Accurate and Efficient Deployment of Large Foundation Models

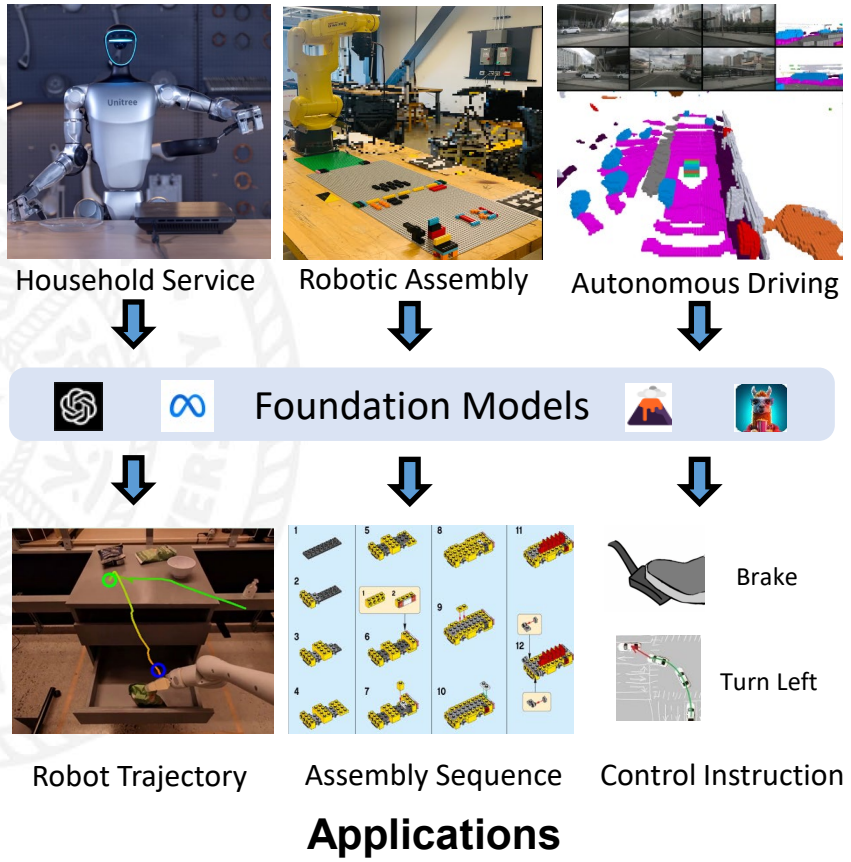
PhD. student Changyuan Wang

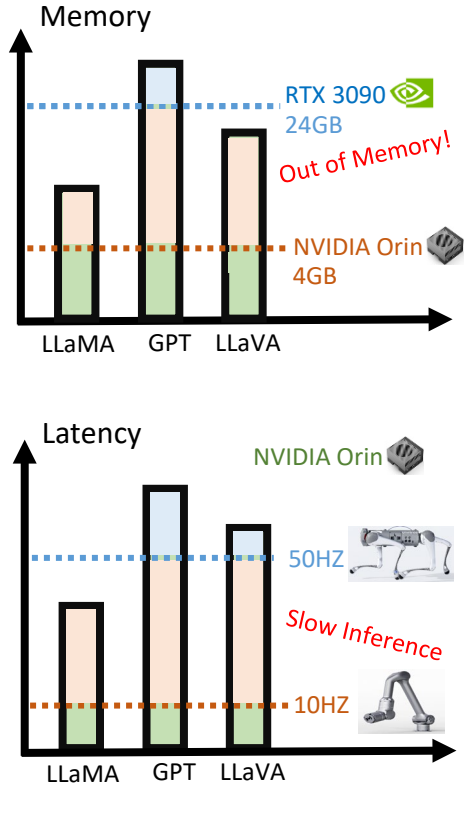
Supervisor: Prof. Yansong Tang

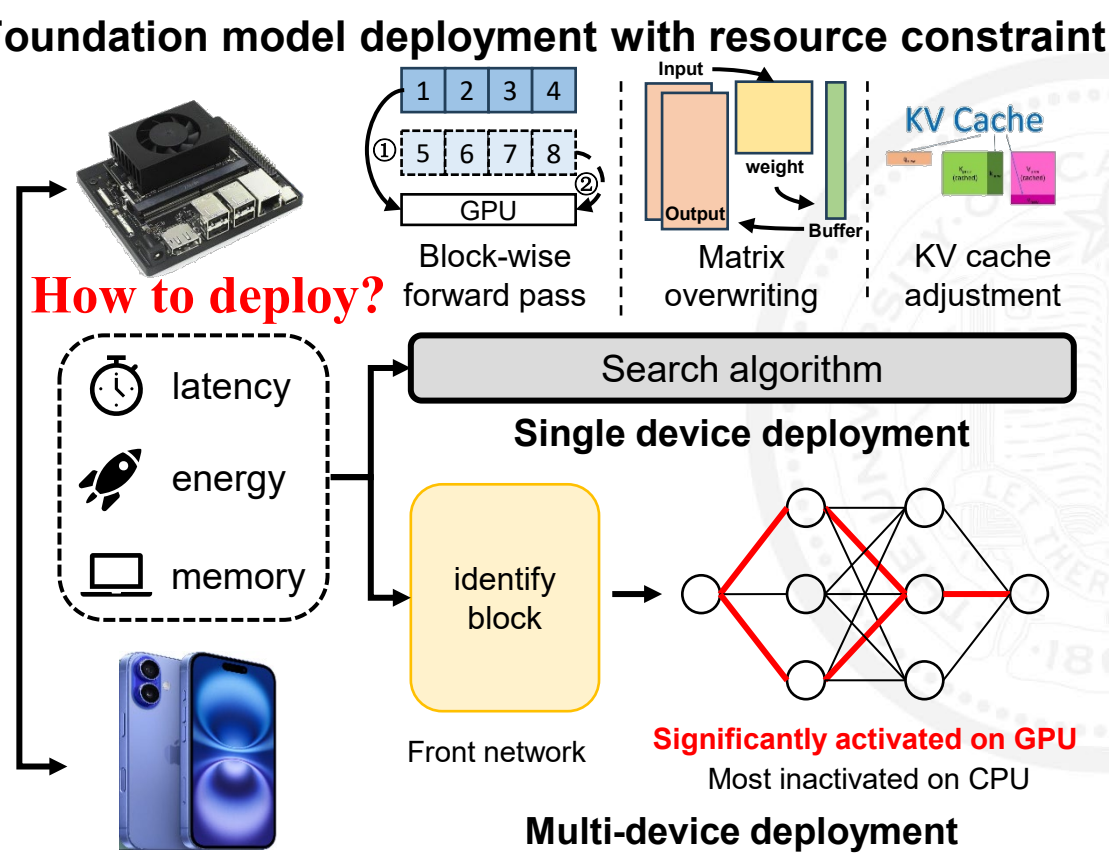
Tsinghua-Berkeley Shenzhen Institude

I. Background – Large Foundation Model

- Deployment of Large foundation models:
  - 1024x1024 diffusion model; 70B parameters vision-language model.
  - How to deploy such large models on single 6GB iPhone 16 pro max ?





