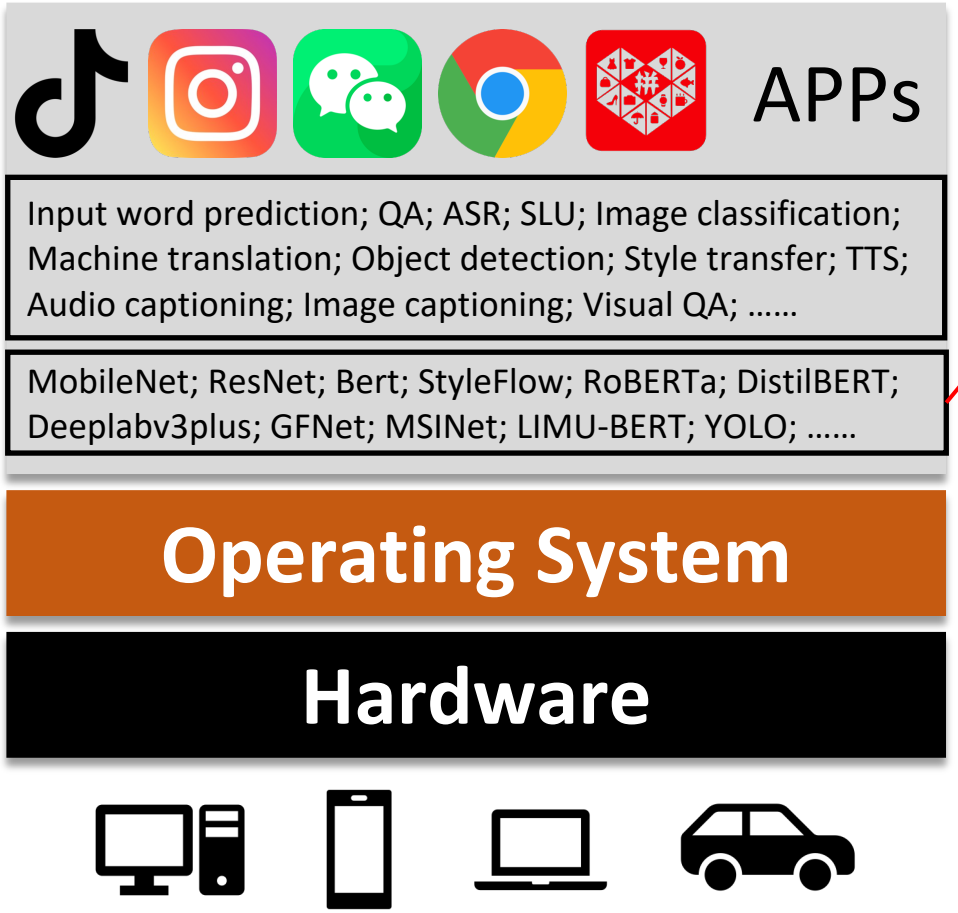
Presenter: Hao Wen (Tsinghua University)

