Exploring the Application Boundaries of Lightweight Multimodal Models

Rui Wang[a]
Faculty Information Science & Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

Keywords: Lightweight Model, Boundary, Artificial Intelligence.

Abstract: With the rapid development of Multimodal Large Language Models (MLLMs), models have shown unprecedented potential in graph comprehension and generation tasks. However, current research mostly focuses on the performance breakthroughs of giga-parameter-level models, and lacks systematic exploration of the capability boundaries of lightweight models (with parameters less than or equal to 3B) in real-world application scenarios. This paper systematically explores the performance boundaries of the lightweight MLLM in different application scenarios based on the Bunny-v1_0-2B-zh open source model. By designing a multi-dimensional evaluation framework covering core capabilities such as text comprehension, visual reasoning, and revealing the practical application potential of the 2 billion parametric scale model. Experiments show that the model can achieve excellent performance in resource-constrained scenarios such as mobile device deployment, but has shortcomings in complex logical reasoning and specialized domain knowledge. This study provides empirical evidence for the selection of applicable scenarios and optimization direction of the lightweight model.

1 INTRODUCTION

As multimodal large language models (MLLMs) show breakthrough performance in tasks such as text generation and visual reasoning, their high computational cost and deployment threshold limit practical application scenarios. Lightweight models provide new possibilities for resource-constrained scenarios such as edge computing and mobile AI through architectural optimization and knowledge distillation techniques, which dramatically reduce the computational requirements while maintaining the basic capabilities (Brown et al., 2020). This study is conducted based on the Bunny-v1_0-2B-zh model, which was developed by the Beijing Academy of Artificial Intelligence (BAAI), and is a lightweight model for Chinese natural language processing tasks (He et al., 2024), which is very suitable for lightweight modeling research.

Existing research focuses on the model compression technology itself, and lacks systematic exploration of the capability boundaries of lightweight models, which often leads to the problem of "model selection mismatch" in practical applications. Currently, lightweight MLLM research is mainly promoted along two paths: based on architectural streamlining methods, such as Meta's LLaMA-7B, which reduces computational complexity through the sparse attention mechanism (Yuan et al., 2025), and achieves 83% of the performance of a 13B model in English unimodal tasks (Touvron et al., 2023); based on the multi-modal adaptation method, Google's MobileViT will be visualized to be a lightweight transformer, which is a lightweight model with the ability of the Transformer. Transformer lightweight and then interfaces it with language models to achieve real-time inference on mobile with an 8% decrease in ImageNet accuracy (Mehta & Rastegari, 2022). However, there are significant limitations in the existing work, for example, most models are only designed for unimodal or bilingual environments and lack specialized optimization for Chinese multimodal scenarios (Liu et al., 2023). The evaluation system is biased towards traditional accuracy metrics and Ignores the need for joint optimization of memory, latency, and energy consumption in real-world deployments (Chen et al., 2023).

[a] https://orcid.org/0009-0000-9013-0586

Wang, R.
Exploring the Application Boundaries of Lightweight Multimodal Models.
DOI: 10.5220/0013699600004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 476-480
ISBN: 978-989-758-765-8

This study aims to explore the application boundaries of the lightweight model, and focuses on the performance exploration when the model is actually deployed. By designing gradient complexity test tasks (Zhang et al., 2024), the performance degradation pattern of the model in core tasks such as text comprehension and visual reasoning is systematically quantified, At the same time, the time spent on each output is collected to visualize the model performance.

Through this study, it can provide a reference for the optimization of lightweight models and the deployment of actual scenarios, which is of practical value for promoting the landing of edge intelligence.

2 RESEARCH OBJECTIVES

This study needs to evaluate the overall performance of the lightweight multimodal language model Bunny-v1.0-2B-zh, focusing on its performance in text generation and image understanding tasks. Through experiments, the applicability and capability of the model in different tasks and scenarios are explored to provide a clear direction for subsequent optimization. Meanwhile, this study hopes to verify whether the lightweight model can achieve more accurate multimodal understanding and generation while maintaining efficient computational performance.

