香港中文大學
The Chinese University of Hong Kong

GenieBlue: Integrating both Linguistic and Multimodal Capabilities for Large Language Models on Mobile Devices

Xudong Lu

Email: luxudong@link.cuhk.edu.hk

Algorithm-System Co-Optimization Empowering Edge AI

香港中文大學
The Chinese University of Hong Kong

Motivation

In the deployment process of (M)LLMs on smartphones, we face the storage and memory limitations.

- We aim to deploy a single model that can efficiently handle both pure language tasks and multimodal tasks simultaneously.

Challenge

- MLLMs still cannot achieve satisfactory pure language capabilities currently.
- Mainstream smartphone NPU platforms currently do not support deploying MoE structures.

Solution

We introduce **GenieBlue**, an efficient MLLM structural design that integrates both linguistic and multimodal capabilities for LLMs on mobile devices.

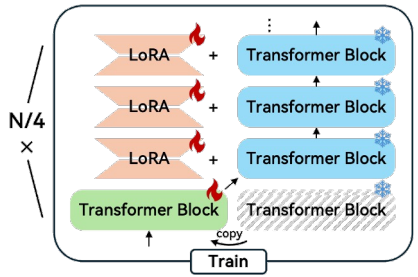
- Training: Freeze the LLM parameters during MLLM training to maintain text-only capabilities.
- Deployment: Employ a non-shared base deployment strategy.

	Model	MATH	AlignBench	MT-Bench
Base LLM	Qwen2.5-3B	61.74	6.00	5.81
MLLM	InternVL2.5-4B	55.20	5.18	4.94
Drop (%)		10.59	13.67	14.97
Base LLM	Qwen2.5-3B	61.74	6.00	5.81
MLLM	Qwen2.5-VL-3B	58.92	5.38	4.72
Drop (%)		4.57	10.33	18.76
Base LLM	Qwen1.5-7B	22.02	5.40	5.77
MLLM	Wings-Qwen1.5-8B	13.96	4.86	4.56
Drop (%)		36.60	10.00	20.97
Base LLM	BlueLM-3B	38.94	5.67	5.42
MLLM	GenieBlue-3B	38.94	5.67	5.42
Drop (%)		**0**	**0**	**0**

Observation 1: MLLM training will lead to a significant decline in pure language capabilities.

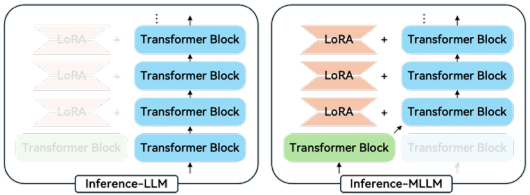
Technical Contribution - Summary

香港中文大學
The Chinese University of Hong Kong



Model Architecture

- ViT: SigLIP-400M
- MLP Projection Layer
- LLM: BlueLM-3B
- **Replicated Transformer Blocks**
- **Added LoRA Module**



Deployment Strategy

- Mixed-Precision Deployment
- **Non-shared base deployment strategy**

1) Problem Discovery and Statement:
We examine the deployment of LLM and MLLMs on current smartphones, identifying performance degradation in text-only tasks and highlighting the limitations of current NPU platforms that do not support the deployment of MoE models.

2) Approach Analyses and Structure Design:
We analyze how to maintain pure language performance during the training of MLLMs from training data and model structure perspectives. Then, we introduce GenieBlue, which integrates linguistic and multimodal capabilities for LLMs on mobile devices through efficient and more hardware-friendly model structural designs.

3) Strong Performance and High Efficiency:
We train GenieBlue with a large amount of multimodal datasets, achieving capabilities comparable to fully fine-tuned MLLMs without compromising any pure language abilities. We also support the deployment of GenieBlue on actual smartphone NPUs.