Motivation: Mobile AI Fragmentation



Diversified on-device DNN types, parameters, and configurations

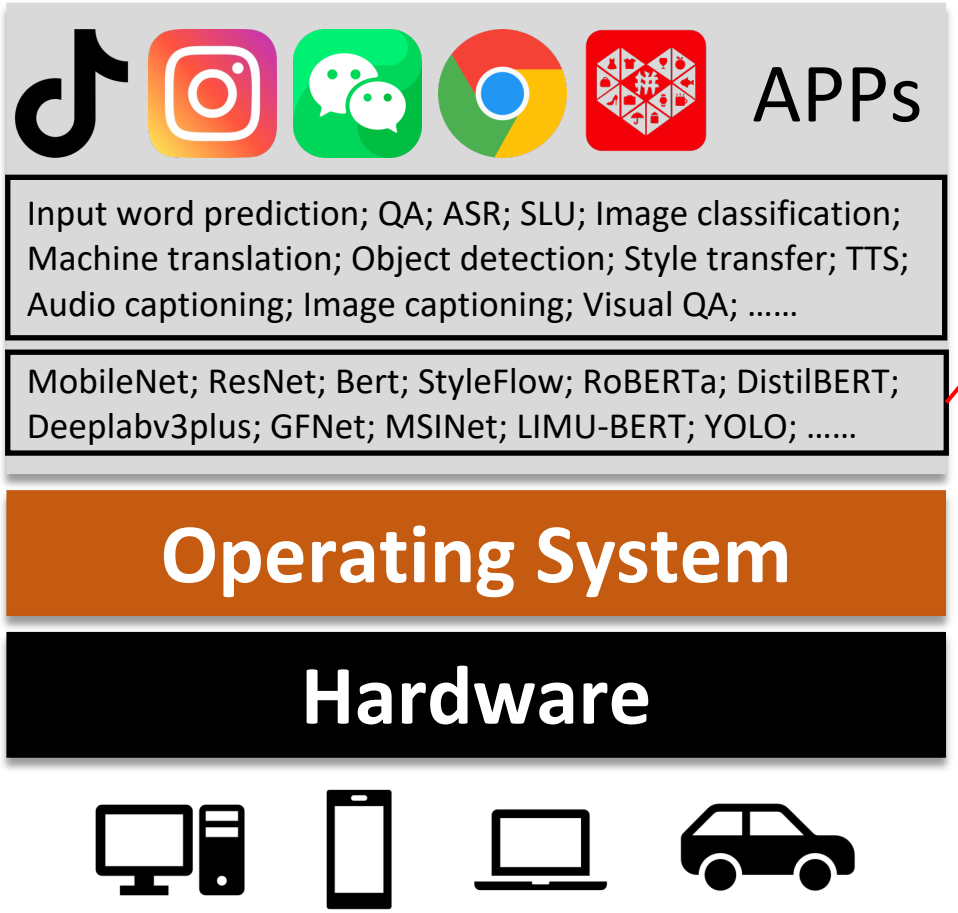
Motivation: Mobile AI Fragmentation



Diversified on-device DNN types, parameters, and configurations

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

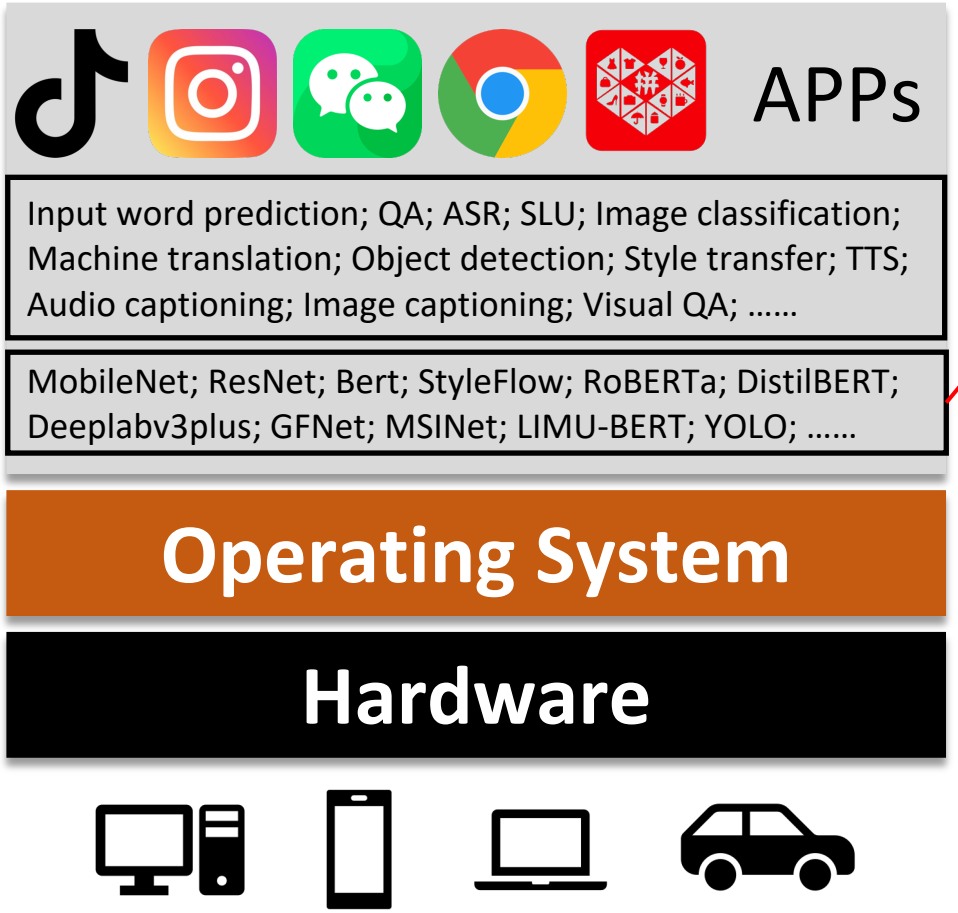
Motivation: Mobile AI Fragmentation



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

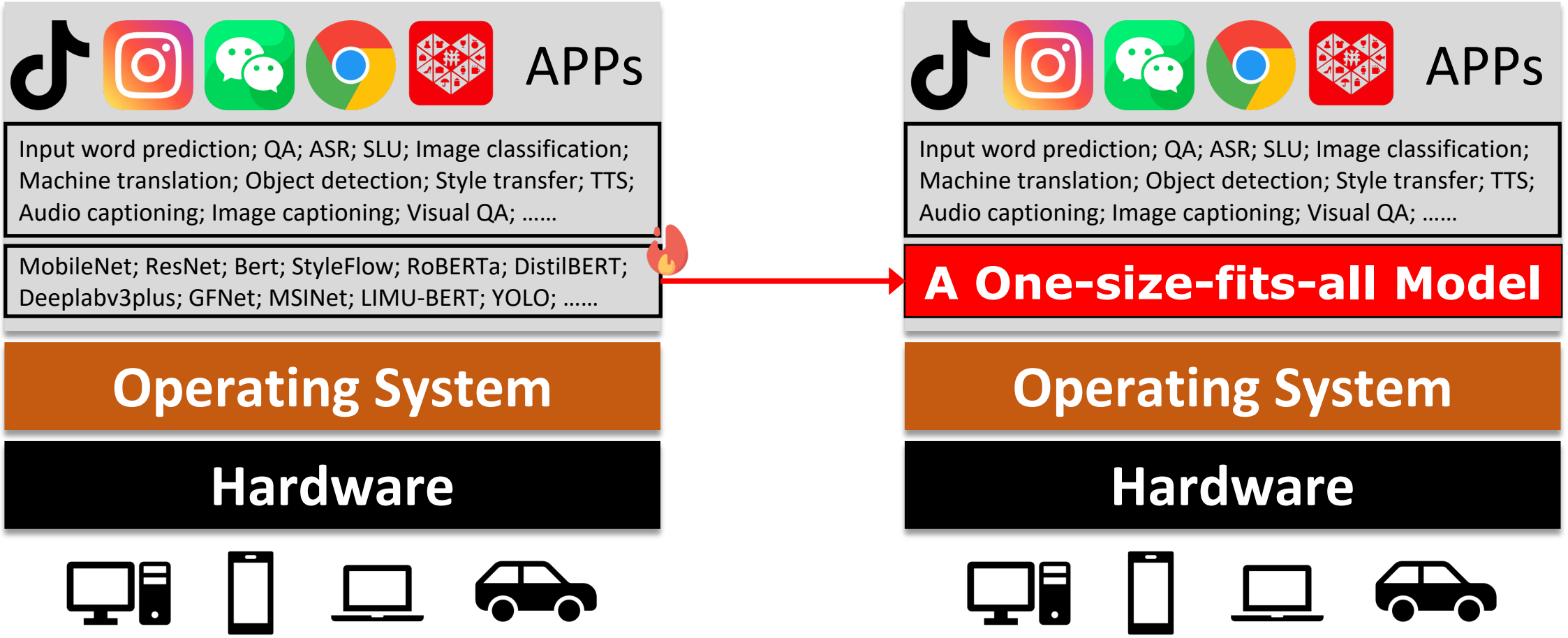
Motivation: Mobile AI Fragmentation



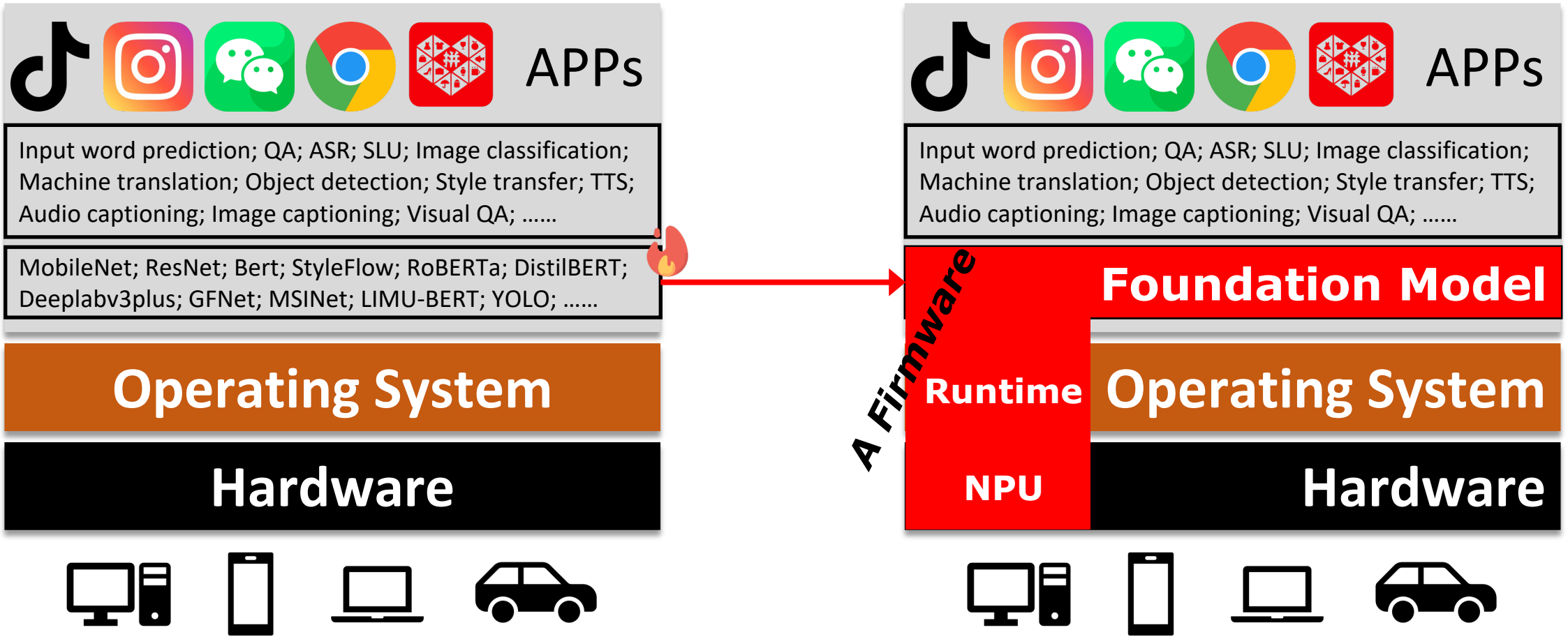
Diversified on-device DNN types, parameters, and configurations

