MuMu-LLaMA: Multi-modal Music Understanding and Generation via Large Language Models

Shansong Liu[1*†], **Atin Sakkeer Hussain**[2*], **Qilong Wu**[2*], **Chenshuo Sun**[2], **Ying Shan**[1]

[1]ARC Lab, Tencent PCG
[2]National University of Singapore

dadinghh2@gmail.com, atin.s@u.nus.edu, qilong_wu@u.nus.edu,
csun@nus.edu.sg, yingsshan@tencent.com

Code: https://github.com/shansongliu/MuMu-LLaMA

Abstract

The landscape of research leveraging large language models (LLMs) has seen remarkable growth, with numerous studies harnessing these models' powerful reasoning capabilities across modalities like text, speech, images, and videos. However, the domain of multi-modal music comprehension and generation remains relatively unexplored, primarily due to the lack of a comprehensive well-annotated multi-modal music dataset. To address this gap, we introduce a novel dataset containing 167.69 hours of multi-modal data, including text, images, videos, and music annotations, tailored for multi-modal music understanding and generation, annotated using advanced visual models like LLaVA and Video-LLaVA. Based on this well-annotated dataset, we propose a multi-modal music understanding and generation model named MuMu-LLaMA. This framework integrates LLMs to comprehend input music and generate music across various modalities, utilizing pretrained models for music, images, and videos. For music generation, we incorporate AudioLDM 2 and MusicGen, connecting multi-modal understanding with music generation through the LLaMA model. Our comprehensive evaluation, encompassing four key tasks—music understanding, text-to-music generation, prompt-based music editing, and multi-modal music generation—demonstrates that MuMu-LLaMA outperforms current state-of-the-art models, highlighting the potential of combining LLMs with multi-modal inputs for innovative music applications.

Introduction

The landscape of research leveraging large language models (LLMs) has seen remarkable growth, with numerous studies harnessing these models' powerful reasoning capabilities across modalities like text, speech, images, and videos. These models facilitate semantic comprehension and interaction within and across modalities, enabling dynamic conversations (OpenAI 2023; Touvron et al. 2023), sophisticated audio and video event recognition (Tang et al. 2023a), and detailed image and 3D data annotation (Xu et al. 2023b). Despite these advancements, the domain of multi-modal music comprehension and generation remains relatively unexplored, primarily due to the lack of a comprehensive, well-annotated multi-modal music-centric dataset, which is crucial for instruction tuning in this domain.

Multi-modal large language models (MLLMs) have emerged as a thriving area of research, captivating the current scholarly landscape (Yin et al. 2023). They primarily serve as a bridge connecting diverse modalities, such as visual (Alayrac et al. 2022; Li et al. 2023a; Xu et al. 2023a), audio (Tang et al. 2023a; Huang et al. 2023; Liu et al. 2023c), 3D (Xu et al. 2023b; Sun et al. 2023) and so on, transcending mere textual interactions. This significant advancement greatly expands the application scenarios of large language models (LLMs).

Addressing this gap is crucial, as the absence of a well-annotated, balanced, and music-centric multi-modal dataset hinders progress in developing models that can effectively understand and generate music based on multi-modal inputs. To overcome this challenge, we introduce a novel dataset comprising 167.69 hours of multi-modal data, including text, images, videos, and music annotations. This dataset is specifically tailored for multi-modal music understanding and generation tasks. Annotated with advanced models such as LLaVA(Liu et al. 2023a) and Video-LLaVA(Lin et al. 2023), it offers a rich and diverse set of examples related to music, ensuring the quality and diversity of the data. The creation of this dataset is a pivotal step in advancing the field, as it equips models with the necessary training data to perform effectively across a wide range of music-related tasks.

Large language models are typically composed of a large number of parameters and trained on extensive datasets, endowing them with powerful comprehension and reasoning capabilities. Leveraging these qualities, researchers have utilized LLMs to achieve semantic understanding across various modalities. Examples include engaging in free-form conversations with humans (OpenAI 2023; Touvron et al. 2023), comprehending audio/video events and performing event-based question answering (Tang et al. 2023a; Huang et al. 2023; Muhammad Maaz and Khan 2023; Zhao et al. 2023), as well as captioning images/3D point cloud data (Chen et al. 2022; Li et al. 2023a; Xu et al. 2023b). In addition to harnessing the capabilities of LLMs for multi-modal understanding, researchers have also strived to utilize these models to grasp the creative intentions of humans. For instance, they have explored generating images (Brade et al. 2023), videos (Hong et al. 2023), audio (Liu et al. 2023d), or music (Copet et al. 2023a) based on textual descriptions, thereby providing valuable assistance in artistic pursuits. Especially for the visual content, some well-performance generative models like Stable Diffusion (Rombach et al. 2022; Blattmann et al. 2023) and Sora (Brooks et al. 2024) can produce high-quality images and videos which even human-eyes cannot distinguish.

[*]Equal contribution.
[†]Corresponding author.
This paper presents an updated version of M^{2}UGen by the same authors. Please reference this version in future citations.

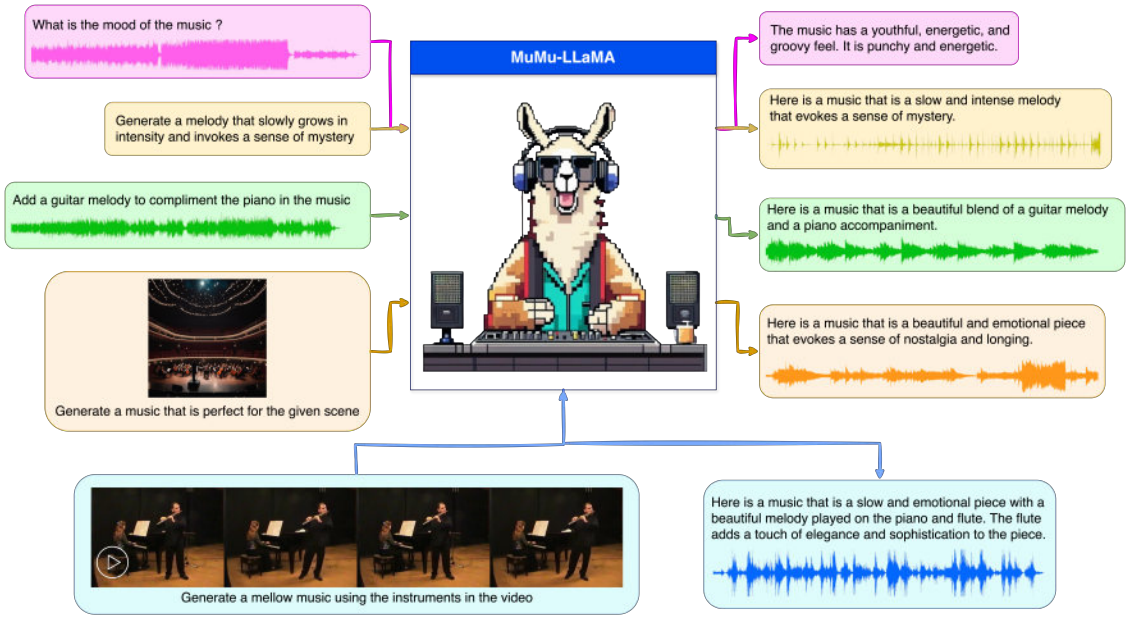


Figure 1: Multi-modal music understanding and generation by our proposed MuMu-LLaMA framework.

Building upon our collected dataset, we propose the **Mu**lti-modal **Mu**sic Understanding and Generation using **LLaMA** (**MuMu-LLaMA**) framework. This innovative framework leverages novel multi-modal adapters that, in conjunction with encoders such as ViT (Dosovitskiy et al. 2021), ViViT (Arnab et al. 2021), and MERT (Li et al. 2023b), effectively capture sequence-level information from various modalities—music, images, and videos—and transform these inputs into rich feature representations. The LLaMA model (Touvron et al. 2023) interprets these features to facilitate both comprehension and generation tasks, tailored to the user's intentions and contextual requirements.

By integrating understanding and generation tasks within the framework of LLMs, we have the potential to significantly enhance the user experience. For example, users can leverage LLMs to summarize videos and generate accompanying audio commentary or suitable background music, thus assisting them in their video creation process. However, research that combines both understanding and generation using LLMs is still limited and in its nascent stage (Moon et al. 2022; Ge et al. 2023a; Huang et al. 2023; Wu et al. 2023a; Guo et al. 2023; Yang et al. 2023; Zhou et al. 2024; Team 2024) especially when it also covers music modality. Among these few existing studies, NExT-GPT (Wu et al. 2023a) stands out as a significant advancement in the field of multi-modal large language models (MLLMs), excelling in both understanding and generation tasks. Notably, it demonstrates impressive capabilities, including music understanding and generation, image and video question answering, text-to-image and text-to-video generation, as well as audio-driven image and video generation. Despite these advancements, the exploration of music understanding and generation leveraging LLMs remains relatively unexplored. While NExT-GPT exhibits some capabilities in music understanding and generation, its proficiency in music-related tasks is modest due to the absence of specialized training on music datasets. To bridge this gap, we explore the use of LLMs for music understanding and multi-modal music generation.