Technical Contribution – Approach Analyses

香港中文大學
The Chinese University of Hong Kong

Approach Analyses – Data Perspective

Type	#Samples	Datasets
General QA	840k	UltraFeedback [22], UltraChat [23], NoRobots [61], LIMA [93], SlimOrca [42], WizardLM-Evol-Instruct-70K [76], Llama-3-Magpie-Pro [77], Magpie-Qwen2-Pro [77], Firefly [81], Dolly [19], OpenAI-Summarize-TLDR [9], Know-Saraswati-CoT [35]
Code	360k	Code-Feedback [92], Glaive-Code-Assistant [26], XCoder-80K [73], Evol-Instruct-Code [56]
Mathematics	830k	GSM8K-Socratic [17], NuminaMath-TIR [37], NuminaMath-CoT [38], InfinityMATH[87], MathQA [2], MetaMathQA [83]

Table 2. We expand the Cambrian-7M dataset with 2M pure text data training samples, primarily sourced from the InternVL2.5 paper [12].

BlueLM-3B	#Samples	AI2D	ChartQA	DocVQA	OCRBench	RealWorldQA	ScienceQA	TextVQA	AVG
MLLM Tasks	7M	74.81	68.32	74.60	55.30	62.35	67.91	60.06	66.19
	7M+2M	74.03	69.36	74.63	56.70	58.04	68.24	62.34	66.19
BlueLM-3B	**#Samples**	**DROP**	**GPQA**	**GSM8K**	**MATH**	**MMLU**	**AlignBench**	**MT-bench**	**AVG**
LLM Tasks	-	81.57	29.46	86.13	38.94	74.13	5.67	5.42	60.16
	7M	62.49	23.21	66.11	19.26	57.50	3.87	3.92	43.78
	7M+2M	64.67	28.80	69.90	30.60	57.67	3.84	3.92	47.03
Qwen2.5-3B	**#Samples**	**AI2D**	**ChartQA**	**DocVQA**	**OCRBench**	**RealWorldQA**	**ScienceQA**	**TextVQA**	**AVG**
MLLM Tasks	7M	77.20	67.36	68.84	54.70	61.05	68.19	57.72	65.01
	7M+2M	76.98	68.48	64.25	56.20	62.09	69.43	55.54	64.71
Qwen2.5-3B	**#Samples**	**DROP**	**GPQA**	**GSM8K**	**MATH**	**MMLU**	**AlignBench**	**MT-bench**	**AVG**
LLM Tasks	-	70.82	30.30	74.75	61.74	66.31	6.00	5.81	60.29
	7M	69.38	20.71	68.54	31.46	63.46	4.61	4.54	49.29
	7M+2M	71.45	27.78	69.37	40.18	64.34	4.36	4.34	51.45

Table 3. We fully fine-tune BlueLM-V-3B from scratch (with SigLIP [86] and BlueLM-3B [53]/Qwen2.5-3B [80]) using Cambrian 2.5M pre-training data and 7M fine-tuning data. We also conduct fine-tuning by adding 2M text-only data to the Cambrian-7M fine-tuning dataset. The inclusion of text-only data does not cause obvious degradation in MLLM performance and partially improves the accuracy on objective NLP tasks, but does not help with subjective NLP tasks (#Samples denotes the number of fine-tuning data samples).

Observation 2: Adding pure-text datasets has little impact on the MLLM performance.

Observation 3: Adding pure-text data leads to an improvement in objective NLP tasks but does not assist with subjective tasks.