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



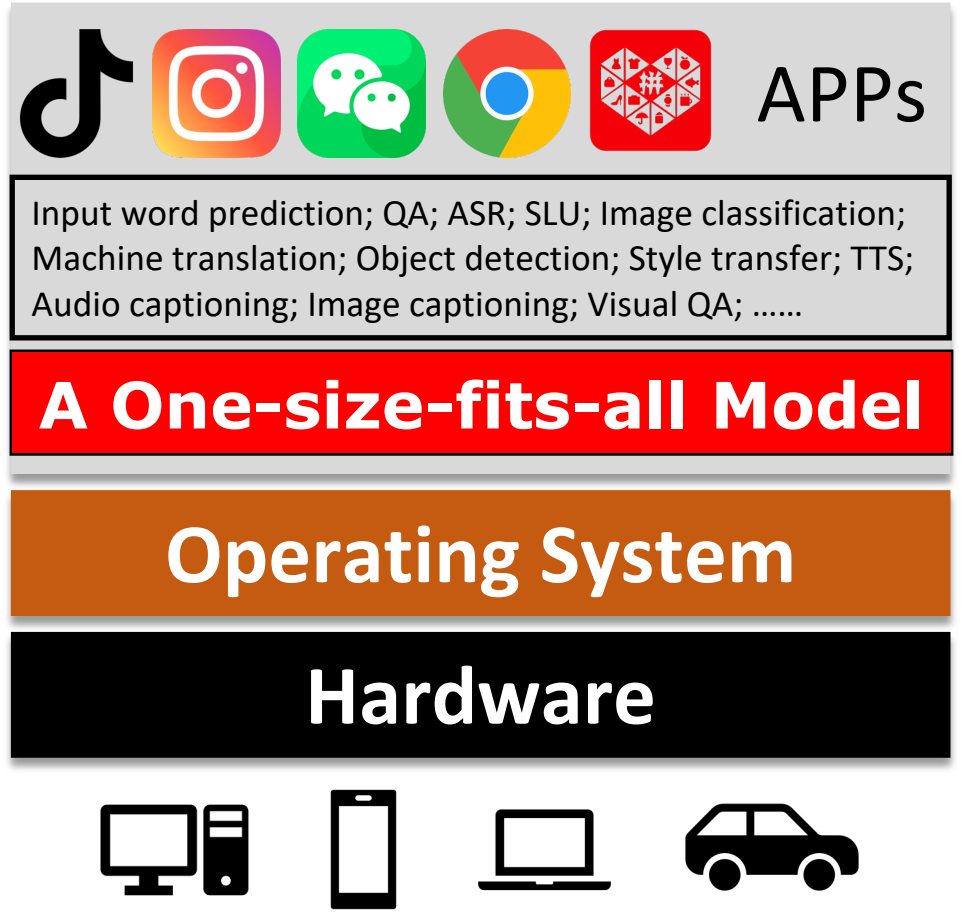
Our Contributions

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

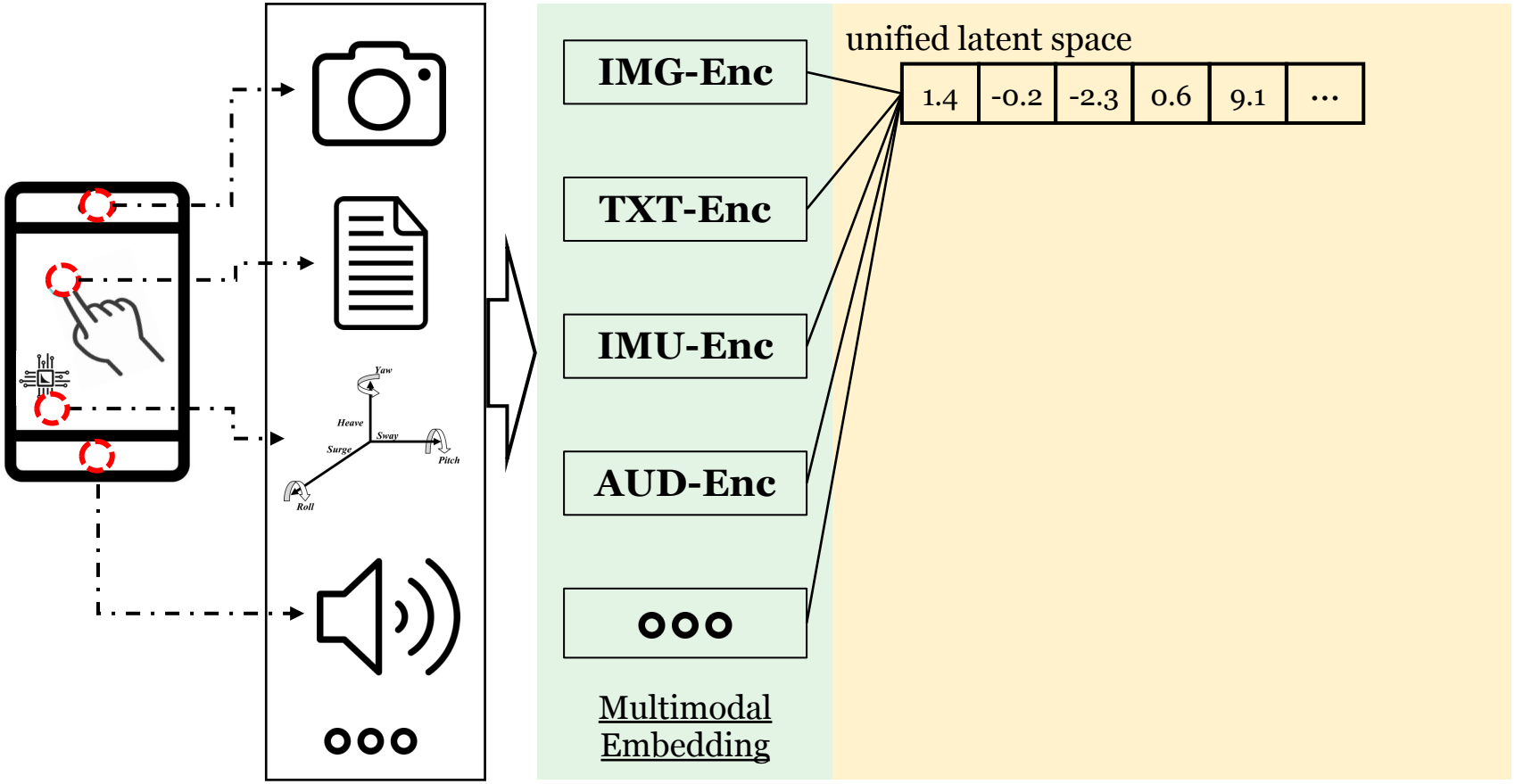
M4 – an any-to-any modality mobile foundation model

