
MMModal - Multi-Images Multi-Audio Multi-turn Multi-Modal

Husein Zolkepli*

Aisyah Razak[†]

Kamarul Adha[‡]

Ariff Nazhan[§]

Abstract

Our contribution introduces a groundbreaking multimodal large language model designed to comprehend multi-images, multi-audio, and multi-images-multi-audio within a single multiturn session. Leveraging state-of-the-art models, we utilize the SigLIP encoder for visual inputs and the Whisper Encoder for audio inputs. Notably, this multimodal large language model is bilingual, proficient in understanding both English and Malay simultaneously. We proudly unveil three versions of this model: Qwen1.5 with 0.5B parameters, TinyLlama with 1.1B parameters, and Mistral with 7B parameters. With its ability to navigate diverse modalities and languages, our model represents a significant advancement for the Malaysian context and beyond. All models released at [HuggingFace Mesolitica Multimodal Malaysian LLM](#).

1 Introduction

Language models trained with instructions have shown remarkable performance across various domains. However, their limitation in handling only text-based data hampers their applicability. Recent advancements in multimodal pre-training have demonstrated the potential to integrate knowledge from diverse modalities into a unified representation. [1] [2] Despite these advancements, there remains a lack of existing multimodal models capable of supporting multiple images/audio inputs and multiturn dialogue in current research. Addressing this gap, our proposal introduces MModal, a multi-modal instruction-tuned Language Learning Model (LLM) that combines image, audio, and text modalities within a single model architecture.

- **Synthetic Audio Instruction Dataset:** In the realm of Natural Language Processing (NLP), research indicates that the quality of instruction-following data significantly impacts the efficacy of instruction-following models. To illustrate this, we create a synthetic dataset comprising multi-dialogue interactions involving multiple images and audio inputs. Using Mistral, we generate a multiturn dialogue instruction dataset aimed at enhancing the language model's capacity to produce precise and contextually relevant responses. This dataset encompasses three distinct types of instruction-following data: multiple images instructions, multi-audio instructions, and combined image and audio instructions.
- **Synthetic Visual Malaysian Context Dataset:** a
- **Synthetic Multi-Images Multi-Audio relationship Dataset:** Our approach adopts a two-step training procedure to integrate multimodal and multiturn capabilities into our model. The initial step entails pretraining the feature alignment module. Through this process, we align the image features and audio data with the pre-trained word embeddings of the Language Learning Model (LLM). Specifically, this step involves training the projection

*husein@mesolitica.com

[†]aisyahrazak171@gmail.com

[‡]kamarul.adha360@gmail.com

[§]ariffnzhn@gmail.com

layer to ensure alignment between the multi-modal features and textual representations. This alignment facilitates seamless integration of diverse modalities within the model architecture.

- **Pretraining Feature Alignment:** a
- **Finetuned Multi-Images Multi-Audio Multi-turn Model:** a

2 Synthetic Data Generation for Audio Instructions

3 Synthetic Visual Malaysian Context Dataset

4 Synthetic Data Generation for Multi-Images, Multi-Audio Multi-turn Instructions

4.1 Synthetic Multi-Images Instruction

4.2 Synthetic Multi-Audio Instruction

4.3 Synthetic Image-Audio Instruction

5 Finetuning Procedure

5.1 Overall Architecture

5.2 Pretraining for Visual Feature Alignment

5.3 Pretraining for Audio Feature Alignment

5.4 Instruction Finetuning

The fine-tuning hyperparameters are detailed below:

Hyperparameter	Value
DeepSpeed	ZeRO-2 Offload
Batch Size	12
Learning Rate	constant 2e-5
Precision	bfloat16

Complete fine-tuning 8192 context length implementation at [here](#).

6 Examples

This section presents examples that highlight the model’s capacity to comprehend and produce responses relating to visual and audio input, showcasing the efficacy and potential of our proposed MMModal. These examples clearly demonstrate how the model handles and combines various information modalities, including audio and pictures, in the context of natural language processing (NLP). MMModal exhibits its capability by producing insightful, pertinent, and cohesive answers to a diverse set of inquiries. This highlights the model’s capacity to create exceptionally efficient interfaces for human-machine communication.

7 Evaluation

8 Future Work

In our future endeavors, we aim to enhance our capabilities by focusing on several key areas. Firstly, we intend to refine our approach to generating synthetic datasets that incorporate multi-images and multi-audio inputs. This will involve expanding the dataset to include more complex relationships

Multi Images Input Example



User

What is related between image 1 and image 2?

Multi Audio Input Example



User

What is related between audio 1 and audio 2?

between inputs and facilitating comparisons involving more than two inputs. Additionally, we recognize the importance of incorporating a wider range of visual Malaysian context datasets into our model training pipeline. By diversifying our data sources, we can ensure that our model is equipped to handle a broader array of real-world scenarios and contexts, ultimately improving its performance and relevance in practical applications.

9 Acknowledgement

Special thanks to Malaysia-AI volunteers especially [Wan Adzhar Faiq Adzlan](#), [Ammar Azman](#), [M. Amzar](#), [Muhammad Farhan](#), [Syafie Nizam](#), [Halim Shukor](#), [Alif Aiman](#), [Azwan Zuharimi](#) and [Haziq Zikry](#) for contributing dataset to train MMMModal.

We would like to express our gratitude to NVIDIA Inception for generously providing us with the opportunity to train our model on the Azure cloud. Their support has played a crucial role in the success of our research, enabling us to leverage advanced technologies and computational resources.

We extend our thanks to the wider research community for their valuable insights and collaborative discussions, which have greatly influenced our work. This paper reflects the collective efforts and contributions from both NVIDIA Inception and the broader research community.

10 Conclusion

In this paper, we introduce MMMModal, a multimodal instruction tuned Model (LLM) specifically designed to handle multiple modalities, including images, audio, and text in a multi-turn dialogue setting. Our novel approach focuses on aligning representations from various modality encoders into a unified space. Unlike existing methods, our approach enables the model to effectively process multi-turn dialogues and incorporate multiple images or audio inputs in its responses. We provide examples demonstrating the multi-modal understanding capabilities of MMMModal.

References

- [1] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.
- [2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.