Our contributions are summarized as follows:

1. We present a novel dataset containing 167.69 hours of well-annotated and balanced multi-modal data, annotated with advanced visual models, to support multi-modal music research.
2. We introduce the MuMu-LLaMA model, a novel data-centric architecture for comprehensive music understanding and multi-modal music generation.
3. Through rigorous evaluations across four key tasks—music understanding, text-to-music generation, prompt-based music editing, and multi-modal music generation—we demonstrate that MuMu-LLaMA outperforms existing state-of-the-art models, highlighting the potential of combining LLMs with multi-modal inputs for innovative music applications.

Related Works

Multi-modal Understanding

The integration of multi-modal data is pivotal in developing AI systems capable of interpreting the complex and heterogeneous information that defines human environments. Research in this domain covers a wide range of tasks, including audio/visual classification (Arnab et al. 2021), question answering (Lei et al. 2018), captioning (Mei et al. 2021), tagging (Gong, Chung, and Glass 2021), event detection (Dinkel, Wu, and Yu 2021), and summarization (Ji et al. 2019). The emergence of Vision Transformer (ViT) (Dosovitskiy et al. 2021) revolutionized the field of computer vision by enabling highly effective visual encoding, leading to the development of methodologies such as ViViT (Arnab et al. 2021), which incorporates both temporal and spatial data for enhanced video representation. Similarly, in the domain of music encoding, recent findings have highlighted the superiority of the MERT encoder (Li et al. 2023b) in downstream music tagging tasks. In our work, we align with these insights and leverage the MERT encoder to enhance the MuMu-LLaMA framework's ability to comprehend music-related data effectively, thereby ensuring robust performance in multi-modal understanding tasks.

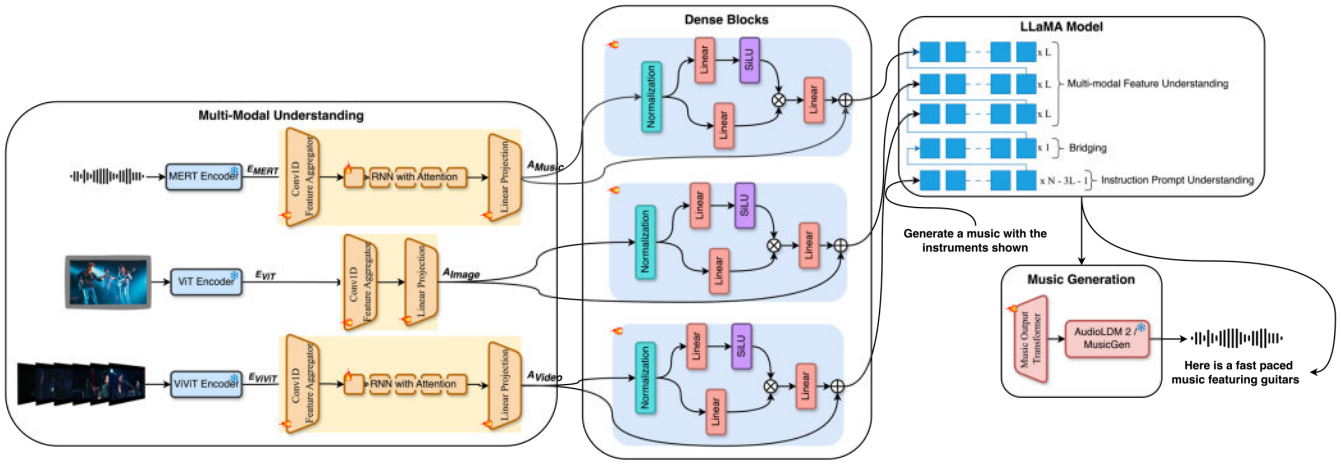


Figure 2: **Multi-modal Music Understanding and Generation Model (MuMu-LLaMA).** This model framework includes four core components: (1) Pre-trained feature encoders that process inputs from diverse modalities including music, images, and videos. (2) Understanding adapters that integrate these features into a coherent representation suitable for the LLaMA model. (3) The LLaMA model, which contextualizes and interprets the integrated information. (4) An output projection layer that translates the contextual understanding into outputs for the music generation decoder.

Multi-modal Music Generation

The field of music generation has witnessed substantial advancements, particularly with the adoption of Transformer (Vaswani et al. 2017) and diffusion models (Ho, Jain, and Abbeel 2020), which have significantly elevated the complexity and quality of generative AI outputs. Models such as MusicLM (Agostinelli et al. 2023) and MusicGen (Copet et al. 2023a) have established new benchmarks in music generation. MusicGen, with its autoregressive Transformer decoder, excels in sequence generation, while AudioLDM 2 (Liu et al. 2023b) utilizes a diffusion process to produce high-fidelity audio outputs. Despite these advancements, previous works like Vis2Mus (Zhang et al. 2022) and CMT (Di et al. 2021) primarily focus on single-modality music generation, limiting their applicability in diverse contexts. In contrast, our MuMu-LLaMA framework expands these concepts into a comprehensive multi-modal approach, integrating text, image, and video inputs. This approach not only enriches the music generation process but also provides a more holistic understanding and generation of music across various modalities, positioning it as a significant advancement in the field.

LLM-assisted Multi-modal Understanding and Generation

Multi-modal Large Language Models (MLLMs) have emerged as a frontier in AI research, aiming to unify the understanding and generation of diverse data modalities within a single framework. Innovations such as Macaw-LLM (Lyu et al. 2023) and DreamLLM (Dong et al. 2023) illustrate the potential of integrated multi-modal systems in enhancing user interactions through dynamic content generation. SEED-LLaMA (Ge et al. 2023b) combines the LLaMA model with diffusion techniques to achieve superior performance in image-related tasks. Similarly, NExT-GPT (Wu et al. 2023a) introduces a novel approach to manage multi-modal conversations, although it exhibits limitations in handling music-related content due to its restricted music training data. Building on these developments, our contribution, MuMu-LLaMA, specifically addresses the challenges and opportunities in multi-modal music understanding and generation. By enabling the modification of input music based on user prompts and integrating multiple modalities, MuMu-LLaMA significantly enhances AI's role in creative and artistic applications, aligning with the broader objectives of advancing innovative AI technologies with substantial societal and cultural impacts.

MuMu-LLaMA Model Architecture & Training

The MuMu-LLaMA model is engineered to harness the synergy of diverse modalities, particularly focusing on music, images, and videos. Figure 2 delineates the architecture which we detail below along with our innovative training methodologies.

Multi-modal Feature Encoders

To minimize training costs while ensuring the multi-modal encoder's robust capability to handle diverse multi-modal data inputs, MuMu-LLaMA leverages state-of-the-art pre-trained encoders to efficiently extract and integrate complex data across different sensory modalities.

Utilizing the MERT model (Li et al. 2023b), which excels in music tagging (Liu et al. 2023c), we encode music features $\mathbf{X}_{\text{Music}}$ into embeddings $\mathbf{E}_{\text{MERT}}$. The Vision Transformer (ViT) (Dosovitskiy et al. 2021) processes images $\mathbf{X}_{\text{Image}}$ into feature embeddings $\mathbf{E}_{\text{ViT}}$. For video data $\mathbf{X}_{\text{Video}}$, the Video Vision Transformer (ViViT) (Arnab et al. 2021) extracts spatio-temporal tokens, resulting in embeddings $\mathbf{E}_{\text{ViViT}}$.

$$\mathbf{E}_{\text{MERT}} = F_{\text{MERT}}(\mathbf{X}_{\text{Music}}) \in \mathbb{R}^{25\times1024} \tag{1}$$

$$\mathbf{E}_{\text{ViT}} = F_{\text{ViT}}(\mathbf{X}_{\text{Image}}) \in \mathbb{R}^{197\times768} \tag{2}$$

$$\mathbf{E}_{\text{ViViT}} = F_{\text{ViViT}}(\mathbf{X}_{\text{Video}}) \in \mathbb{R}^{3137\times768} \tag{3}$$

where $F_{\text{MERT}}(\cdot)$, $F_{\text{ViT}}(\cdot)$, and $F_{\text{ViViT}}(\cdot)$ represent feature encoders specifically designed for music, image, and video modalities. The resulting embeddings $\mathbf{E}_{\text{MERT}}$, $\mathbf{E}_{\text{ViT}}$, and $\mathbf{E}_{\text{ViViT}}$ have dimensions of 25×1024, 197×768, and 3137×768.