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

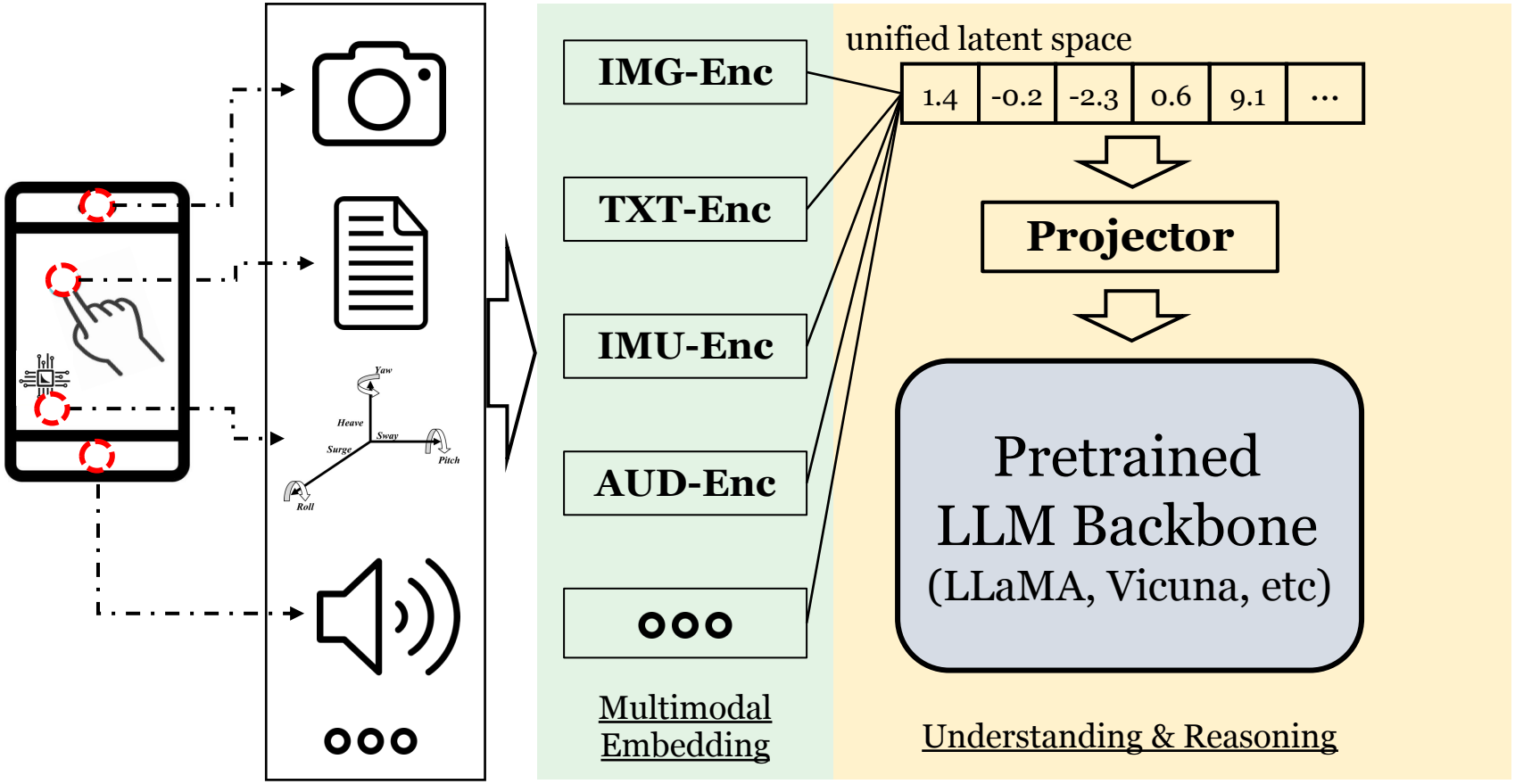
eAIBench – a comprehensive mobile AI benchmark



