香港中文大學
The Chinese University of Hong Kong

Algorithm and System Co-Design for Multimodal Large Language Models on Mobile Devices

Xudong Lu
Multimedia Laboratory
The Chinese University of Hong Kong
Email: luxudong@link.cuhk.edu.hk

Outline

- Algorithm-System Co-Optimization Empowering Edge AI
- BlueLM-V-3B Technical Contribution
  - Algorithm and System Co-Design
  - Training Recipe
- Demos and Empirical Validation
  - State-of-the-art Performance in Multimodal Understanding
  - Optimized Inference for Edge Deployment
- Widespread Dissemination
- Future Work

Algorithm-System Co-Optimization Empowering Edge AI

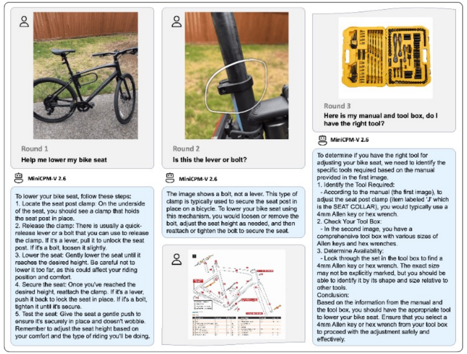
香港中文大學 The Chinese University of Hong Kong

Motivation As "intelligent companion", The smartphone is the most ideal deployment platform for MLLMs.

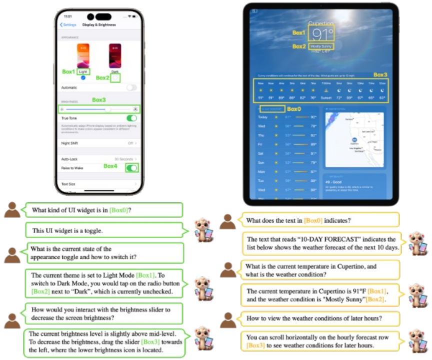
Challenge Deploying MLLMs on smartphones is constrained by limited memory and computational power, making user satisfactory heavily dependent on parameter number and deployment efficiency.

Solution We introduce **BlueLM-V-3B**, an MLLM specifically tailored for mobile platforms:

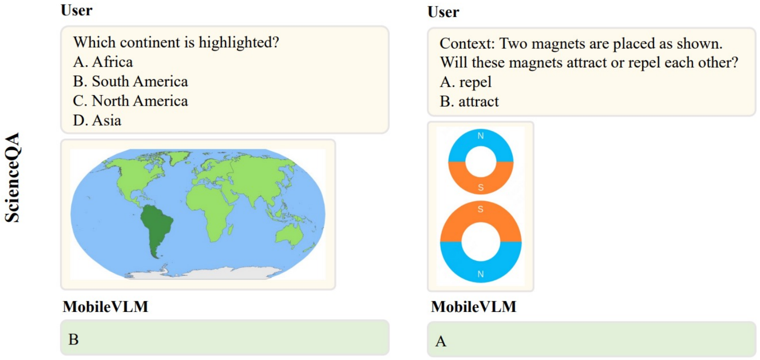
- Limited memory: Train and deploy a 3B-parameter compact model.
- Limited computational power: Design algorithm-system co-optimization framework.



MiniCPM-V



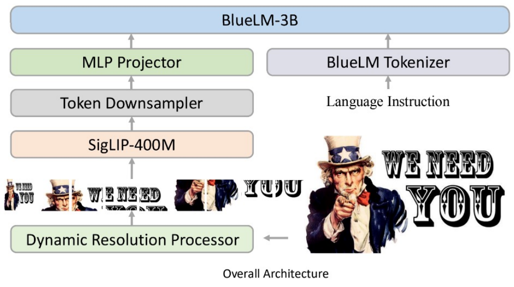
Ferret-UI 2



MobileVLM

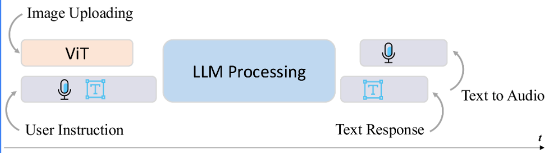
Technical Contribution - Summary

香港中文大學
The Chinese University of Hong Kong



Model Architecture

- ViT: SigLIP-400M
- MLP Projection Layer
- LLM: BlueLM-3B
- Dynamic Resolution Module
- Token Down-Sampling Module



Deployment Strategy

- Batched Image Patch Encoding
- Pipeline Parallelism in Image Encoding
- Chunked Processing of Input Tokens
- Mixed-Precision Deployment

1) Algorithm and System Initiative:
We identify and address the excessive image enlargement issue in the dynamic resolution scheme used by classical MLLMs. Additionally, we implement a series of system designs and optimizations for hardware-aware deployment.