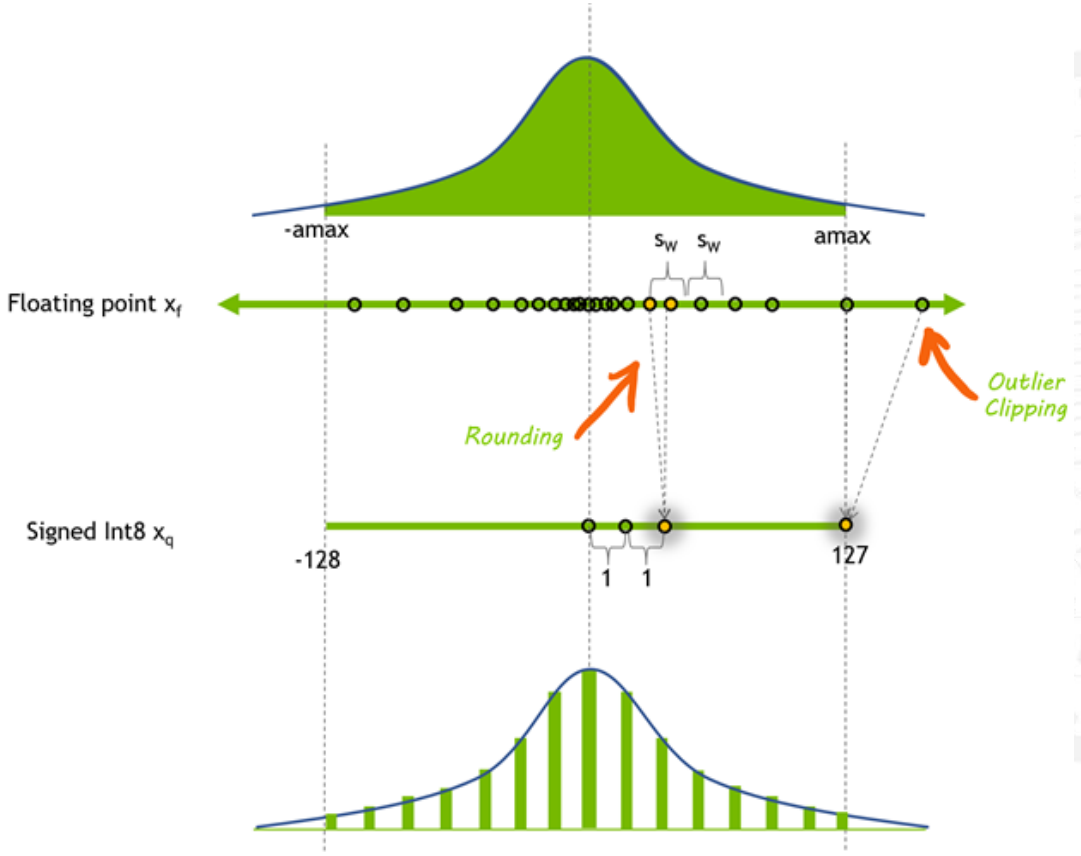


I. Background – Model Compression

- Quantization: replaces FP16 with quantized INT4, INT2
  - Popular compression method with significant efficiency enhancement.
  - 2.78x memory compression, 1.44x generate speed up but accurate drop.

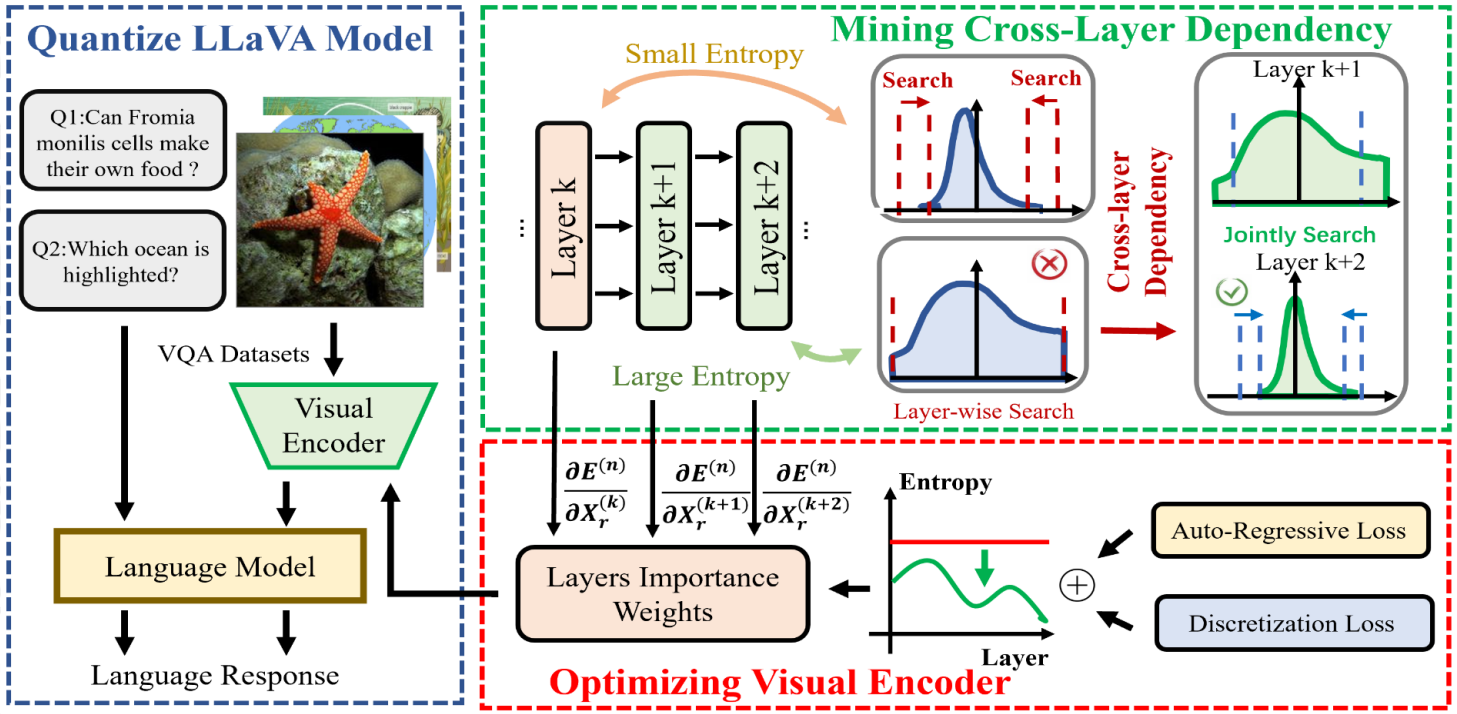
Technical Parameters		Existing Metrics	Expected Metrics
Model Parameters		7B	3500M
Mobile Platform (iPhone 14 Pro Max)	Model Inference Speed (token/s)	50 token/s	80 token/s
	Peak Memory Usage (GB)	12G	6G
	Model Power Consumption (W)	100W	50W
Robot Platform (Jetson Xavier)	Model Inference Speed (token/s)	10 token/s	20 token/s
	Peak Memory Usage (GB)	12G	4G
	Model Power Consumption (W)	20W	6W