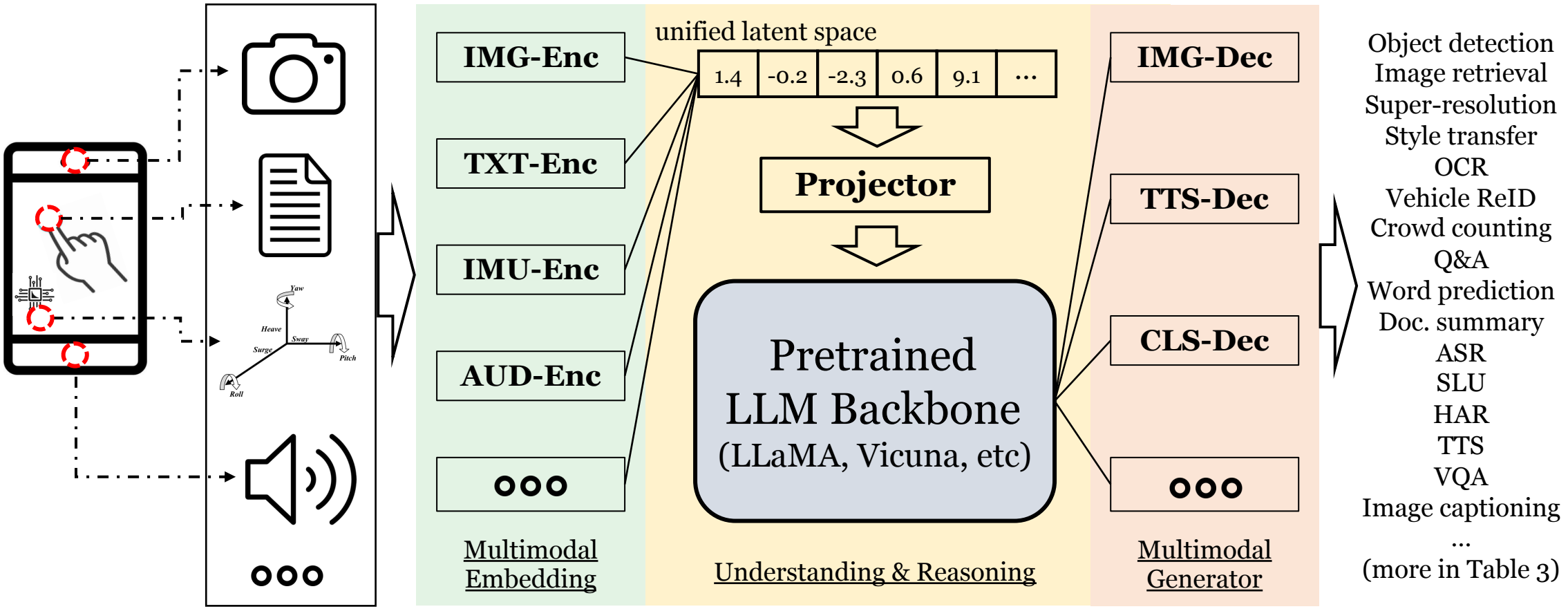
M4 Design


IMG-Enc
TXT-Enc
IMU-Enc
AUD-Enc
ooo
Multimodal Embedding
unified latent space
1.4
-0.2
-2.3
0.6
9.1
…
Yaw
Heave
Surge
Sway
Pitch
Roll
ooo

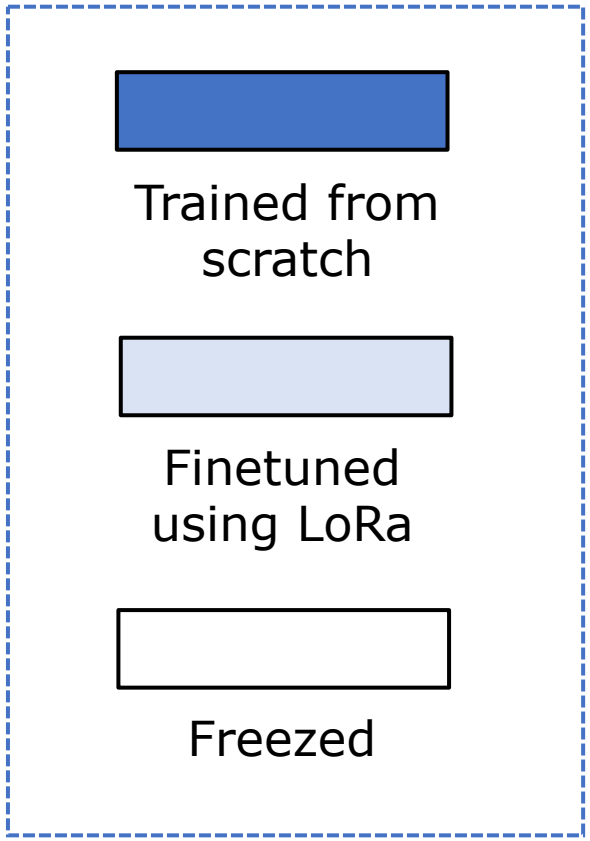
M4 Design

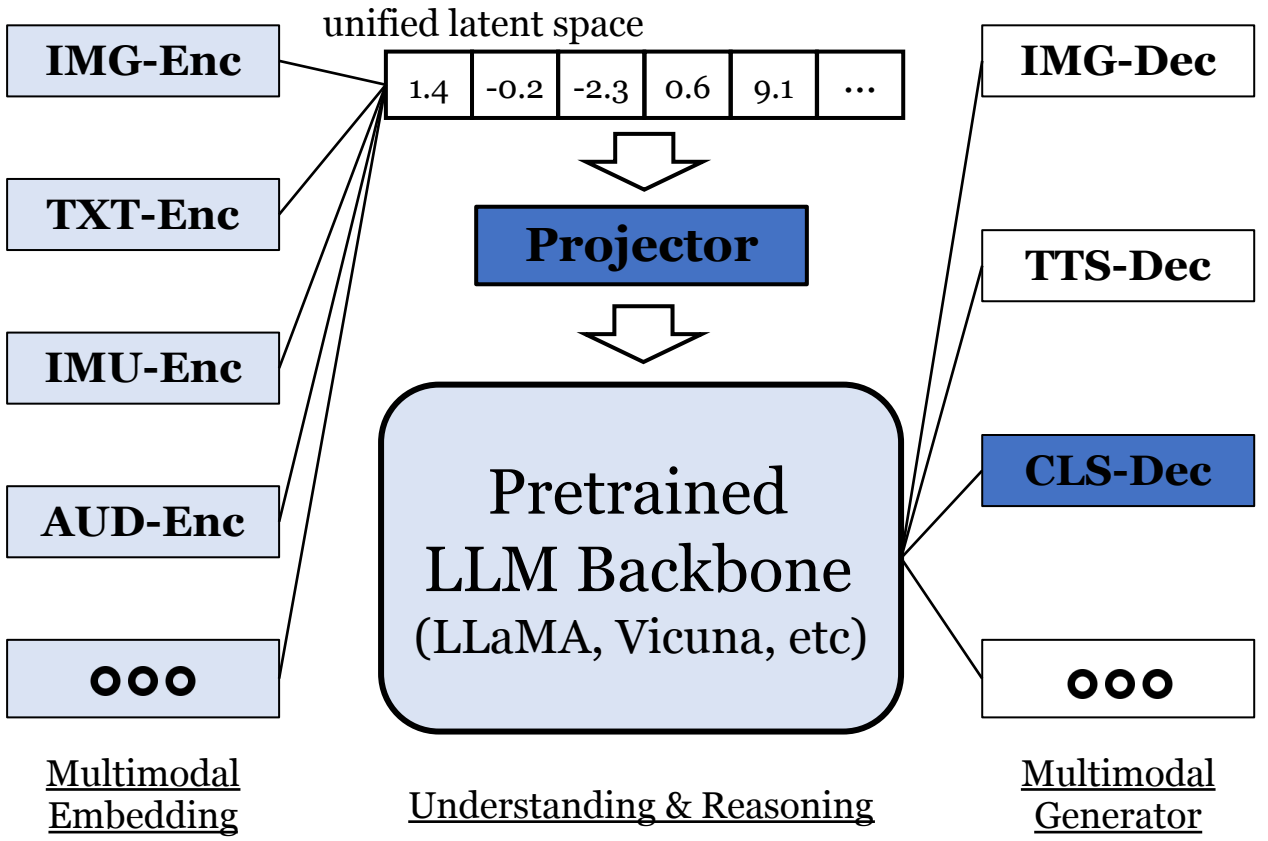

IMG-Enc
TXT-Enc
IMU-Enc
AUD-Enc
ooo
Multimodal Embedding
unified latent space
1.4
-0.2
-2.3
0.6
9.1
…
Projector
Pretrained LLM Backbone
(LLaMA, Vicuna, etc)
Understanding & Reasoning
Yaw
Heave
Surge
Sway
Pitch
Roll