Approach Analyses – Model Structure Perspective

BlueLM-3B	#Param	AI2D	ChartQA	DocVQA	OCRBench	RealWorldQA	ScienceQA	TextVQA	AVG	Retention (%)
Full-Finetune	3161.26M	74.03	69.36	74.63	56.70	58.04	68.24	62.34	66.19	-
LoRA	458.06M	68.23	61.24	66.17	48.70	55.56	68.57	56.97	60.78	91.82
CogVLM-Post	1005.69M	67.81	60.80	66.49	51.00	57.12	67.00	58.58	61.26	92.55
CogVLM-Pre	1005.69M	69.04	64.28	70.23	51.50	52.29	67.67	60.42	62.20	93.98
CogVLM-Skip	1005.69M	70.01	66.36	71.97	54.60	56.34	68.91	59.37	63.94	96.60
Qwen2.5-3B	**#Param**	**AI2D**	**ChartQA**	**DocVQA**	**OCRBench**	**RealWorldQA**	**ScienceQA**	**TextVQA**	**AVG**	**Retention (%)**
Full-Finetune	3527.81M	76.98	68.48	64.25	56.20	62.09	69.43	55.54	64.71	-
LoRA	456.84M	65.35	54.32	55.84	48.10	55.56	72.72	58.40	58.61	90.58
CogVLM-Post	1146.75M	68.72	60.48	65.14	51.30	48.89	64.76	59.85	59.88	92.53
CogVLM-Pre	1146.75M	68.88	62.12	67.95	52.30	53.73	72.87	57.36	62.17	96.08
CogVLM-Skip	1146.75M	69.30	65.92	71.10	54.10	50.59	69.48	59.62	62.87	97.16

Table 4. Evaluation results on MLLM benchmarks. We fine-tune all the models using the 9M dataset, comparing full fine-tuning, LoRA fine-tuning, and CogVLM fine-tuning. **Post**, **Pre**, and **Skip** means adding the visual expert module to the last quarter of the layers, the first quarter of the layers, and at every quarter interval of the layers. Apart from full fine-tuning, other methods can maintain pure language capability consistent with the original LLM during inference through the use of the non-shared base deployment strategy. CogVLM-Skip achieves the best MLLM performance retention. We also provide the trainable parameter numbers (#Param) during MLLM training.

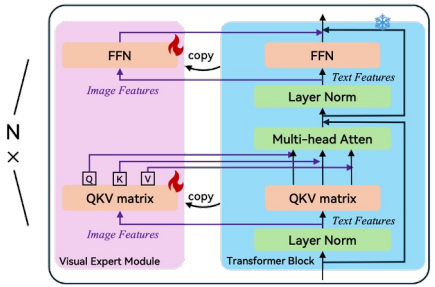


Figure 1. CogVLM [71] replicates an identical visual expert module alongside each transformer block to handle multimodal inputs.

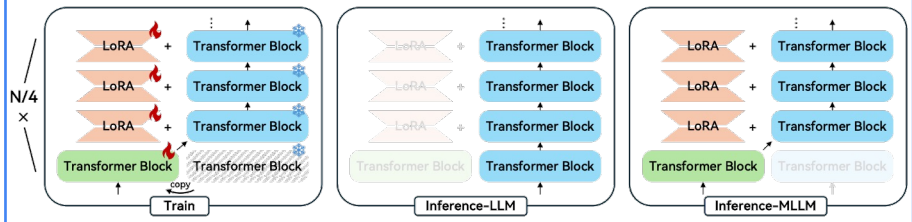
Observation 4: Compared to full fine-tuning the LLM, LoRA and CogVLM lead to a decrease in the multimodal performance of the trained MLLM.

Observation 5: For CogVLM, the addition of visual expert modules at every quarter interval of the layers results in the best MLLM performance.

Technical Contribution – Approach Analyses

香港中文大學
The Chinese University of Hong Kong

Approach Analyses – GenieBlue



Training: We replicate the transformer blocks at every quarter interval throughout the layers of the LLM while integrating LoRA modules into the remaining transformer blocks. During multimodal training, we freeze the original LLM, allowing ViT, the replicated transformer blocks, and the LoRA parameters to be fully trained.

Deployment: Non-shared base deployment strategy

- For pure-text inference, we utilize the original, unmodified LLM to perform all calculations.
- For multimodal inference, we replace the original blocks with the trained transformer blocks at every quarter interval and incorporate LoRA into the remaining transformer blocks.