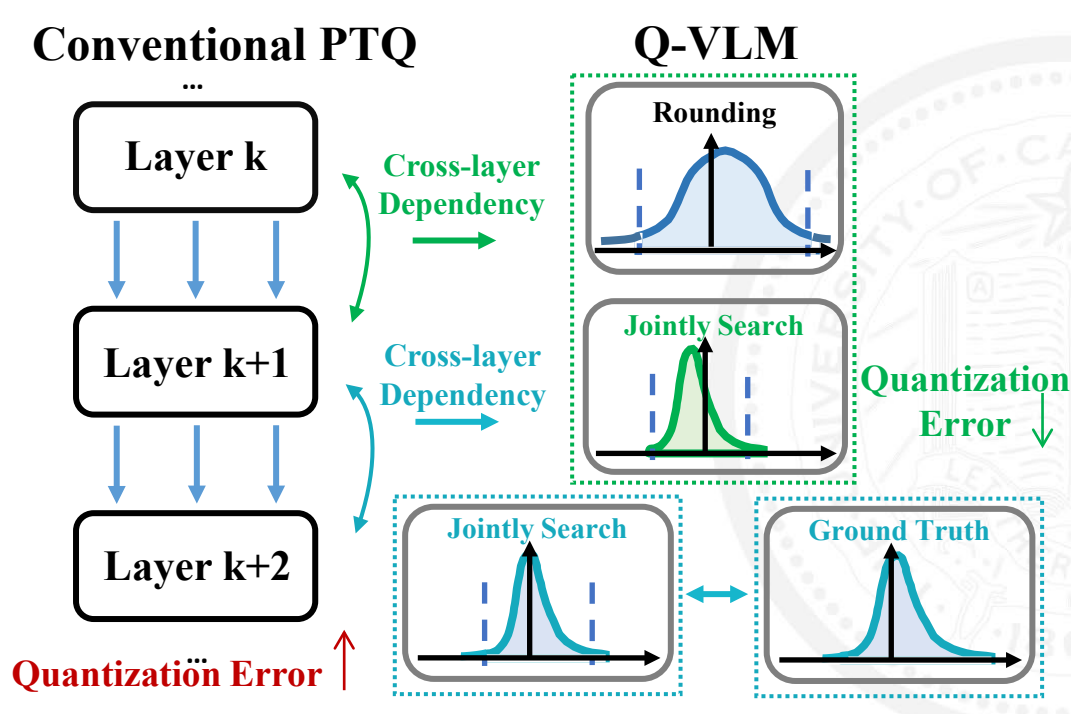
After quantization



III. Efficient Sampling – V-QLM

- Deployment of 4-bits Large Vision-Language models :
  - Conventional methods search the layer-wise rounding functions.
  - Leverage Entropy to mine cross-layer dependency for block-wise search.