M4 Design


IMG-Enc
TXT-Enc
IMU-Enc
AUD-Enc
ooo
Multimodal Embedding
unified latent space
1.4
-0.2
-2.3
0.6
9.1
...
Projector
Pretrained LLM Backbone (LLaMA, Vicuna, etc)
Understanding & Reasoning
IMG-Dec
TTS-Dec
CLS-Dec
ooo
Multimodal Generator
Object detection
Image retrieval
Super-resolution
Style transfer
OCR
Vehicle ReID
Crowd counting
Q&A
Word prediction
Doc. summary
ASR
SLU
HAR
TTS
VQA
Image captioning
...
(more in Table 3)

M4 Implementation





Total params:
~8.5B

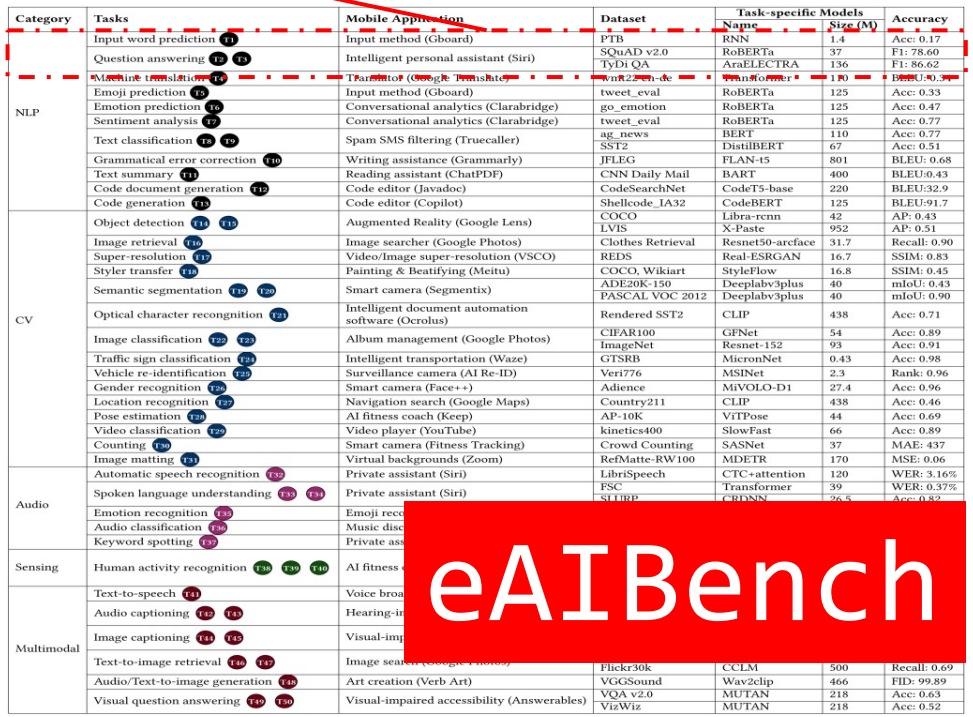
Trainable params:
5M-33M

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 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	1.0	BLEU: 0.36
	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 recognition T21	Intelligent document automation software (Ocrolus)	Rendered SST2	CLIP	438	Acc: 0.71
	Image classification T22 T23	Album management (Google Photos)	CIFAR100	GFNet	54	Acc: 0.89
			ImageNet	Resnet-152	93	Acc: 0.91
	Traffic sign classification T24	Intelligent transportation (Waze)	GTSRB	MicronNet	0.43	Acc: 0.98
	Vehicle re-identification T25	Surveillance camera (AI Re-ID)	Veri776	MSINet	2.3	Rank: 0.96
	Gender recognition T26	Smart camera (Face++)	Adience	MiVOLO-D1	27.4	Acc: 0.96
	Location recognition T27	Navigation search (Google Maps)	Country211	CLIP	438	Acc: 0.46
	Pose estimation T28	AI fitness coach (Keep)	AP-10K	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 recc				
	Audio classification T36	Music disc				
	Keyword spotting T37	Private ass				
Sensing	Human activity recognition T38 T39 T40	AI fitness				
Multimodal	Text-to-speech T41	Voice broa				
	Audio captioning T42 T43	Hearing-in				
	Image captioning T44 T45	Visual-imp				
	Text-to-image retrieval T46 T47	Image sear	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

Evaluation: End-to-end Performance

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

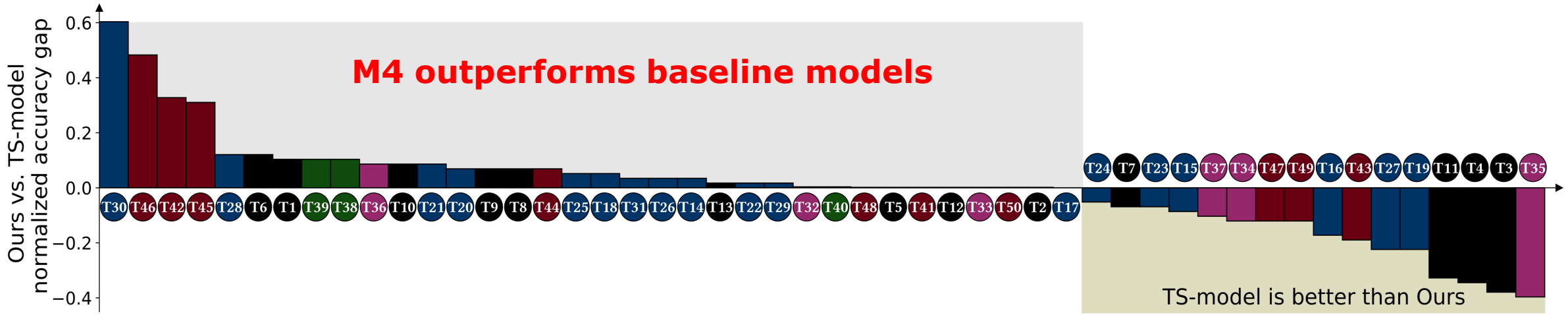
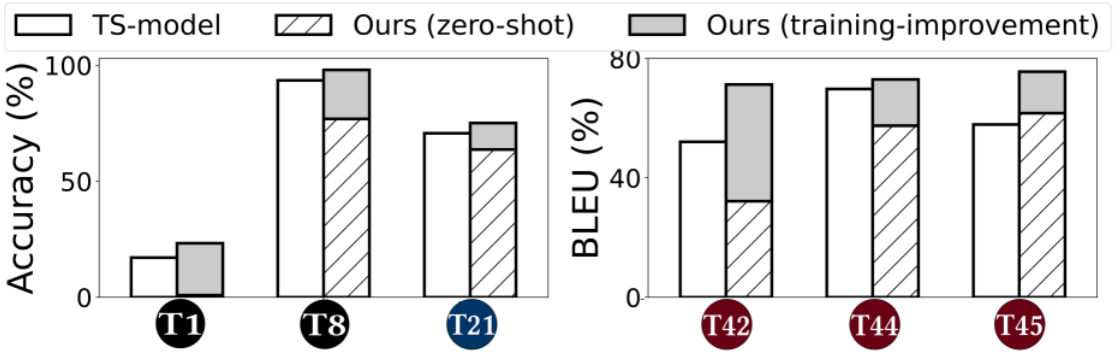


Figure 1. Normalized accuracy comparison of M4 and TS-models on 50 popular mobile tasks and datasets.