Specific tasks include: comprehensively evaluating the model performance and clarifying the difference between the model performance under standard tasks and extreme scenarios. Then, identify bottlenecks and deficiencies, and identify shortcomings in the current model design by exploring the limitations under different parameters and task conditions. Quantitative and qualitative analyses are conducted through a series of experiments to analyze model performance across multiple tasks, including consistency, accuracy, and variety of text generation, as well as precision and semantic consistency of image understanding. Key parameters affecting the quality of the model output, such as generation length and repetition penalties, are also identified. Finally, an application potential exploration is conducted to validate the model's ability to operate in practice in low-resource computing devices. Its adaptability in multimodal combinatorial tasks is tested and its application scenarios in real-world deployments are explored.

3 EXPERIMENTAL

3.1 Experimental Objectives

To test the generation quality of the model for different lengths of text output. To investigate the effect of the repetition penalty mechanism on the generated content. Evaluate the model's ability to understand image content and combine it with text generation. Test the model's performance in a specific domain. To test the model's ability to handle non-canonical inputs such as spelling errors, incomplete inputs.

3.2 Introduction to the Model

This experiment uses Bunny-v1.0-2B-zh, a model based on the bunny architecture built with the SigLIP visual coder and the Qwen1.5-1.8B language backbone, which accepts high-resolution images up to 1152x1152. The compression of large-scale datasets into high-quality core sets through data condensation (Wu et al., 2024) techniques plays a key role in model lightweight.

3.3 Experimental Design

3.3.1 Experiment Prep

The experiment uses Graphics Processing Unit (GPU) to run the model, in order to reflect the advantages of the lightweight model, and to simulate the more demanding model deployment scenarios, the mobile GPU RTX 4080 Laptop is used in this experiment.

The image input uses the example_2 provided by BAAI as the input image for most of the experiments, which avoids errors caused by unclear images and blind zones in model understanding. Meanwhile, the use of officially provided images facilitates research and comparison by other researchers. The images are shown in Figure 1.

Figure 1: Official image example_2. (Beijing Zhiyuan Artificial Intelligence Research Institute, 2024)

For BERT scoring in later experiments, a manual text description of the image is needed, which is described as follows: "A black and white cat poses with wide-open eyes and wide-open mouth, revealing a shocked look, with a yellow 'HUH?' in the middle of the image, a humorous Internet emoticon, the cat sits in the corner of a painted wall, and the wall in the lower left of the image has peeling, revealing a gray interior."

In order to test the effect of different parameters on the model-generated content, the input text "Please describe the content of this image and reply in Chinese" was fixed, and then the length restriction experiment and the repetition penalty parameter experiment were conducted respectively.

3.3.2 Text Length Limitation Experiment

The text length limitation parameter is increased from 10 to 500, 10 words are added each time, and each output and output time are recorded. The data with large differences and reference values are selected, and finally the relevance score and accuracy score are calculated.

Below is the formula for relevance score:

$$S = \left(\frac{a + 0.5b}{T} \times 4\right) + (c \times 0.2) \quad (1)$$

The element a is the number of element hits, b is the number of partial descriptions, T is the total number of elements, and c is the coherence coefficient.

Number of element hits: the number of elements that exactly match the standard answer (maximum 5), Total number of elements: fixed to 5 (black and white fur color / HUH text / cat / emoji attributes / spatial relations), Coherence factor: paragraph articulation score based on BERT-score (0-5).

Accuracy scoring formula:

$$A = 5 - \sum_{i=1}^{n} w_i \cdot e_i \quad (2)$$

wi: error type weight (subject error w1=3.0, fictional element w2=1.0, attribute contradiction w3=0.5). ei: number of occurrences of the corresponding error type. Base score: 5 points, with a minimum of 0 points after deductions.

3.3.3 Repeat Penalty Experiment

The penalty parameter is increased from 0.5 to 2.5 by 0.1 each time, other operations are the same as the text length limitation experiment, and finally the repetition rate, content accuracy and semantic fluency are calculated.