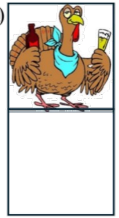
2) State-of-the-art MLLM Performance:
BlueLM-V-3B achieves SOTA performance (e.g., 66.1 on the OpenCompass benchmark) among models with similar parameter sizes, surpassing a series of MLLMs with much more parameters (e.g., MiniCPM-V-2.6, InternVL2-8B).

3) High Deployment Efficiency:
BlueLM-V-3B is highly efficient when deployed on mobile phones. Take the MediaTek Dimensity 9300 processor as an example, with a memory requirement of just 2.2GB, it can encode images with a resolution of 768×1536 in approximately 2.1 seconds and achieves a token throughput speed of 24.4 token/s.

Technical Contribution - Algorithm and System Co-Design

Dynamic Image Resolution - Algorithm Design

Problem: Traditional dynamic resolution approaches lead to exaggerated image upscaling.

(A)


(B)


In Figure A, LLaVA-NeXT selects a resolution of 384×768 for an original image sized 380×393. In Figure B, InternVL 1.5 chooses a resolution of 1920×384 for an original image sized 500×102.

Solution: BlueLM-V-3B implements a relaxed aspect ratio matching algorithm.

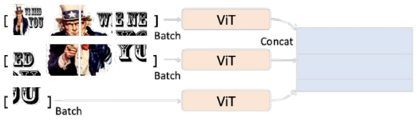
Algorithm 1 Relaxed Aspect Ratio Matching

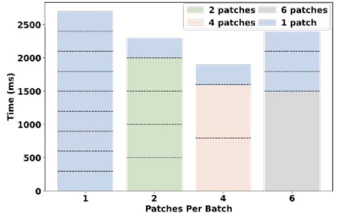
function RELAXED_ASPECT_RATIO_MATCHING(original_size: $(W_{\text{orig}}, H_{\text{orig}})$, possible_ratios: List of (m, n))
  Initialize: best_fit ← None, $R_{e,\max} \leftarrow 0$, $R_{w,\min} \leftarrow \infty$
  $(W_{\text{orig}}, H_{\text{orig}}) \leftarrow$ original_size
  for each (m, n) in possible_ratios **do**
    $(W, H) \leftarrow (384 \times m, 384 \times n)$
    scale $\leftarrow \min\left(\frac{W}{W_{\text{orig}}}, \frac{H}{H_{\text{orig}}}\right)$
    $\delta W \leftarrow \text{int}(W_{\text{orig}} \cdot \text{scale})$
    $\delta H \leftarrow \text{int}(H_{\text{orig}} \cdot \text{scale})$
    $R_e \leftarrow \min(\delta W \cdot \delta H, W_{\text{orig}} \cdot H_{\text{orig}})$
    $R_w \leftarrow W \cdot H - R_e$
    if $(R_e - R_{e,\max}) > \alpha \cdot R_{e,\max}$ **or** $((R_{e,\max} - R_e) < \alpha \cdot R_{e,\max}$ **and** $R_w < R_{w,\min})$ **then**
      $R_{e,\max} \leftarrow R_e$
      $R_{w,\min} \leftarrow R_w$
      best_fit ← (m, n)
    end if
  end for
  return best_fit
end function

Dynamic Image Resolution - System Design

Problem: Serial image encoding schemes fail to fully utilize NPU performance.

Solution: Design multi-patch parallel encoding on the NPU.

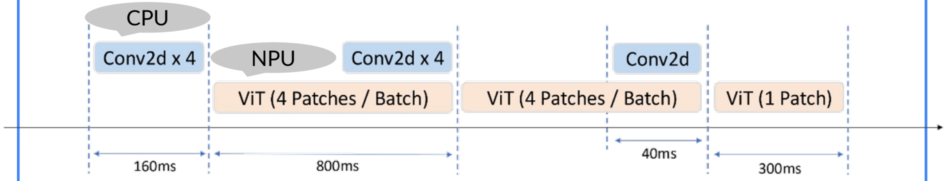




Dynamic Image Resolution - System Design

Problem: Single-threaded encoding causes CPU and NPU to wait on each other.