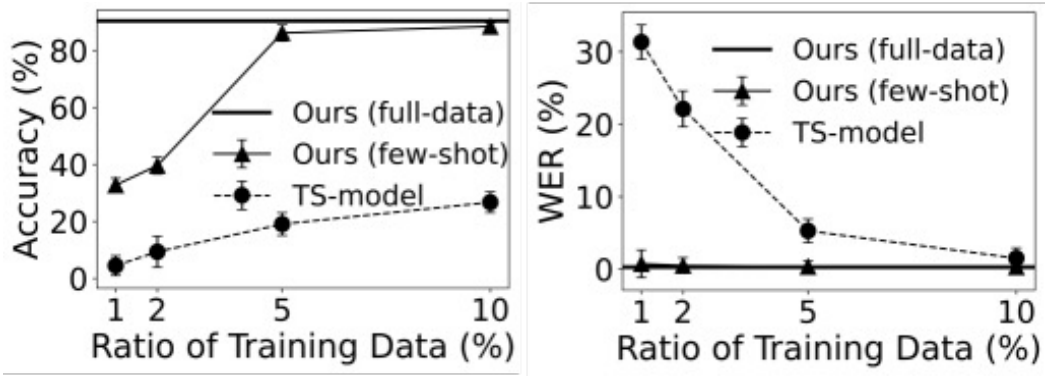
M4 can achieve comparable performance across **85%** of tasks, with over **50%** of these tasks showcasing considerable performance improvement.

Evaluation: Zero/Few-shot Ability

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

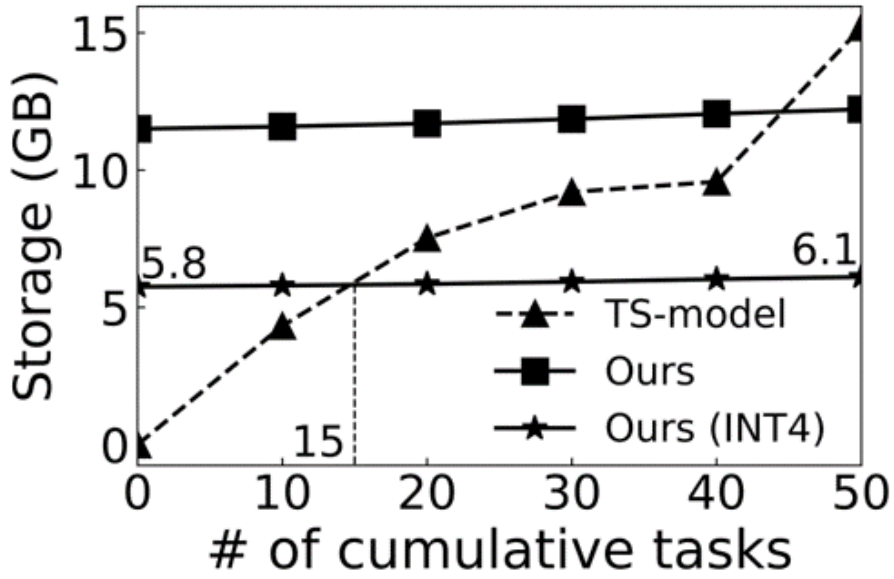


- **M4 has better few-shot ability than TS-models that are trained from scratch.**