Multi-modal Understanding Adapters

To integrate diverse modal outputs within the LLaMA framework (Touvron et al. 2023), we employ multi-modal understanding adapters, which consist of a 1D convolutional layer, linear projection, and a dense network. Since music and video data inherently possess a time dimension and are highly time-dependent, we utilize an RNN coupled with an attention mechanism to project these temporal inputs into a shared 4096-dimensional space. In contrast, image data does not require temporal processing, so we bypass the RNN and attention components for this modality, ensuring efficient and effective cross-modal interaction.

For music and video understanding, the input features $\mathbf{E}_{\text{MERT}}$ for music and $\mathbf{E}_{\text{ViViT}}$ for video are first encoded using the MERT and ViViT encoders, respectively. These encoded features are then passed through a 1D convolutional layer (Conv1D feature aggregator) to capture local temporal dependencies for better multi-modal alignment. Following this, the features are further processed through an RNN with attention mechanisms to model the sequential nature of the data. The RNN output, $\mathbf{A}_{\text{RNN}} \in \mathbb{R}^{L\times d}$, where L denotes the sequence length and d is the feature dimension, is used to compute the Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) matrices:

$$\mathbf{Q} = \mathbf{A}_{\text{RNN}}\mathbf{W}^{Q}, \quad \mathbf{K} = \mathbf{A}_{\text{RNN}}\mathbf{W}^{K}, \quad \mathbf{V} = \mathbf{A}_{\text{RNN}}\mathbf{W}^{V} \tag{4}$$

where $\mathbf{W}^{Q} \in \mathbb{R}^{d\times d}$, $\mathbf{W}^{K} \in \mathbb{R}^{d\times d}$, and $\mathbf{W}^{V} \in \mathbb{R}^{d\times d}$ are learnable weight matrices. The attention scores are then calculated as:

$$\text{Attention\_Scores} = \text{Softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}^{T}}{\sqrt{d}}\right) \tag{5}$$

Finally, the attention-weighted output is linearly projected to generate the final modality-specific embeddings:

$$\mathbf{A}_{\text{Music}} = \text{LinearProjection}(\text{Attention\_Scores}\cdot\mathbf{V}) \tag{6}$$

$$\mathbf{A}_{\text{Video}} = \text{LinearProjection}(\text{Attention\_Scores}\cdot\mathbf{V}) \tag{7}$$

In contrast, for image understanding, the input features $\mathbf{E}_{\text{ViT}}$ for images are encoded using the ViT encoder. Since image data does not have a temporal dimension, the encoded features skip the RNN and attention layers. Instead, the features are directly passed through the Conv1D Feature Aggregator and then linearly projected to generate the final image-specific embedding:

$$\mathbf{A}_{\text{Image}} = \text{LinearProjection}(F_{\text{Conv1D}}(\mathbf{E}_{\text{ViT}})) \tag{8}$$

This method for multi-modal understanding ensures that each modality's unique characteristics are appropriately handled to produce rich, unified representations for subsequent tasks.

LLM as a Bridge

MuMu-LLaMA strategically integrates multi-modal data into LLaMA to enhance context-aware processing. Modality-specific features are introduced every 6 layers within the 32-layer structure, with the last 18 layers divided into three sets of 6 layers (L_{18-23}, L_{24-29}, L_{30-35}) corresponding to each modality. In each of these sets, hidden states from previous layers are combined with modality-specific features, ensuring that the model effectively processes the input from different modalities.

For each set of layers, the hidden states $\mathbf{H}_{\text{LLaMA}}$ from the previous layers are combined with the modality-specific features $\mathbf{A}_{\text{Modality}}$ (i.e., $\mathbf{A}_{\text{Video}}$, $\mathbf{A}_{\text{Image}}$, or $\mathbf{A}_{\text{Music}}$) and a corresponding prefix query $\mathbf{P}_{\text{query},i}$. The equations governing this integration process are as follows:

$$\mathbf{H}_{\text{LLaMA}}^{(18-23)} = \text{Layer}_{18-23}\left(\mathbf{H}_{\text{LLaMA}}^{(12-17)}, \mathbf{A}_{\text{Video}} + \mathbf{P}_{\text{query},1}\right) \tag{9}$$

$$\mathbf{H}_{\text{LLaMA}}^{(24-29)} = \text{Layer}_{24-29}\left(\mathbf{H}_{\text{LLaMA}}^{(18-23)}, \mathbf{A}_{\text{Image}} + \mathbf{P}_{\text{query},2}\right) \tag{10}$$

$$\mathbf{H}_{\text{LLaMA}}^{(30-35)} = \text{Layer}_{30-35}\left(\mathbf{H}_{\text{LLaMA}}^{(24-29)}, \mathbf{A}_{\text{Music}} + \mathbf{P}_{\text{query},3}\right) \tag{11}$$

Here, $\mathbf{P}_{\text{query},i} \in \mathbb{R}^{d}$ represents the prefix query for each modality-specific integration step, which provides an additional contextual signal during the multi-modal fusion. The prefix query is crucial for maintaining consistency in the integration process across different layers and modalities. If a modality is unavailable at a particular layer, the model defaults to using only the prefix query for that integration step. The final hidden state $\mathbf{H}_{\text{LLaMA}}^{35}$ is then normalized and passed through the LLaMA output layer, completing the multi-modal integration process and ensuring that the combined features are well-prepared for subsequent tasks.

Music Understanding and Generation

Inspired by models like NExT-GPT (Wu et al. 2023a), our framework employs discrete audio tokens $[AUD_i]$ ($i \in \{0, 1, \cdots, 7\}$) to enable dynamic music understanding and generation. This strategy allows for context-sensitive generation of music or text, depending on the task requirements during inference.

Training Method

Considering the computational demands of training from scratch, we utilize the LoRA fine-tuning approach (Hu et al. 2022), which allows us to effectively adapt MuMu-LLaMA's capabilities while freezing the base encoders and generative models. This not only conserves computational resources but also accelerates the training process. The loss function is designed to optimize the following components depending on the task:

$$\text{Loss} = \begin{cases} L_{CE}(y_{\text{tokens}}, f(y)_{\text{logits}}) \\ \quad + \|y_{\text{embeddings}} - g(f(x)_{\text{hidden}})\|, & \text{if music} \\ L_{CE}(y_{\text{tokens}}, f(y)_{\text{logits}}), & \text{else} \end{cases} \tag{12}$$

Here, L_{CE} denotes the cross-entropy loss, crucial for refining text token generation and $g(\cdot)$ denotes the Music Output Transformer. For music generation tasks, we employ mean squared error (MSE) to align the generated embeddings with target music captions, ensuring the high fidelity of generated audio content. An additional regularization term penalizes improper generation of audio tokens, promoting precision in both textual and musical outputs.

Music Oriented Instruction Dataset

Training MLLMs requires extensive data, but there is a shortage of multi-modal datasets focused on music-related tasks. MusicCaps (Agostinelli et al. 2023) and MusicQA (Liu et al. 2023c) are the largest public datasets for music captioning and question answering, but they are insufficient for multi-modal music understanding and generation. For our MuMu-LLaMA model, we need multi-modal instruction datasets for any-to-music generation and extensive datasets such as text-image pairs for alignment training. We use datasets such as Alpaca (Taori et al. 2023) for instruction following and COCO (Lin et al. 2014) for image encoder alignment. Additionally, we collect our own dataset using automated methods inspired by previous works (Liu et al. 2023c; Gong et al. 2023), leveraging models like MU-LLaMA (Liu et al. 2023c), Mistral-7B-Instruct (Jiang et al. 2023), LLaVA (Liu et al. 2023a), and VideoLLaVA (Lin et al. 2023) to perform data annotation.

We create a comprehensive multi-modal dataset with a total of 167.69 hours to enhance MuMu-LLaMA's performance. This

Table 1: Descriptions of Our Proposed Music Dataset. A/D/R* represents Add/Delete/Replace.

Dataset		# Audios	Avg. Time (s)/Audio	Total Time (h)		Data Source
MUCaps		18,515	10.00	51.43		AudioSet (Gemmeke et al. 2017)
MUImage		14,520	10.00	40.33		Balanced-AudioSet (Gemmeke et al. 2017)
MUVideo		14,504	10.00	40.29		Balanced-AudioSet (Gemmeke et al. 2017)
MUEdit	Speed	2384	15.11	10.01		Looperman (Looperman 2000)
	Pitch	2369	15.20	10.00	35.64	Looperman (Looperman 2000)
	A/D/R*	229	245.66	15.63		Slakh (Manilow et al. 2019)

fine-grained captioned dataset supports multi-modal understanding and generation. Detailed descriptions are in Table 1. Figure 3 shows balanced instrument distributions in MUImage and MUVideo datasets. Figures 3b and 3c indicate a generally balanced distribution in MUEdit, while Figure 3a shows a long-tail distribution due to the rarity of certain instruments during data collection and processing.