Repetition rate formula:

$$R = \frac{r - a}{T - a} \quad (3)$$

r: repetition count, a: allowed repetitions, T: total token number.

The accuracy score formula is the same as the text length limitation experiment.

Fluency scoring formula:

$$F = \text{BERTScore} \times 5 \quad (4)$$

3.4 Analysis of Experimental Results

By filtering and scoring, the table of text length limitation experiment and the table of repetition penalty experiment were drawn.

Table 1: Text Length Limit Experimental Data.

Word limit	Relevance score (1-5)	Accuracy score (0-5)	Time
10 words	1.2	2.0	2m 34.9s
30 words	1.2	2.0	2m 33.7s
89 words	1.2	2.0	2m 33.5s
90 words	3.8	5.0	2m 34.5s
120 words	4.1	4.5	3m 13.1s
200 words	4.3	4.0	3m 30.3s
500 words	4.0	4.0	3m 32.4s
Unlimited	4.5	3.5	3m 24.2s

As shown in Table 1, the outputs under the limit of 10 to 89 words are exactly the same and of poor quality, which may be the drawback brought by the forced restriction of the number of text output by the model. After the length limit reaches 90, the output quality returns to normal, and the output quality gradually decreases when the length limit continues to be increased. The reason for this phenomenon is that the model deliberately matches the text length limit, and too short a text limit leads to insufficient generation steps to complete multiple rounds of reasoning; too long a text limit leads to fictitious elements in the model, which are prone to produce erroneous details.

Compare the unrestricted output with the model's default parameters. Its output is strongly correlated with the image, but the accuracy is slightly worse compared to the output with the restricted over-length parameter, which is exactly the opposite when the parameter is 90. In summary, the best performance parameter is 120, below 90 is not recommended, from 90 to unlimited each has its own advantages, can be adjusted according to the actual situation.

Table 2: Repeat Penalty Experiment Data.

Penalty coefficient	Repetition rate (0-1)	Accuracy score (0-5)	Semantic fluency (1-5)	Time
1.0	0.62	3.5	2.4	1m 53.3s
1.2	0.24	4.0	3.9	2m 24.2s
1.5	0.18	3.0	3.2	2m 35.6s
1.7	0.15	2.0	3.0	2m 42.3s
2.0	0.15	2.0	3.0	3m 7.6s
2.5	0.15	2.0	3.0	2m 41.2s

After the experiment, the model of the repetition penalty parameter is lower than 1.0 repeat output meaningless content, from 1.0 normal output content, parameter 1.2 to reach the best performance, parameter 1.5 when the performance of the decline, the repetition is reduced; parameter 1.7 when the performance continues to decline, the repetition continues to decline, the output is unchanged after the output is greater than 1.7, it is an invalid modification.

By observing Table 2, it is concluded that the model's repetition penalty boundary is from 1.0 to 1.7, and the comprehensive performance is best at 1.2, which is suitable as the default parameter; when it is greater than 1.2, the model's content becomes dispersed in order to avoid repetition, and the accuracy of the description of the picture decreases, but its creativity is richer, and it is suitable for dealing with the creative task, therefore, the repetition penalty parameter is recommended to be set between 1.2 and 1.5 .

3.5 Advanced Experiment

3.5.1 Special Domain Adaptation Experiment

Following the optimal parameters derived from the first two experiments, a special domain adaptation experiment can be done. In this paper, take the medical field as an example, input an X-ray chest photo. The photo is shown in Figure 2.

Figure 2: X-ray Chest Picture. (Youlai Doctor, 2024)

Ask the question "Does this patient have steel nails in his chest, reply in Chinese." Under the repetition penalty parameter 1.2, the model recognizes a steel nail in the patient's chest and adds that it may be a tool for fixing the chest wall during treatment or surgery; under parameter 1.5, the previous output remains unchanged, but the sentence "X-ray shows L-1023" is fictionalized at the end. If the model is not prompted, but is simply asked, "Do you know what is wrong with this patient?", the model assumes that the patient has small nodules and masses in the lungs.