Solution: Design a pipeline parallel encoding scheme.



Technical Contribution - Algorithm and System Co-Design

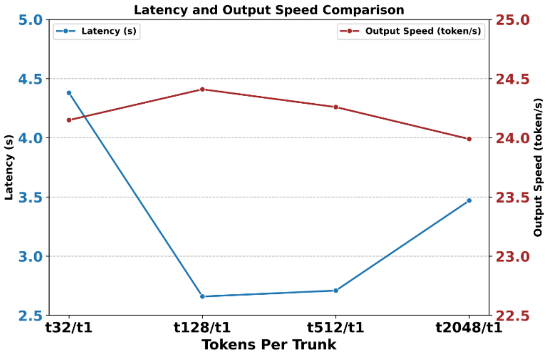
香港中文大學 The Chinese University of Hong Kong

Token Down-Sampling - Algorithm and System Design

Problem: Dynamic resolution results in very long input tokens.

Solution:

- Algorithm: Merge every 2×2 image tokens into a single token and use a linear layer for information fusion.
- System: Process 128 input tokens (t128) in parallel per iteration, then merge the results.



t{x}/t1 implies processing x input tokens in parallel

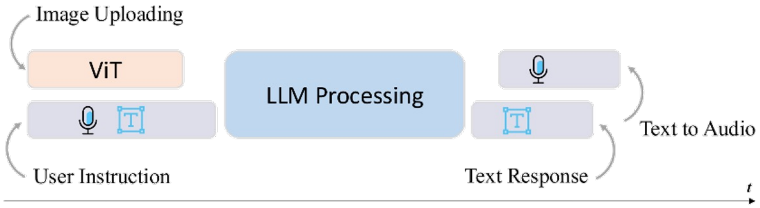
System Deployment – Overall Design

Solution 1: Mixed-Precision Parameter Deployment

- Weights: Use INT8 for SigLIP and MLP linear projection layers, and INT4 for the LLM.
- Activations: Use FP16 for SigLIP and MLP activations, INT16 for LLM activations, and INT8 for KV cache.

Solution 2: Decoupling Image Processing and User Input

- Once the user uploads an image, SigLIP starts processing it, while the user can simultaneously input commands.
- After image processing completes, the user's commands are forwarded to the LLM to generate responses.



Technical Contribution – Training Recipe

香港中文大學 The Chinese University of Hong Kong

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.

Training Data:

- **Stage 1:** Use a comprehensive 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 - State-of-the-art MLLM Performance

香港中文大學 The Chinese University of Hong Kong

- **OpenCompass Benchmark** （By December 2024, ≤ 10B params）：

Model	Params	Avg.	MMBench	MMStar	MMMU	MathVista	HallusionBench	AI2D	OCRBench	MMVet
Qwen2-VL [125]	8B	**67**	81	60.7	**53.7**	61.4	**50.4**	83	843	**61.8**
MiniCPM-V-2.6 [134]	8B	65.2	78	57.5	49.8	60.6	48.1	82.1	**852**	60
InternVL2 [22]	8B	64.1	79.4	61.5	51.2	58.3	45	83.6	794	54.3
POINTS-Qwen2.5 [74]	8.3B	62.5	78	60.9	51.4	**63**	45.6	81.2	717	47.9
BlueLM-V (Ours)	**3B**	66.1	**82.7**	**62.3**	45.1	60.8	48	**85.3**	829	**61.8**

BlueLM-V-3B achieves the highest scores in four tasks and ranks second on average, demonstrating strong MLLM performance.

- **Text-centric/OCR Capacity**

Model	Params	$\text{TextVQA}_{\text{val}}$	$\text{DocVQA}_{\text{test}}$	MTVQA
Phi-3-Vision [2]	4.2B	72.4	84.6	13.9
MiniCPM-V-2 [134]	2.8B	73.2	71.9	9.3
InternVL2 [22]	4B	74.7	89.2	15.5
Qwen2-VL [125]	**2B**	**79.9**	**90.1**	20.7
BlueLM-V (Ours)	3B	78.4	87.8	**32.7**

In OCR-related tasks, BlueLM-V-3B achieves highly competitive results and significantly outperforms mainstream multimodal models in multilingual evaluations.