In the following subsections, we provide a comprehensive overview of the methodologies employed in crafting the datasets used for training the MuMu-LLaMA model.

MUCaps Dataset

We develop the MUCaps dataset, composed of text-music pairs (Table 1), encompassing 51.43 hours of 10-second music files sourced from AudioSet (Gemmeke et al. 2017) and publicly accessible music websites. The MU-LLaMA model captions the music files with the question: *"Describe the music in detail, including aspects such as instruments used, tempo, and the mood of the song"*. The MUCaps dataset is used for encoder and decoder alignment training.

MUEdit Dataset

To enable music editing in response to prompts, we curated the MUEdit dataset, which includes 35.64 hours of music pairs (Table 1). The dataset generation involves:

1. Use the WSOLA algorithm (Grofit and Lavner 2008) in the sox (Klauer 1999) tool to generate speed and pitch-changed music files for origin-to-target pairs, and manipulate individual tracks in the Slakh (Manilow et al. 2019) dataset for origin-to-Add/Delete/Replace pairs.
2. Employ the Mistral-7B-Instruct (Jiang et al. 2023) model to generate an instruction-response pool with hundreds of templates for each MUEdit subtype, diversifying the dataset and enhancing model robustness.
3. For each origin-to-X music pair, randomly select instructions and responses from the template pool to construct the final MUEdit dataset.

Speed The WSOLA algorithm modifies music speed without changing pitch to create the Speed split of MUEdit. Supported duration changes are 0.5, 0.7, 1.3, and 1.5 times the original. Instruction-response pairs represent these changes explicitly with numerical values and by degrees of speed variation.

Pitch The Pitch split of MUEdit uses the sox tool's pitch shift feature, allowing pitch changes without altering duration. Permitted pitch variations are ± 100 and ± 200 cents, representing a semitone and a whole tone. Instructions express pitch variations through verbal expressions, not just numerical values.

Add/Delete/Replace For this branch, we use MIDI data with explicit individual instrument tracks from the Slakh dataset, suitable for tasks like music source separation. In the "Add" sub-split, different tracks from the same MIDI are combined. In the "Delete" sub-split, the mixed track is the input, and individual tracks are the output. In the "Replace" sub-split, tracks from different instruments are selected as input and output.

MUImage Dataset

The MUImage dataset generates fitting music for a given image by pairing music samples from AudioSet with corresponding images. Key steps include:

1. Use the MU-LLaMA (Liu et al. 2023c) model to generate music captions for the sampled music files.
2. Generate captions for the corresponding images using the LLaVA-v1.6-34B model (Liu et al. 2023a, 2024) with detailed instructions to describe the image, focusing on instruments and other relevant elements.
3. Employ the Mistral-7B-Instruct model (Jiang et al. 2023) to produce approximately 200 unique instructions that begin with "Generate," such as: "Generate music to match the image."
4. Generate the model side of the conversation to integrate information from ground-truth tags, music, and image captions, following specific instructions to describe and match the music to the image.

MUVideo Dataset

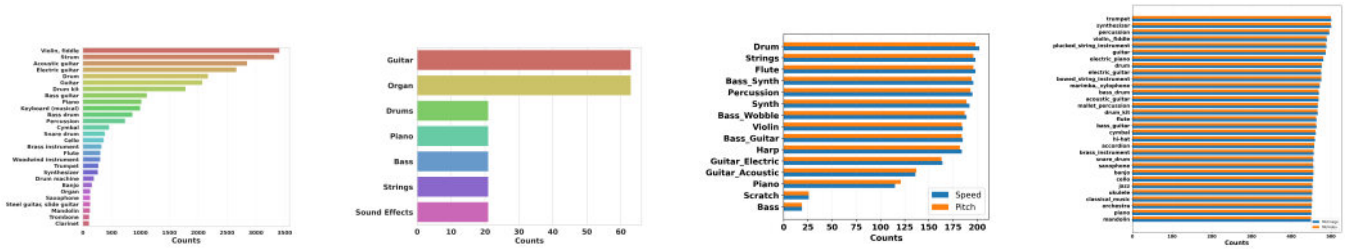
The MUVideo dataset facilitates video-to-music generation and understanding, sourcing music samples and corresponding videos from the Balanced-AudioSet (Gemmeke et al. 2017). Key steps include:

1. Use the MU-LLaMA model to generate captions for all acquired music files.
2. Generate captions for the corresponding videos using the VideoLLaVA captioning model (Lin et al. 2023) with detailed instructions to describe the video, focusing on dynamic changes, storyline progression, and visual cues.
3. Generate both the human and model sides of the conversation using a process similar to the MUImage dataset.

Efforts are made to minimize overlaps among the music files in all datasets. Evaluation sets are established to compare our model's performance with current state-of-the-art (SOTA) models.

Model Evaluation

We extensively evaluate MuMu-LLaMA on tasks involving music understanding and generation from multi-modal inputs, and compare its performance against other state-of-the-art models. Direct comparison with NExT-GPT (Wu et al. 2023a) was not feasible due to issues with accessing the required checkpoints. For a fair



(a) MUCaps (b) MUEdit - A/D/R (c) MUEdit - Speed & Pitch (d) MUImage & MUVideo

Figure 3: **Distribution of instrument categories in our four curated datasets:** (a) MUCaps reveals a broad diversity of instruments with a long-tail distribution. (b) MUEdit - A/D/R shows a relatively even distribution of add, delete, and replace manipulations across various instruments. (c) MUEdit - Speed & Pitch demonstrates a consistent distribution of speed and pitch modifications, suggesting balanced attention to tempo and tonal adjustments. (d) MUImage & MUVideo illustrates a balanced pairing of instruments with corresponding images and videos, ensuring a wide representation within these multi-modal components.

evaluation, MuMu-LLaMA's hyperparameters were set to a temperature of 0.6, top_p of 0.8, and a maximum target length of 512 tokens. These hyperparameters were consistently applied across other models in the evaluation, including LLaMA-Adapter (Gao et al. 2023), MU-LLaMA (Liu et al. 2023c), and SALMONN (Tang et al. 2023a). Notably, MuMu-LLaMA was paired with the MusicGen decoder, which showed superior performance compared to the AudioLDM 2 decoder. To further understand the contributions of each component within MuMu-LLaMA, an ablation study was conducted.

Music Understanding

MuMu-LLaMA's music understanding capabilities were evaluated using the MTG-eval-QA subset of the MusicQA dataset (Liu et al. 2023c), consisting of 4,500 music-related question-answer pairs. The evaluation was conducted against several state-of-the-art (SOTA) models including LTU (Gong et al. 2023), LLaMA-Adapter (Gao et al. 2023), SALMONN (Tang et al. 2023a), and MU-LLaMA (Liu et al. 2023c), the latter being specifically trained on music-related datasets. We employed well-established evaluation metrics such as BLEU (B-U) (Papineni et al. 2002a), METEOR (M-R) (Banerjee and Lavie 2005a), ROUGE$_L$ (R-L) (Lin 2004a), and BERT-Score (BERT-S) (Zhang* et al. 2020a) to quantify the model's performance.

Data presented in Table 2 indicate that MuMu-LLaMA outperforms the other models significantly across all metrics. This superior performance is largely attributed to MuMu-LLaMA's advanced music understanding adapter, which incorporates an additional RNN layer and attention mechanism. These components are particularly effective in capturing the temporal information inherent in musical sequences, enabling MuMu-LLaMA to produce more accurate and contextually relevant responses.

Table 2: **Evaluation of Models on Music Understanding**. The best values of different metrics are made **bold**.

Music Understanding				
Model	**B-U↑**	**M-R↑**	**R-L↑**	**BERT-S↑**
LTU	0.242	0.274	0.326	0.887
LLaMA Adapter	0.273	0.334	0.413	0.895
SALMONN	0.286	0.332	0.371	0.898
MU-LLaMA	0.306	0.385	0.466	0.901
MuMu-LLaMA	**0.341**	**0.442**	**0.491**	**0.908**

Text-to-Music Generation

For the task of text-to-music generation, we utilize the MUCaps dataset's 5,000 text-music pairs, comparing MuMu-LLaMA with state-of-the-art models like CoDi (Tang et al. 2023b), AudioLDM 2 (Liu et al. 2023b), and MusicGen (Copet et al. 2023a), employing Fréchet Audio Distance (FAD) (Kilgour et al. 2019), Kullback-Leibler divergence (KL), and CLAP score (Wu et al. 2023b) for evaluation. As shown in Table 3, MuMu-LLaMA demonstrates superior performance, particularly when paired with the MusicGen decoder, which enhances the relevance of the generated music to the input instructions, evidenced by higher CLAP scores. This improvement is largely due to the integration of Large Language Models (LLMs), which enhance the model's comprehension and effective use of input instructions for guiding music generation.