This experiment illustrates a number of points. One is that the repetition penalty parameter 1.5 would fudge the content and is not suited to the rigors of the medical field; the other is that the model's image parsing is so powerful that it is able to identify foreign objects in the chest cavity on monochrome X-rays, even though it requires some prompting to identifythem as steel nails, which only shows that the model has not been trained in the medical field. What is certain is that lightweight models have unlimited potential for future deployment in various domains. In fact, a researcher has already trained a lightweight model for the gaming domain, called VideoGameBunny (VGB), based on the Bunny model architecture, using game screenshots as a training set, and its performance is also excellent (Taesiri & Bezemer, 2024).

3.5.2 Abnormal Input Robustness

This final experiment challenges the model's Chinese language processing capabilities, with anomalous input consisting of missing key prepositions, homophone substitutions, mixing of non-literal symbols, mixing of Chinese and English, and extremely long single sentences without punctuation. The model is able to handle most of the abnormal inputs correctly, but only outputs English when English is used to request Chinese responses. When the model is not prompted to reply in Chinese, the model is likely to reply in English, which indicates that although the model has been optimized for Chinese, English is still a higher priority.

4 CONCLUSIONS

Based on BAAI's Bunny-v1_0-2B-zh model, this paper explores the application boundaries of this lightweight model through stepwise parameter tuning. This experiment has been designed in a targeted way, without using public datasets for large-scale testing, and the experimental data may have some errors; the experiments in this paper are only tested using the Bunny-v1_0-2B-zh model, which cannot represent other lightweight models, and the experimental results are for reference only. As a model with only 2 billion parameters, the Bunny-v1_0-2B-zh model is smaller than general lightweight models, but it shows amazing performance. It not only outputs results in an average of 2 minutes and 30 seconds on mobile devices, but is also able to show image recognition capabilities that are not inferior to those of larger models, which gives it application potential in a wide range of domains , and bodes very well for the development of lightweight models. This paper provides a reference for the optimization and deployment of lightweight models in the future, and contributes to the promotion of lightweight development of AI.

REFERENCES

Beijing Zhiyuan Artificial Intelligence Research Institute. 2024. Bunny-v1_0-2B-zh. Retrieved April 1, 2025, from https://huggingface.co/BAAI/Bunny-v1_0-2B-zh/tree/main/images

Brown, T. B., Mann, B., Ryder, N., et al. 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chen, W., Li, M., Zhou, Y., et al. 2023. EcoMLM: A holistic efficiency evaluation framework for lightweight multimodal models. In Proceedings of the 2023 ACM International Conference on Mobile Systems, Applications, and Services (MobiSys) (pp. 1–15).

He, M., Liu, Y., Wu, B., et al. 2024. Efficient multimodal learning from data-centric perspective. arXiv preprint arXiv:2402.11530.

Liu, Y., Chen, Z., Wang, Y., et al. 2023. CPM-M3: Multilingual multimodal pre-training with curriculum learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 10234–10248).

Mehta, S., & Rastegari, M. 2022. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In Proceedings of the 10th International Conference on Learning Representations (ICLR).

Taesiri, M. R., & Bezemer, C.-P. 2024. VideoGameBunny: Towards vision assistants for video games. arXiv preprint arXiv:2407.15295.

Touvron, H., Lavril, T., Izacard, G., et al. 2023. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Wu, X., Zhang, B., Deng, Z., et al. 2024. Vision-language dataset distillation. arXiv preprint arXiv:2308.07545v4.

Yuan, J., Gao, H., Dai, D., et al. 2025. Native sparse attention: Hardware-aligned and natively trainable sparse attention. arXiv preprint arXiv:2502.11089v2.

Youlai Doctor. 2024. Sternum fracture (33). Retrieved April 1, 2025, from https://www.youlai.cn/dise/imagedetail/343_72859.html

Zhang, Y., Liu, Z., Wang, X., et al. 2024. Gradient complexity profiling: A systematic approach to lightweight model capability boundary analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(3), 1452–1467.