Approach Analyses – GenieBlue

BlueLM-3B	#Param	AI2D	ChartQA	DocVQA	OCRBench	RealWorldQA	ScienceQA	TextVQA	AVG	Retention (%)
Full-Finetune	3161.26M	74.03	69.36	74.63	56.70	58.04	68.24	62.34	66.19	-
CogVLM-Skip	1005.69M	70.01	66.36	71.97	54.60	56.34	68.91	59.37	63.94	96.60
GenieBlue-Post	1005.73M	68.49	61.68	67.78	49.80	55.42	69.96	61.59	62.10	93.82
GenieBlue-Pre	1005.73M	72.90	66.20	71.11	46.50	58.30	73.20	60.03	64.03	96.74
GenieBlue-Skip	1005.73M	73.67	69.32	74.26	55.30	57.39	68.34	60.37	65.52	98.99
Qwen2.5-3B	**#Param**	**AI2D**	**ChartQA**	**DocVQA**	**OCRBench**	**RealWorldQA**	**ScienceQA**	**TextVQA**	**AVG**	**Retention (%)**
Full-Finetune	3527.81M	76.98	68.48	64.25	56.20	62.09	69.43	55.54	64.71	-
CogVLM-Skip	1146.75M	69.30	65.92	71.10	54.10	50.59	69.48	59.62	62.87	97.16
GenieBlue-Post	1146.79M	67.29	59.80	60.70	49.30	56.47	75.35	59.88	61.26	94.66
GenieBlue-Pre	1146.79M	69.01	58.44	56.65	43.90	58.04	75.01	62.19	60.46	93.44
GenieBlue-Skip	1146.79M	72.99	63.04	62.74	53.90	57.39	71.05	61.68	63.26	97.76

Table 5. Evaluation results on MLLM benchmarks after training with the 9M fine-tuning dataset. Similar to the experiment setting of CogVLM, we replicate transformer blocks at the last, first, and every interval quarter of layers. Results show that GenieBlue-Skip demonstrates the best MLLM performance, yielding over 97% retention in MLLM performance compared to full fine-tuning.

Observation 6: For GenieBlue structure, GenieBlue-Skip achieves the best multimodal performance, it also outperforms CogVLM-Skip.

BlueLM-3B	Shared Base	DROP	GPQA	GSM8K	MATH	MMLU	AlignBench	MT-bench	AVG	Retention (%)
BlueLM-3B	-	81.57	29.46	86.13	38.94	74.13	5.67	5.42	60.16	-
Full-Finetune	-	64.67	28.80	69.90	30.60	57.67	3.84	3.92	47.03	78.18
LoRA	✓	79.71	29.02	84.46	39.08	69.76	4.62	4.61	56.33	93.63
GenieBlue-Post	✓	78.64	28.13	85.37	37.08	70.77	4.51	4.65	55.94	92.98
GenieBlue-Pre	✓	76.95	29.24	74.98	35.66	65.26	4.61	4.71	53.61	89.12
GenieBlue-Skip	✓	75.36	29.02	76.27	38.16	67.78	4.66	4.76	54.40	90.42

Table 6. Comparison of pure language capabilities using the shared base versus non-shared base deployment strategies, trained with 9M fine-tuning data. The non-shared base approach can maintain the pure text capabilities of the original LLM. In the shared-base strategy, training with BlueLM-3B indicates that the fewer trainable parameters involved in multimodal training, the better the retention of pure text capabilities.

Observation 7: Deploying with the non-shared base strategy results in significantly better pure-text capabilities compared to the shared base strategy.

Technical Contribution – Training and Deployment Recipe

香港中文大學
The Chinese University of Hong Kong

Training Recipe

Approach: Adopt a two-stage training strategy.

- **Stage 1:** Pretrain the linear projection layer while keeping the ViT and LLM frozen.
- **Stage 2:** Finetune the entire model using a large-scale image-text paired dataset.