Results - High Edge Deployment Efficiency

香港中文大學
The Chinese University of Hong Kong

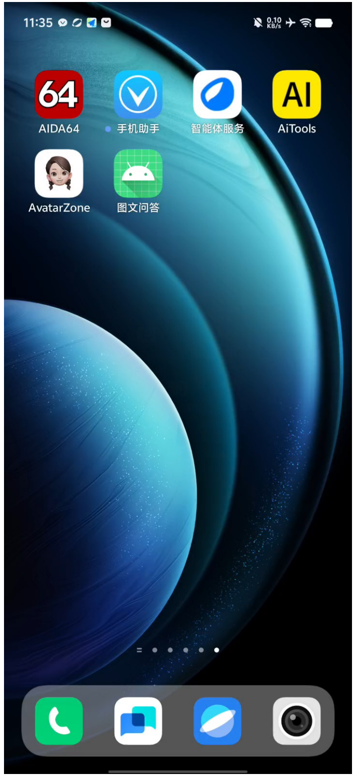
- **Time Consumption of Each Component (MediaTek Dimensity 9300)**

		MLLM Prefilling Time (s)			Output Speed (token/s)
Image Resolution	Patch Number	Init (s)	ViT (s)	LLM Prefilling (s)	
384×768	3	0.47	0.8	0.80	≈ 20token/s
768×768	5		1.1	1.28	
768×1152	7		1.6	1.76	
768×1536	9		1.9	2.70	
768×1920	11		2.4	4.42	

- **Comparison with MiniCPM-V**

Model Name	Params	Processor	Solution	Image Processing	LLM Prefilling	Throughput
MiniCPM-V 2.5 [134]	8B	MediaTek Dimensity 9300	CPU (llama.cpp) ☹	4.0s	13.9s	4.9 token/s
BlueLM-V-3B (Ours)	3B	MediaTek Dimensity 9300	NPU ☺	**2.53s** (0.47+2.06)	**2.7s**	**24.4** token/s

BlueLM-V-3B, with its smaller parameter size and algorithm-system co-design, demonstrates advantages in latency and token throughput.



Widespread Dissemination

香港中文大學
The Chinese University of Hong Kong

• CVPR 2025

BlueLM-V-3B: Algorithm and System Co-Design for Multimodal Large Language Models on Mobile Devices

Xudong Lu[*2,1†], Yinghao Chen[*1], Cheng Chen[*1], Hui Tan[*1], Boheng Chen[1], Yina Xie[1], Rui Hu[1], Guanxin Tan[1], Renshou Wu[1], Yan Hu[1], Yi Zeng[1], Lei Wu[1], Liuyang Bian[1], Zhaoxiong Wang[1], Long Liu[1], Yanzhou Yang[1], Han Xiao[2,1†], Aojun Zhou[2], Yafei Wen[1], Xiaoxin Chen[1], Shuai Ren[1‡✉], Hongsheng Li[2✉]
[1]vivo AI Lab [2]CUHK MMLab
{luxudong@link,hsli@ee}.cuhk.edu.hk
shuai.ren@vivo.com

https://arxiv.org/pdf/2411.10640

• Synced China

算法系统协同优化，vivo与港中文推出BlueLM-V-3B，手机秒变多模态AI专家

机器之心 2024年11月29日 14:15 北京

机器之心 | AIxiv

> AIxiv专栏是机器之心发布学术、技术内容的栏目。过去数年，机器之心AIxiv专栏接收报道了2000多篇内容，覆盖全球各大高校与企业的顶级实验室，有效促进了学术交流与传播。如果您有优秀的工作想要分享，欢迎投稿或者联系报道。
> 投稿邮箱：liyazhou@jiqizhixin.com；zhaoyunfeng@jiqizhixin.com