Prompt-Based Music Editing

MuMu-LLaMA stands out as one of the few models supporting music editing through natural language commands, unlike AUDIT (Wang et al. 2023) and InstructME (Han et al. 2023), which require specific prompt words like "Add" or "Remove." Although Loop Copilot (Zhang et al. 2023) also offers natural language-based editing, it is not open-sourced and thus excluded from our comparison. For AUDIT and InstructME, which are also not open-sourced, we relied on sample outputs available on InstructME's official website for comparison purposes.

To evaluate music editing capabilities, we adopted AUDIT's evaluation metrics, including Fréchet Audio Distance (FAD) and Kullback-Leibler divergence (KL), and introduced log spectral distance (LSD) (Gray and Markel 1976) for additional assessment. The results in Table 3 show MuMu-LLaMA's superior performance over AUDIT and InstructME, attributed to its use of the LLaMA model for interpreting natural language prompts and the MERT Encoder for understanding source music, significantly enhancing its editing capabilities.

Multi-modal Music Generation

MuMu-LLaMA can generate music from images and videos, a significant feature that sets it apart in the realm of multi-modal music generation. In our experiments, we compare MuMu-LLaMA with CoDi (Tang et al. 2023b), an any-to-any generation model capable of both image-to-music (I2M) and video-to-music (V2M) tasks, as well as with CMT (Di et al. 2021) specifically for V2M tasks. The evaluation sets for these tasks consist of 2,500 pairs each of image-music and video-music, providing a robust basis for comparison.

Table 3: **Comparison of Models for Music Generation**. The best values of different metrics are made **bold**.

Model	Text-to-Music Generation			Prompt-based Music Editing			Image-to-Music Generation			Video-to-Music Generation		
	$FAD_{vgg}\downarrow$	$KL\downarrow$	$CLAP_{score}\uparrow$	$FAD_{vgg}\downarrow$	$KL\downarrow$	$LSD\downarrow$	$FAD_{vgg}\downarrow$	$KL\downarrow$	*IB Rank*↑	$FAD_{vgg}\downarrow$	$KL\downarrow$	*IB Rank*↑
CoDi	16.201	6.021	0.143	N/A	N/A	N/A	10.788	9.925	0.493	11.273	6.267	0.212
AudioLDM 2	11.619	4.074	0.238	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MusicGen	10.697	3.909	0.289	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
AUDIT	N/A	N/A	N/A	2.855	9.925	0.987	N/A	N/A	N/A	N/A	N/A	N/A
InstructME	N/A	N/A	N/A	2.442	6.018	0.846	N/A	N/A	N/A	N/A	N/A	N/A
CMT	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	9.021	5.991	0.629
MuMu-LLaMA	**9.982**	**3.191**	**0.312**	**1.911**	**5.028**	**0.705**	**6.289**	**5.021**	**0.882**	**7.959**	**4.784**	**0.891**

To assess the performance of MuMu-LLaMA, we employ traditional metrics such as Fréchet Audio Distance (FAD) and Kullback-Leibler divergence (KL), and introduce ImageBind Ranking (IB Rank) (Girdhar et al. 2023), a novel metric designed to evaluate the alignment between the input modality (image/video) and the generated music. This is achieved by using the ImageBind model to generate embeddings for both the visual input and the corresponding music output, allowing for the calculation of similarity scores that reflect how well the music matches the visual content.

As evidenced by the results in Table 3, MuMu-LLaMA demonstrates exceptional performance in multi-modal music generation, both in terms of the quality of the music produced and its relevance to the input modality. The model consistently outperforms other state-of-the-art (SOTA) models, highlighting its advanced capability to generate music that is not only high in quality but also closely aligned with the visual content, whether it be from images or videos.

Ablation Study

We evaluated the contributions of the multi-modal understanding adapter's components through an ablation study focused on the dense network and the RNN with attention mechanism for the music understanding task. The results, presented in Table 4, provide clear insights into the importance of each component in enhancing MuMu-LLaMA's performance. Specifically, the model variant that excluded both the dense network and the attention RNN showed the lowest performance across all evaluation metrics, indicating the critical role these components play in the architecture.

Adding the dense network to the model led to noticeable improvements in the evaluation metrics, demonstrating its effectiveness in refining the feature representations. Incorporating the RNN component further enhanced the model's performance, with a more significant impact than the dense network alone, suggesting the RNN's crucial role in capturing temporal dependencies in music data. The inclusion of the attention mechanism on top of the RNN provided an additional boost, emphasizing its importance in focusing on the most relevant musical features. Ultimately, the complete MuMu-LLaMA model, utilizing all these components, achieved the best performance, underscoring the synergistic contributions of the dense network, RNN, and attention mechanism in facilitating comprehensive music understanding.

Table 4: **Ablation study of our MU-LLaMA model on the music understanding task**. The best values of different metrics are made **bold**.

Model	B-U↑	M-R↑	R-L↑	BERT-S↑
w/ Projection layer	0.277	0.302	0.326	0.876
w/ Dense Network	0.303	0.354	0.401	0.880
w/ RNN	0.313	0.367	0.411	0.886
w/ Attn. RNN	0.336	0.375	0.439	0.894
MuMu-LLaMA	**0.341**	**0.442**	**0.491**	**0.908**

Subjective Evaluation for Music Generation

To assess our model's music generation capabilities, we conducted a subjective evaluation involving 45 participants. For this evaluation, we designed 13 questions covering three distinct tasks: text-to-music (T2M), image-to-music (I2M), and video-to-music (V2M) generation. Each question presented participants with options generated by different models, and these options were randomly shuffled to eliminate any potential preference bias. This approach ensures that participants' choices were based solely on the quality and relevance of the generated music, rather than any preconceived notions about the models.

In this evaluation, AudioLDM 2 (Liu et al. 2023b) and MusicGen (Copet et al. 2023b) were assessed exclusively on the T2M task, while NExT-GPT (Wu et al. 2023a) and MuMu-LLaMA were evaluated across all three tasks. The CoDi model (Tang et al. 2023c), however, was not included in the T2M task due to its limited performance in that area. The results, as presented in Table 5, indicate that our proposed MuMu-LLaMA model consistently received the highest preference among participants across all three music generation tasks. This strong preference highlights MuMu-LLaMA's superior ability to generate music that aligns closely with user inputs across various modalities.

Table 5: **Subjective comparison of models for three music generation tasks**. The best values of different metrics are made **bold**.

Model	T2M	I2M	V2M
AudioLDM 2	11.6%	N/A	N/A
MusicGen	21.3%	N/A	N/A
CoDi	N/A	5.8%	4.4%
NExT-GPT	8.9%	12.9%	14.8%
MuMu-LLaMA	**58.2%**	**81.3%**	**80.7%**

Conclusion

This paper presents the MuMu-LLaMA model, a novel framework utilizing a large language model (LLM) for integrated music comprehension and multi-modal music generation. Our contributions include not only the development of the model but also a comprehensive methodology for generating specialized datasets to train it. Experimental results demonstrate that MuMu-LLaMA surpasses existing state-of-the-art models in tasks such as music comprehension, music editing, and music generation from text, image, and video inputs. Future work will focus on refining the model's understanding of complex musical nuances and improving the alignment of generated music with diverse multi-modal inputs, further advancing AI's capabilities in creative and cultural applications.

Limitations

The reliance on the pre-trained MusicGen/AudioLDM 2 model for music generation introduces several challenges that can act as a bottleneck for the overall performance and flexibility of the MuMu-LLaMA model. While these pre-trained models are state-of-the-art in their respective domains, they come with inherent limitations that may affect the MuMu-LLaMA's ability to generate high-quality, context-aware music.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; et al. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; et al. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lucic, M.; and Schmid, C. 2021. ViViT: A Video Vision Transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6816–6826.

Banerjee, S.; and Lavie, A. 2005a. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop*, 65–72.

Banerjee, S.; and Lavie, A. 2005b. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop*, 65–72.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; et al. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127.

Brade, S.; Wang, B.; Sousa, M.; Oore, S.; and Grossman, T. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–14.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; et al. 2024. Video generation models as world simulators.

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18030–18040.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023a. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284*.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; et al. 2023b. Simple and Controllable Music Generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Di, S.; Jiang, Z.; Liu, S.; Wang, Z.; et al. 2021. Video Background Music Generation with Controllable Music Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2037–2045.

Dinkel, H.; Wu, M.; and Yu, K. 2021. Towards Duration Robust Weakly Supervised Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 887–900.