Deployment Recipe

- **Device:** iQOO 13 smartphone with the Qualcomm Snapdragon 8 Elite SoC.
- **SDK:** Qualcomm QNN SDK.
- **Precision:**
  - ✓ **ViT and projector layer:** W8A16
  - ✓ **LLM:** W4A16
  - ✓ **LoRA:** W8A16

Training Data

- **Stage 1:** Use a pretraining dataset of 2.5 million image-text pairs, including LLaVA, ShareGPT4V, and ALLaVA.
- **Stage 2:** Build a dataset of 645 million image-text pairs, comprising both open-source and internal datasets. This dataset covers a wide range of downstream tasks and diverse data types such as image captioning, visual question answering, text-image recognition, and pure text data.

Type	Public (M)	In-House (M)	In-House / Public
Pure Text	2.2	64.7	29.4
Caption	10.0	306.3	30.6
VQA	20.3	44.4	2.2
OCR	23.3	173.9	7.5
Total	**55.8**	**589.3**	**10.6**

Results - Strong LLM/MLLM Performance

香港中文大學
The Chinese University of Hong Kong

- **OpenCompass Benchmark**（By March 2025）：

Model	#Params	AVG	MMBench	MMStar	MMMU	MathVista	HallusionBench	AI2D	OCRBench	MMVet
BlueLM-V-3B [53]	3.2B	66.1	82.7	62.3	45.1	60.9	48.0	85.3	82.9	61.8
Ovis2-2B [52]	2.46B	65.2	76.9	56.7	45.6	64.1	50.2	82.7	87.3	58.3
Qwen2.5-VL-3B [7]	3.75B	64.5	76.8	56.3	51.2	61.2	46.6	81.4	82.8	60.0
SAIL-VL-2B [24]	2.1B	61.0	73.7	56.5	44.1	62.8	45.9	77.4	83.1	44.2
InternVL2.5-2B-MPO [72]	2B	60.9	70.7	54.9	44.6	53.4	40.7	75.1	83.8	64.2
GenieBlue	3.2(+0.55)B	64.2	78.2	59.4	47.6	58.0	46.3	83.1	82.9	58.1
InternVL2-8B [13]	8B	64.1	79.4	61.5	51.2	58.3	45.0	83.6	79.4	54.3

GenieBlue retains over 97% accuracy of BlueLM-V-3B while outperforming InternVL2-8B on average.

- **Pure-text Tasks**

	#Params	DROP	GPQA	GSM8K	MATH	MMLU	AlignBench	MT-bench	AVG	Retention (%)
BlueLM-3B	2.7B	81.57	29.46	86.13	38.94	74.13	5.67	5.42	60.16	-
GenieBlue	3.2(+0.55)B	81.57	29.46	86.13	38.94	74.13	5.67	5.42	60.16	100.00
Qwen2.5-3B	3.1B	70.82	30.30	74.75	61.74	66.31	6.00	5.81	60.29	-
Qwen2.5VL-3B	3.75B	72.72	24.24	70.43	58.92	65.07	5.38	4.72	56.05	92.98

GenieBlue retains 100% performance of the original LLM, whereas Qwen2.5VL-3B exhibits some degradation.

Results – High Deployment Efficiency

香港中文大學
The Chinese University of Hong Kong

- **GenieBlue vs. BlueLM-V-3B**:

Model	Context (token)	Load Time (s)	ViT Time (s)	Input Speed (token/s)	Output Speed (token/s)	Storage (GB)	Memory (GB)
BlueLM-V-3B	2048	0.51	0.4	1515.15	33.00	1.77	1.73
GenieBlue	2048	0.80	0.4	1666.67	31.00	1.92	2.10

With the inclusion of additional LoRA parameters, GenieBlue incurs longer model loading times, slightly larger storage and memory requirements, and a marginally slower token output speed. However, a token output speed of 30 token/s is fully sufficient for daily use on mobile devices.

Thanks for your listening!