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

Supplementary Material

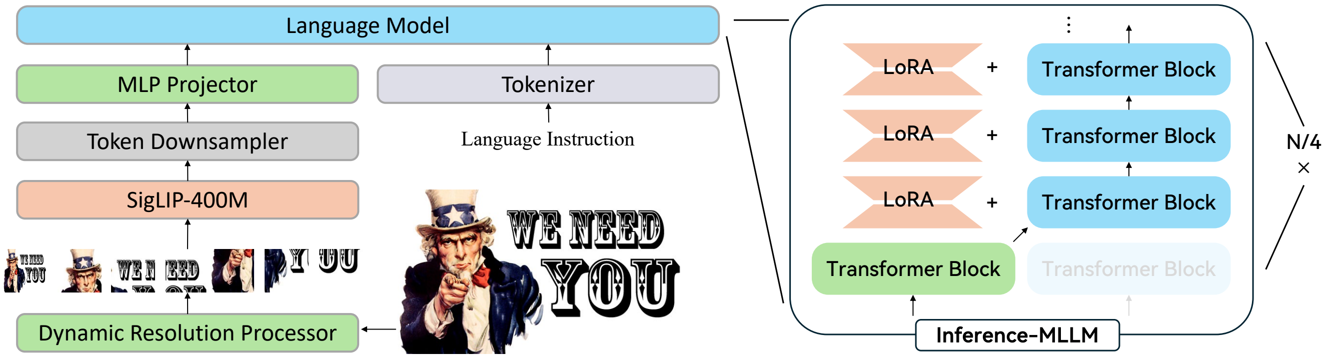

Figure 1. Structure detail of GenieBlue during the MLLM inference process.

1. Structure Details of GenieBlue

We here provide the detailed structure of GenieBlue during the MLLM inference process based on the BlueLM-V-3B [3] architecture (Fig. 1). BlueLM-V-3B is modified from the classical LLaVA approach [2], incorporating a redesigned dynamic resolution processor and a token downsampler [1] to optimize for better on-device deployment. GenieBlue further focuses on the structural design of the transformer blocks within the language model.

2. Training Data Composition

We here provide the data composition of Cambrian-7M [4]. It has already included approximately 1.5M pure text training samples. The data composition of the 645M fine-tuning data for GenieBlue can be found in [3].

Type	**OCR**	**General**	**Language**	**Counting**	**Code**	**Math**	**Science**
Ratio (%)	27.22	34.52	21.00	8.71	0.87	7.20	0.88

Table 1. Data composition of the Cambrian-7M [4] fine-tuning dataset (with approximately 1.5M pure-text data).

3. More Discussions

GenieBlue is a plug-and-play training approach that efficiently decouples multimodal training parameters from the original language model. This design allows GenieBlue to achieve good multimodal performance without compromising the language model's performance. In addition, this structural design requires minimal hardware-side adaptation and reduces the engineering difficulty during practical end-side deployment, making it a relatively reasonable approach at the current stage. In the future, we will validate the feasibility of GenieBlue on a wider range of SoC platforms.

4. Latency on More Devices

We provide the performance of GenieBlue vs. BlueLM-V-3B on the NPU of MediaTek 9400.

	Load (s)	**Vit (s)**	**Input (token/s)**	**Output (token/s)**
BlueLM-V-3B	0.6	0.3	1242.7	28.4
GenieBlue	0.84	0.3	1229.8	27.6

5. Hyper Parameters

Here, we provide the hyper-parameters used in the pre-training and fine-tuning stage of the final GenieBlue model. We use the same 2.5M pre-training data and 645M fine-tuning data as in BlueLM-V-3B [3].

5.1. Pre-training Stage

Configuration	Stage 1
LLM Sequence Length	4096
Dynamic Resolution	None (384×384)
Optimizer	AdamW
Optimizer Hyperparams	$\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$
Peak LR	10^{-3}
LR Schedule	Cosine Decay
Weight Decay	0.05
Training Steps	3.434k
Warm-up Steps	34
Global Batch Size	720
Gradient Accumulation	1
Numerical Precision	`bfloat16`

Table 2. Hyper-parameters for the pre-training stage (stage 1) of GenieBlue with 2.5M training samples.

5.2. Fine-tuning Stage

In the process of fine-tuning, to enhance the speed of training, we concatenate training samples to achieve a sequence length of 4096.

Configuration	Stage 2
LLM Sequence Length	4096
Dynamic Resolution	Up to 16 patches (1536×1536)
Optimizer	AdamW
Optimizer Hyperparams	β_1 = 0.9, β_2 = 0.98, ϵ = 10^{-6}
Peak LR	10^{-4}
LR Schedule	Cosine Decay
Weight Decay	0.05
ViT Layer-wise LR Decay	0.9
Training Steps	53k
Warm-up Steps	530
Global Batch Size	6800
Gradient Accumulation	10
Numerical Precision	`bfloat16`

Table 3. Hyper-parameters for the fine-tuning stage (stage 2) of GenieBlue with 645M training samples.

References

[1] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 1

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 1

[3] Xudong Lu, Yinghao Chen, Cheng Chen, Hui Tan, Boheng Chen, Yina Xie, Rui Hu, Guanxin Tan, Renshou Wu, Yan Hu, et al. Bluelm-v-3b: Algorithm and system co-design for multimodal large language models on mobile devices. *arXiv preprint arXiv:2411.10640*, 2024. 1, 2

[4] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1