Dong, R.; Han, C.; Peng, Y.; Qi, Z.; Ge, Z.; Yang, J.; Zhao, L.; Sun, J.; Zhou, H.; Wei, H.; et al. 2023. DreamLLM: Synergistic Multimodal Comprehension and Creation. *arXiv preprint arXiv:2309.11499*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; Li, H.; and Qiao, Y. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

Ge, Y.; Ge, Y.; Zeng, Z.; Wang, X.; and Shan, Y. 2023a. Planting a Seed of Vision in Large Language Model. *arXiv preprint arXiv:2307.08041*.

Ge, Y.; Zhao, S.; Zeng, Z.; Ge, Y.; Li, C.; Wang, X.; and Shan, Y. 2023b. Making LLaMA SEE and Draw with SEED Tokenizer. *arXiv preprint arXiv:2310.01218*.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An Ontology and Human-labeled Dataset for Audio Events. In *Proc. IEEE ICASSP 2017*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One Embedding Space To Bind Them All. In *CVPR*.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3292–3306.

Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; and Glass, J. 2023. Listen, Think, and Understand. *arXiv preprint arXiv:2305.10790*.

Gray, A.; and Markel, J. 1976. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5): 380–391.

Grofit, S.; and Lavner, Y. 2008. Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients. *IEEE Transactions on Audio, Speech, and Language Processing*, 16: 106–115.

Guo, Z.; Zhang, R.; Zhu, X.; Tang, Y.; Ma, X.; Han, J.; Chen, K.; Gao, P.; Li, X.; Li, H.; et al. 2023. Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following. *arXiv preprint arXiv:2309.00615*.

Han, B.; Dai, J.; Song, X.; Hao, W.; He, X.; Guo, D.; Chen, J.; Wang, Y.; and Qian, Y. 2023. InstructME: An Instruction Guided Music Edit And Remix Framework with Latent Diffusion Models. *arXiv preprint arXiv:2308.14360*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2023. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In *The Eleventh International Conference on Learning Representations*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, R.; Li, M.; Yang, D.; Shi, J.; Chang, X.; Ye, Z.; Wu, Y.; Hong, Z.; Huang, J.; Liu, J.; et al. 2023. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. *arXiv preprint arXiv:2304.12995*.

Ji, Z.; Xiong, K.; Pang, Y.; and Li, X. 2019. Video Summarization with Attention-based Encoder-Decoder Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6): 1709–1717.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech*.

Klauer, U. 1999. SoX - Sound eXchange.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1369–1379. Association for Computational Linguistics.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; et al. 2023b. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. *arXiv preprint arXiv:2306.00107*.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*.

Lin, C.-Y. 2004a. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, 74–81.

Lin, C.-Y. 2004b. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; et al. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; Li, B.; et al. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. arXiv:2304.08485.

Liu, H.; Tian, Q.; Yuan, Y.; Liu, X.; Mei, X.; et al. 2023b. AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining. *arXiv preprint arXiv:2308.05734*.

Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2023c. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. *arXiv preprint arXiv:2308.11276*.

Liu, X.; Zhu, Z.; Liu, H.; Yuan, Y.; Huang, Q.; Liang, J.; Cao, Y.; Kong, Q.; Plumbley, M. D.; and Wang, W. 2023d. WavJourney: Compositional Audio Creation with Large Language Models. *arXiv preprint arXiv:2307.14335*.

Looperman. 2000. Looperman - Free Loops, Beats, Samples, Acapellas.

Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; and Tu, Z. 2023. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv preprint arXiv:2306.09093*.

Manilow, E.; Wichern, G.; Seetharaman, P.; and Le Roux, J. 2019. Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE.

Mei, X.; Liu, X.; Huang, Q.; Plumbley, M. D.; and Wang, W. 2021. Audio Captioning Transformer. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, 211–215.

Moon, J. H.; Lee, H.; Shin, W.; Kim, Y.-H.; and Choi, E. 2022. Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. *IEEE Journal of Biomedical and Health Informatics*, 26(12): 6070–6080.

Muhammad Maaz, S. K., Hanoona Rasheed; and Khan, F. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*.

OpenAI. 2023. ChatGPT (Mar 14 version) [Large language model].

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002a. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002b. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Sun, C.; Han, J.; Deng, W.; Wang, X.; Qin, Z.; and Gould, S. 2023. 3D-GPT: Procedural 3D Modeling with Large Language Models. *arXiv preprint arXiv:2310.12945*.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2023a. SALMONN: Towards Generic Hearing Abilities for Large Language Models. *arXiv preprint arXiv:2310.13289*.

Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; and Bansal, M. 2023b. Any-to-Any Generation via Composable Diffusion. *arXiv preprint arXiv:2305.11846*.

Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; and Bansal, M. 2023c. Any-to-Any Generation via Composable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Team, C. 2024. Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv:2405.09818.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. *Advances in neural information processing systems*, 30.

Wang, Y.; Ju, Z.; Tan, X.; He, L.; Wu, Z.; Bian, J.; and Zhao, S. 2023. AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models. *arXiv preprint arXiv:2304.00830*.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2023a. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05519*.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023b. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Xu, L.; Liu, B.; Khan, A. H.; Fan, L.; and Wu, X.-M. 2023a. Multimodal Pre-training for Medical Vision-language Understanding and Generation: An Empirical Study with A New Benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, 117–132.

Xu, R.; Wang, X.; Wang, T.; Chen, Y.; Pang, J.; and Lin, D. 2023b. PointLLM: Empowering Large Language Models to Understand Point Clouds. *arXiv preprint arXiv:2308.16911*.

Yang, Z.; Zhang, Y.; Meng, F.; and Zhou, J. 2023. TEAL: Tokenize and Embed ALL for Multi-modal Large Language Models. *arXiv preprint arXiv:2311.04589*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; et al. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

Zhang, R.; Zhang, Y.; Shao, K.; Shan, Y.; and Xia, G. 2022. Vis2Mus: Exploring Multimodal Representation Mapping for Controllable Music Generation. *arXiv preprint arXiv:2211.05543*.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020a. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Zhang, Y.; Maezawa, A.; Xia, G.; Yamamoto, K.; and Dixon, S. 2023. Loop Copilot: Conducting AI Ensembles for Music Generation and Iterative Editing. *arXiv preprint arXiv:2310.12404*.

Zhao, Y.; Misra, I.; Krähenbühl, P.; and Girdhar, R. 2023. Learning Video Representations from Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6586–6597.

Zhou, C.; Yu, L.; Babu, A.; Tirumala, K.; Yasunaga, M.; Shamis, L.; Kahn, J.; Ma, X.; Zettlemoyer, L.; and Levy, O. 2024. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. arXiv:2408.11039.

Appendix

In the appendix, comprehensive information is provided concerning the model's training dataset and training methodology, encompassing explicit insights into the utilized training approach and the corresponding model hyperparameters. Additionally, a thorough exposition is given regarding the composition of the evaluation sets employed in our study, accompanied by a delineation of the evaluation methodology and metrics applied to assess the performance of our model. To elucidate the diverse capabilities of our model, illustrative demo examples are also included.

Music Oriented Dataset Information

We generate 4 different datasets to train the MuMu-LLaMA model: MUCaps, MUImage, MUVideo and MUEdit datasets. An example of each from the 4 datasets are shown in Figure 4.

Model Training

In this section, we detail the training strategy for the MuMu-LLaMA model along with parameters used for training.

Model Training Strategy

The MuMu-LLaMA model adopts the adapter training strategy, implementing a three-step training regimen. In the first phase, all parameters, with the exception of those associated with the Multi-modal Understanding Adapters, undergo freezing. The training dataset is configured to incorporate the MUCaps dataset for music understanding, the COCO dataset for image comprehension, and the captions sourced from the MUVideo dataset for video understanding. During this training stage, the Cross Entropy Loss function is applied to compute the disparity between the caption generated by the LLaMA model and the target caption corresponding to the input modality. This process is illustrated in Figure 5.

In the second training stage, the output projector is trained to generate conditional embeddings using input captions processed by the LLaMA model. The LLaMA model produces specialized audio tokens, denoted as [AUDi] where $i \in \{1, 2, \ldots, K\}$ (with K as a hyperparameter representing the number of special audio tokens added to the LLaMA model's vocabulary) when processing input captions. The special audio tokens serve as signaling indicators, aiding the model in determining whether to generate text+music or solely text. In training, these audio tokens are added to the end of the text output in datasets requiring music output. During inference, if the MuMu-LLaMA model generates audio tokens, downstream music decoders (MusicGen/AudioLDM 2) will perform music generation, otherwise, solely text will be produced.

The hidden embeddings corresponding to these audio tokens from the last layer of the LLaMA model is then input into the output projection layer, generating the conditional embedding for the Music Generation model. The MUCaps dataset is utilized to train this stage, with captions serving as inputs to the model and the target output tokens set as the special audio tokens.