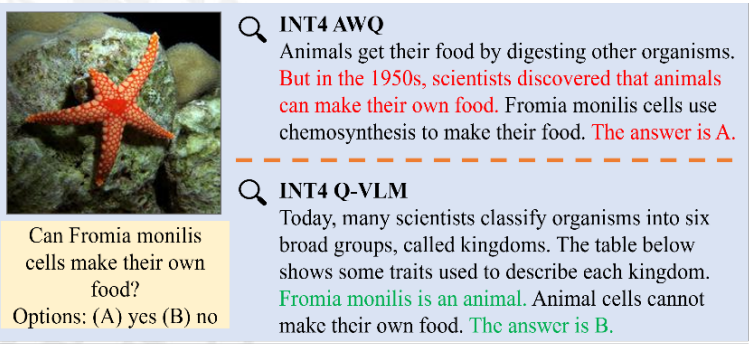


Wang, Changyuan, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. "Q-VLM: Post-training Quantization for Large Vision-Language Models." (**NeurIPS**), 2024. (GitHub ***33*** stars.)

III. Efficient Sampling – V-QLM

Experiments:

- We compresses the memory by 2.78x and increase the generate speed by 1.44x about 13B LLaVA model without performance degradation on diverse multi-modal reasoning tasks.



	Bacteria	Archaea	Protists	Fungi	Animals	Plants
How many cells do they have?	one	one	one or many	one or many	many	many
Do their cells have a nucleus?	no	no	yes	yes	yes	yes
Can their cells make food?	some species can	some species can	some species can	no	no	yes

	Bits	Method	Subject NAT	Subject SOC	Subject LAN	Context Modality TXT	Context Modality IMG	Context Modality NO	Average
LLaVA-7B	FP	-	89.39	96.06	85.64	88.71	87.65	88.50	89.81
	W6A6	AWQ	85.39	92.01	83.27	84.80	83.54	85.99	86.23
		QLoRA	88.45	94.71	84.45	87.63	86.07	87.87	88.43
		Q-VLM	89.43	95.73	84.00	88.71	87.51	87.25	**89.34**
	W4A4	AWQ	74.33	72.22	74.82	73.41	67.13	77.98	74.02
		QLoRA	77.53	75.48	79.18	76.64	70.70	81.95	77.53
		Q-VLM	80.86	75.93	80.73	80.01	72.48	83.90	**79.79**
LLaVA-13B	FP	-	90.19	93.14	87.09	89.39	87.06	89.83	90.00
	W6A6	AWQ	88.03	92.60	84.00	86.02	85.18	86.41	87.57
		QLoRA	88.87	92.89	85.64	87.59	86.56	87.53	88.87
		Q-VLM	89.54	93.18	86.50	88.12	87.01	88.85	**89.70**
	W4A4	AWQ	80.71	70.61	78.49	79.46	70.76	81.82	77.91
		QLoRA	79.62	71.43	82.45	78.25	68.42	85.30	78.64
		Q-VLM	82.55	73.32	83.18	81.03	70.82	86.74	**80.78**

layer-wise search

block-wise search

Table 2: Comparisons with the state-of-the-arts post-training quantization methods for LLaVA-v1.3 and MoE-LLaVA models across bitwidth setting.Results (accuracy) on Science QA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context.

Method	FP Time	FP Memory	FP Accuracy	W8A8 Time	W8A8 Memory	W8A8 Accuracy	W4A4 Time	W4A4 Memory	W4A4 Accuracy
QLoRA	12.9h	24.0G	90.00	16.7h	16.5G	89.32	17.0h	10.7G	78.64
AWQ				11.2h	17.2G	88.94	8.9h	11.2G	77.91
Q-VLM				11.2h	15.7G	89.82	8.9h	9.6G	80.78

Table 4: Comparisons with the state-of-the-arts post-training quantization methods for LLaVA-v1.3-13B models about inference time, memory and accuracy in Science QA dataset.

Thanks for your attention !

PhD. student Changyuan Wang

Supervisor: Yansong Tang

Tsinghua-Berkeley Shenzhen Institude