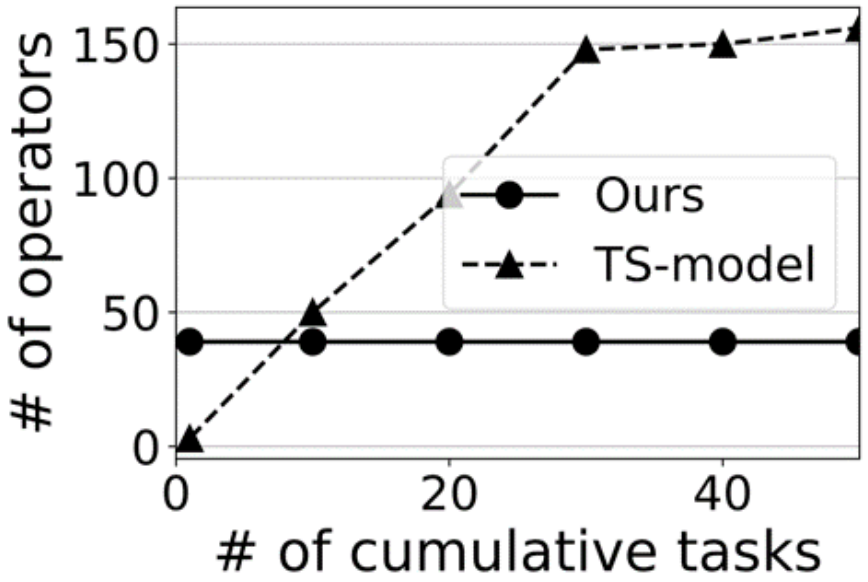


Evaluation: Runtime Cost

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

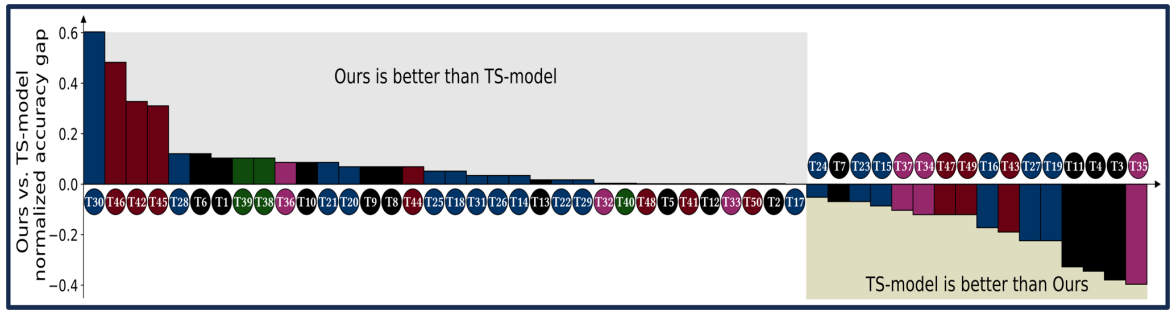


- **M4 greatly simplify accelerator design.**



M4 evolves with better backbones

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



Default Backbone

LLaMA 1

2023.02

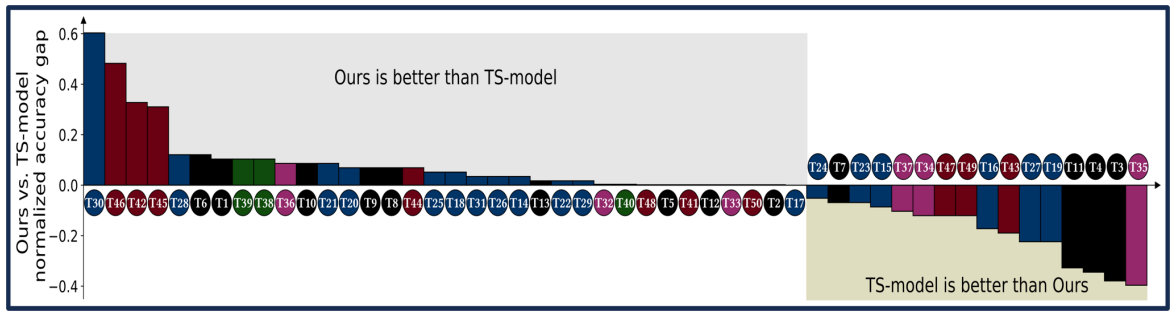
2023.07

LLaMA 2

M4 can be further enhanced with enhanced foundation models.

M4 evolves with better backbones

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



Default Backbone

LLaMA 1

2023.02

2023.07

LLaMA 2

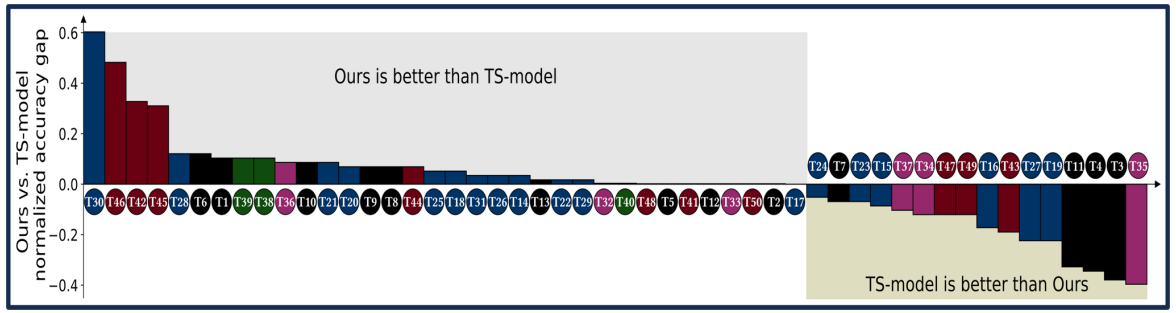
UPGRADE

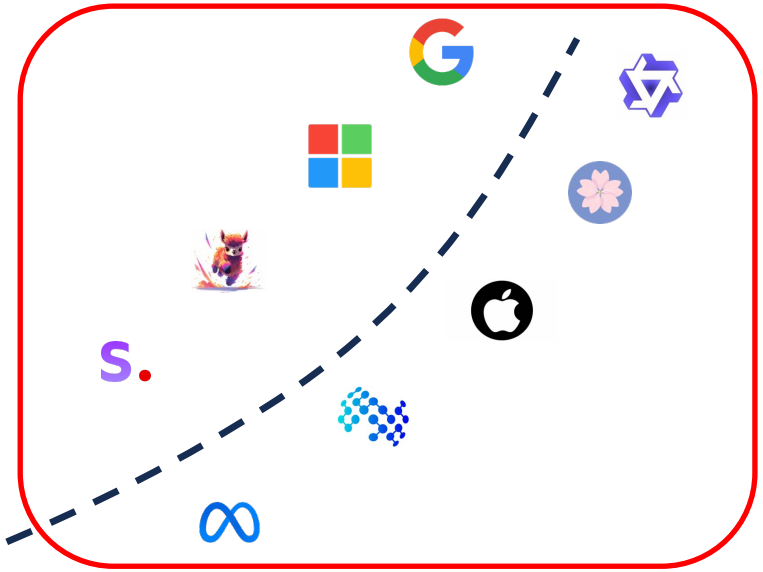
Up to **15%** accuracy improvement

M4 can be further enhanced with enhanced foundation models.

M4 evolves with better backbones

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





Default Backbone

LLaMA 1 → 2023.02

2023.07 → LLaMA 2

UPGRADE

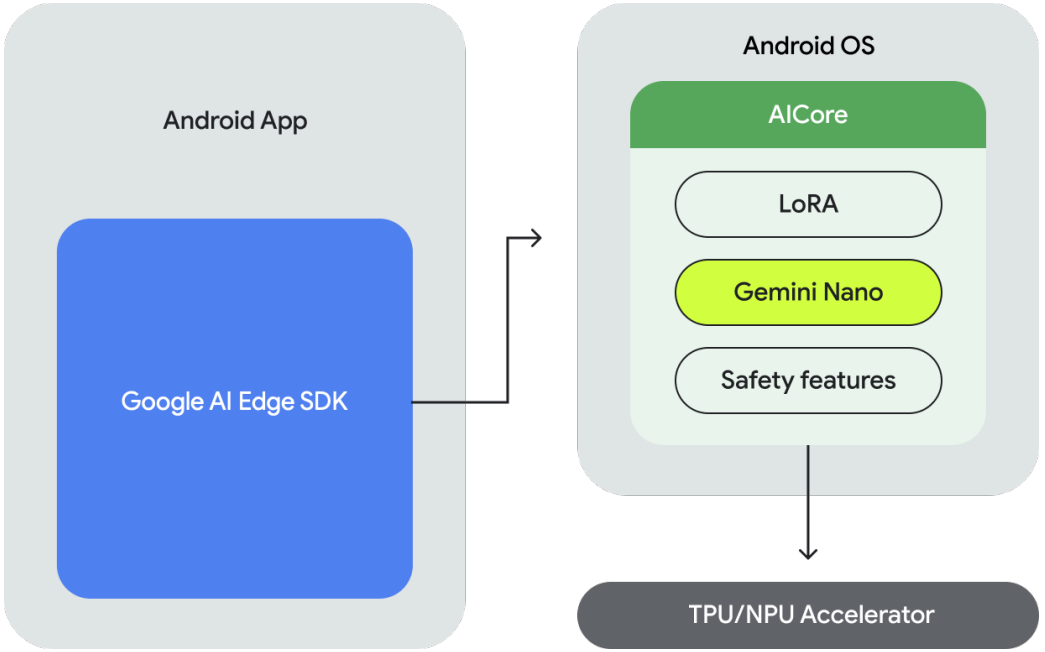
Up to **15%** accuracy improvement

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

Jinliang Yuan*, Chen Yang*, Dongqi Cai*,..., Shangguang 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