Assuming a total of N tokens generated by the LLaMA model, where [AUDi] with $i \in \{0, 2, \ldots, 7\}$ constitutes the last 8 tokens. The hidden embeddings size is $(1, N, 4096)$, and the last 8 tokens are extracted along dimension -1, resulting in an input embedding size of the Output Projection layer as $(1, 8, 4096)$. The output size from the projection layer varies based on the Music Generation model: for AudioLDM2, it is $(1, 512)$, and for MusicGen, it is $(512, 768)$.

In the final training stage, the LoRA training strategy is employed to train the LLaMA model, concurrently fine-tuning the Multi-modal Understanding Adapter and Output Projection layer. This stage utilizes datasets including Alpaca, MusicQA, MUImage, MUVideo, and MUEdit. To signal the MuMu-LLaMA model to generate both music and text, the output text in MUImage, MUVideo, and MUEdit datasets is extended with the special audio tokens.

Table 6: **Evaluation Dataset Statistics**. The number of instructions in the evaluation dataset and total hours of music files in the dataset

Dataset	Number of Instructions	Hours of Music
MUCaps Eval	4000	265.35
MUImage Eval	2500	6.94
MUVideo Eval	2500	6.94
MUEdit Eval	2000	5.55

Reasoning for Training Strategy

Using the initial training stage, the model undergoes training with the objective of comprehending diverse modalities by leveraging extensive captioning datasets for music, image, and video. The subsequent training stage focuses on refining the LLaMA model's capability to condition music generation based on input captions.

These dual training stages equip the model with the ability to grasp various modalities and generate distinct music conditions based on input captions. This proficiency significantly contributes to the final training stage. In this ultimate phase, the model leverages the trained Multi-modal Understanding Adapters and Output Projection layers to bridge the gap between them, honing the LLaMA model's skills through the utilization of multi-modal music generation and music understanding datasets.

Model Training Parameters

We conduct training for the three stages of our model, employing 5, 5, and 2 epochs, respectively. The training process incorporates the following hyper-parameters: $N = 32$, $L = 6$, number of Audio Tokens $= 8$, and $lr = 10^{-4}$. This choice of hyper-parameters, coupled with our training strategy, allows for the effective use of a reduced number of epochs and a smaller dataset during the final stage of model training.

Model Evaluation

In this section, we elaborate on the datasets employed to assess the various capabilities of the MuMu-LLaMA model, followed by a discussion of the evaluation metrics utilized.

Evaluation Datasets

For each of the training datasets generated—MUCaps, MUImage, MUVideo, and MUEdit—we create a corresponding evaluation set. The methodology employed for generating the evaluation dataset mirrors that of the training dataset generation. Detailed statistics for the evaluation datasets are provided in Table 6. However, for the MUEdit dataset, the evaluation set could not be utilized for model evaluation due to the unavailability of code bases and trained checkpoints for InstructME(Han et al. 2023) and AUDIT(Wang et al. 2023). Consequently, we resort to utilizing samples from InstructME's demo website, which includes samples from both AUDIT and InstructME, to assess our model's performance. For evaluating the MuMu-LLaMA's music understanding capabilities we utilize the evaluation split of the MusicQA dataset.

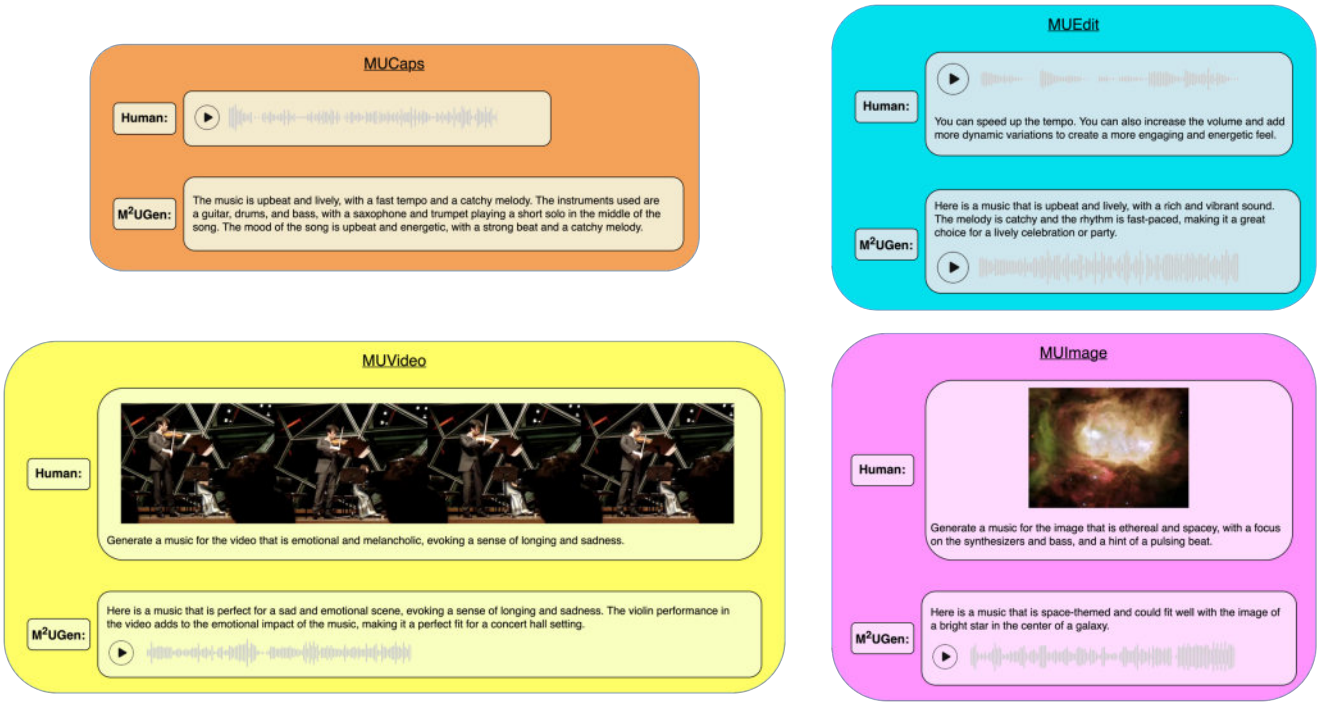


Figure 4: **Music Oriented Dataset.** Examples from the MUCaps, MUEdit, MUImage and MUVideo datasets used to train the MuMu-LLaMA model.

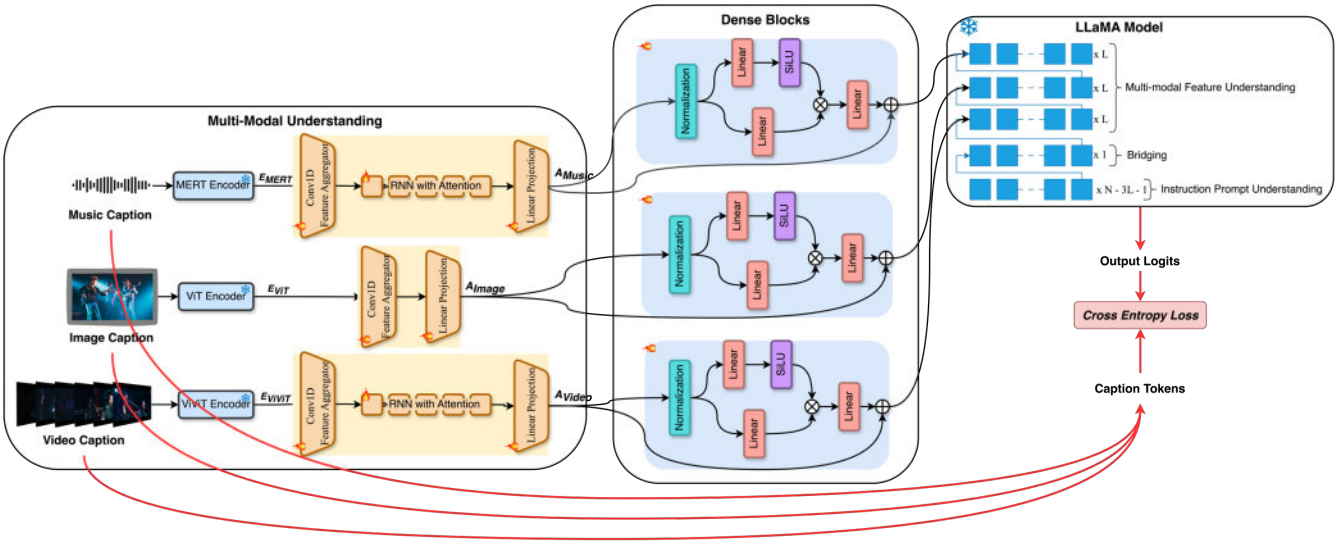


Figure 5: **Training Stage 1**: The Multi-modal Understanding Adapters are trained to integrate multi-modal features into the different layers of the LLaMA model.