BlueLM-V-3B 是一款由 vivo AI 研究院与香港中文大学联合研发的端侧多模态模型。该模型现已完成对天玑 9300 和 9400 芯片的初步适配，未来将逐步推出手机端应用，为用户带来更智能、更便捷的体验。

https://www.jiqizhixin.com/articles/2024-11-29-5

• Other Media Outlets

澎湃 要闻 深度 直播 视频 时事 更多 下载客户端

vivo 提出 BlueLM-V-3B：在移动平台上高效部署 MLLM

多模态大语言模型（MLLM）的出现和日益普及，在改善交流、促进学习和解决问题等日常生活的各个方面都具有巨大的潜力。手机作为人们日常生活中必不可少的伴侣，是最有效、最方便的 MLLM 部署平台，可将其无缝集成到日常任务中。然而，由于内存大小和计算能力的限制，在手机上部署 MLLM 面临着挑战，如果不进行大量优化，就很难实现流畅的实时处理。

在这项工作中，来自 vivo AI Lab 和香港中文大学 MMLab 的研究团队提出了一种算法和系统协同设计方法——BlueLM-V-3B，专门用于在移动平台上高效部署 MLLM。具体来说，他们重新设计了主流 MLLM 采用的动态解析方案，并针对硬件感知部署实施了系统优化，以优化手机上的模型推理。

https://www.thepaper.cn/newsDetail_forward_29401005

把大象放冰箱！算法与系统协同优化，vivo与港中文推出BlueLM-V-3B，手机秒变多模态AI专家

我爱计算机视觉 2024年11月29日 23:30 江苏

关注公众号，发现CV技术之美

我爱计算机视觉
探寻CV新知，发现AI价值
705篇原创内容
公众号

本文转载自机器之心。

BlueLM-V-3B 是一款由 vivo AI 研究院与香港中文大学联合研发的端侧多模态模型。该模型现已完成对天玑 9300 和 9400 芯片的初步适配，未来将逐步推出手机端应用，为用户带来更智能、更便捷的体验。

https://mp.weixin.qq.com/s/QHGG98mQ0H-iaWzz2Xndig

端侧多模态 | vivo手机端侧多模态大模型技术解读

Original 卖热干面的小女孩 小窗幽记机器学习 2024年11月27日 07:25 广东

0. 引言

随着多模态大型语言模型的快速发展，如何在移动设备上高效部署这些模型成为关键挑战。Vivo提出BlueLM-V-3B，通过算法与系统协同设计，实现了高性能的移动端部署方案。

1. 简介

Vivo提出的**BlueLM-V-3B**是一种专门为移动设备（如手机）优化的多模态大型语言模型（MLLM）。通过算法和系统的协同设计，从模型小型化、推理速度优化和高效性能提升等角度，成功将 BlueLM-V-3B 部署到移动平台上。BlueLM-V-3B 在具有约 3B 参数规模的模型中实现了优异的性能表现，同时在手机端实现了高效的实时推理。

https://mp.weixin.qq.com/s/jrD8YwmNlzkS2QJfkmln0A

论文解读系列：BlueLM-V-3B:面向移动设备的多模态大语言模型的算法与系统协同优化

Original chal1ce 橙陈编程笔记 2024年11月20日 08:00 广东

蓝厂最近在移动端的设备推出了个新的模型 BlueLM-V-3B，据说是一个多模态模型，他们的主要研究点在于对大语言模型的算法与系统协同优化，也因为这个推出了这个移动端的小参数模型，优化了内存、处理技术等等，其实就是针对硬件做优化，让大模型能更适应蓝厂的手机。这是他们的论文报告：

> BlueLM-V-3B: Algorithm and System Co-Design for Multimodal Large Language Models on Mobile Devices
>
> 链接：https://arxiv.org/pdf/2411.10640

https://mp.weixin.qq.com/s/8vsGPoOFQ0MbbW5Dp3VBOQ

Future Work

香港中文大學
The Chinese University of Hong Kong

- **Post-training Optimization for Edge Deployment**

  Conduct instruction tuning, RLHF, and QAT on the base multimodal model to produce a higher-performance, more user-friendly 3B model optimized for on-device deployment.

- **Integration with Mobile Agents**

  Enhance the model's planning and tool-usage capabilities by integrating with mobile agent technologies, enabling a more intelligent and responsive on-device assistant.

- **Exploration of Model Size Limits**

  Leverage techniques such as pretraining and knowledge distillation to explore parameter efficiency, targeting 1B and 0.5B models with performance comparable to 3B and 7B models.

- **Modular Multimodal Model via LoRA Training**

  Investigate LoRA-based training to decouple the language and vision components, aiming to develop edge models that combine strong language understanding with advanced multimodal capabilities.

Thanks for your listening!