Evaluation Metrics

To assess music question answering, we adopt the metrics employed in (Liu et al. 2023c), namely BLEU (B-U) (Papineni et al. 2002b), METEOR (M-R) (Banerjee and Lavie 2005b), ROUGE$_L$ (R-L) (Lin 2004b), and BERT-Score (BERT-S) (Zhang* et al. 2020b). These metrics are widely used for evaluating text generation. For all music generation tasks, we employ Fréchet Audio Distance (FAD) (Kilgour et al. 2019) and Kullback-Leibler divergence (KL), as these metrics are commonly utilized to assess the quality of generated audio. In addition to these general metrics, task-specific metrics are applied for each of the music generation tasks, namely Text-to-Music, Image-to-Music, Video-to-Music, and Music Editing.

In the context of Text-to-Music, we employ the CLAP(Wu et al.

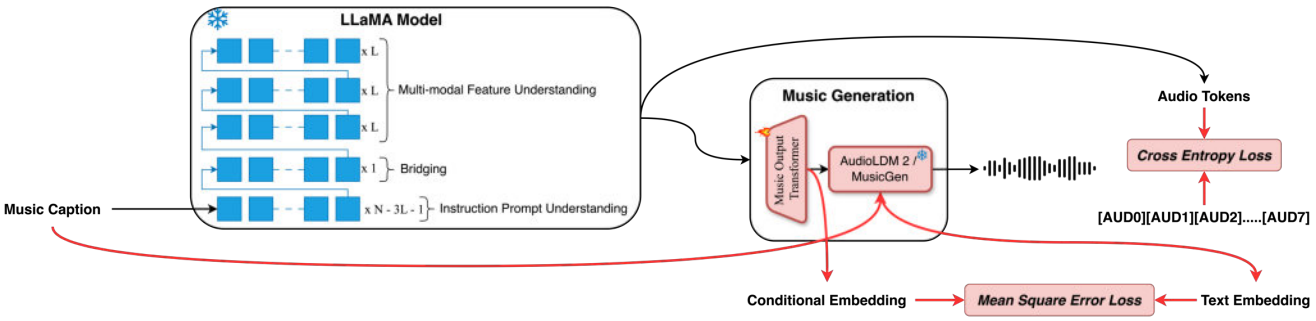


Figure 6: **Training Stage 2**: The Output Projection Layer is trained to generate the conditioning embedding for the MusicGen/AudioLDM 2 model.

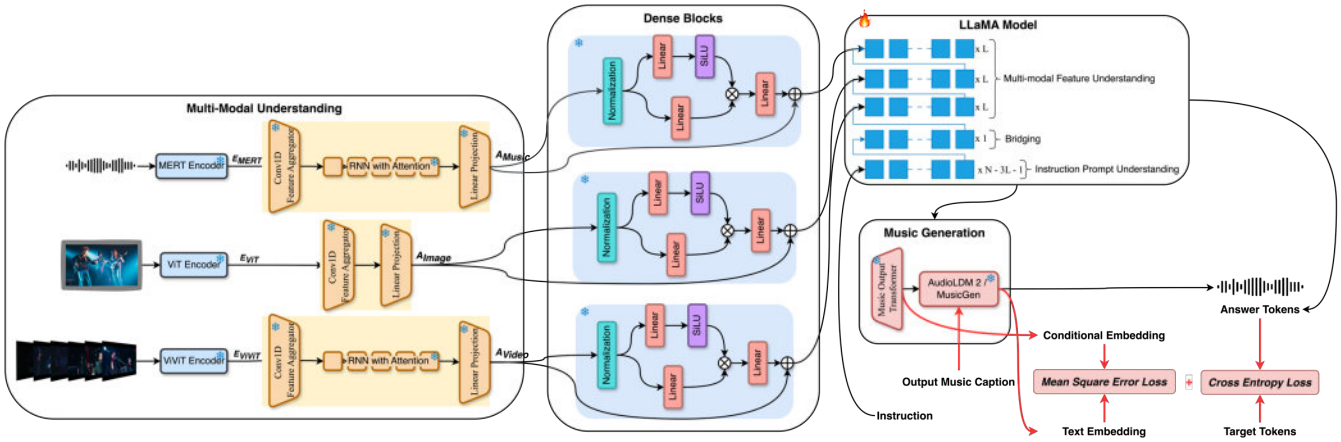


Figure 7: **Training Stage 3**: The Multi-modal Understanding Adapter and Output Projection Layer are fine-tuned while the LoRA-enabled LLaMA model is trained in this stage.

2023b) score, calculated by determining the cosine similarity between the CLAP embedding for the generated music and the text input:

$$CLAPScore(M, T) = \max(100 \times \cos(E_M, E_T), 0)$$

Here, M represents the generated music, T denotes the text input, and E_M, E_T represent the CLAP embeddings for the music and text, respectively.

For the Music Editing task, we leverage the Log Spectral Distance (LSD) to assess the disparity between the generated music and the target music. This metric facilitates the evaluation of whether the frequencies in the generated music, post-editing, align with those in the target music.

For Image-to-Music and Video-to-Music tasks, we introduce the ImageBind(Girdhar et al. 2023) Ranking (IB Rank), akin to the CLAP score, to quantify the alignment between the input modality and the generated music. Considering N distinct models producing N music files, we generate ImageBind embeddings for the music files, denoted as $E_{M1}, E_{M2}, \ldots, E_{MN}$, as well as the ImageBind embedding for the input modality, denoted as $E_{I/V}$. The embeddings for the music files are ranked based on their cosine similarity to $E_{I/V}$. Subsequently, after ranking all the samples in the evaluation set, the ImageBind Ranking is computed by calculating the ranking score using the individual rankings.

Using these evaluation metrics, we are able to evaluate the MuMu-LLaMA model against other state-of-the-art models for the different tasks.

Model Demonstration

In this section, we present screenshots of the MuMu-LLaMA model demo, illustrating various capabilities of the model.

Figures 9, 8, and 10 showcase the MuMu-LLaMA model's ability to generate music directly from textual prompts and draw inspiration from images and videos, both with and without textual guidance. Figure 11 exemplifies MuMu-LLaMA's proficiency in music editing guided by natural language prompts. Additionally, Figures 12, 13, and 14 illustrate the utilization of MuMu-LLaMA's editing capabilities to further refine music generated from different modalities. Collectively, the MuMu-LLaMA model proves to be a robust framework for Music Understanding, Generation, and Editing.

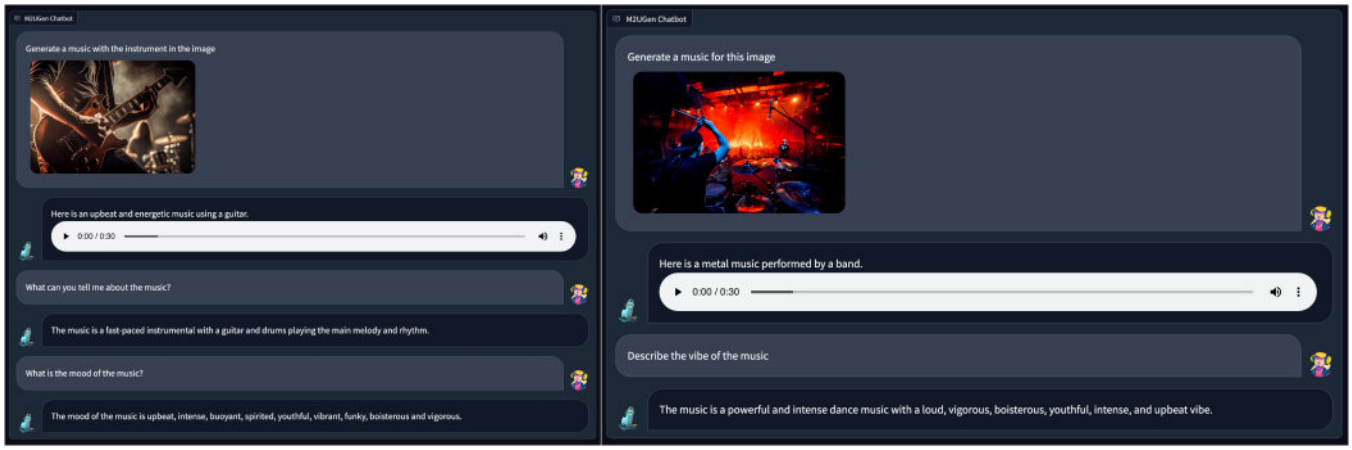


Figure 8: **Image-To-Music Generation and Understanding:** The MuMu-LLaMA model is capable of generating music for images and also answering questions regarding the generated music.

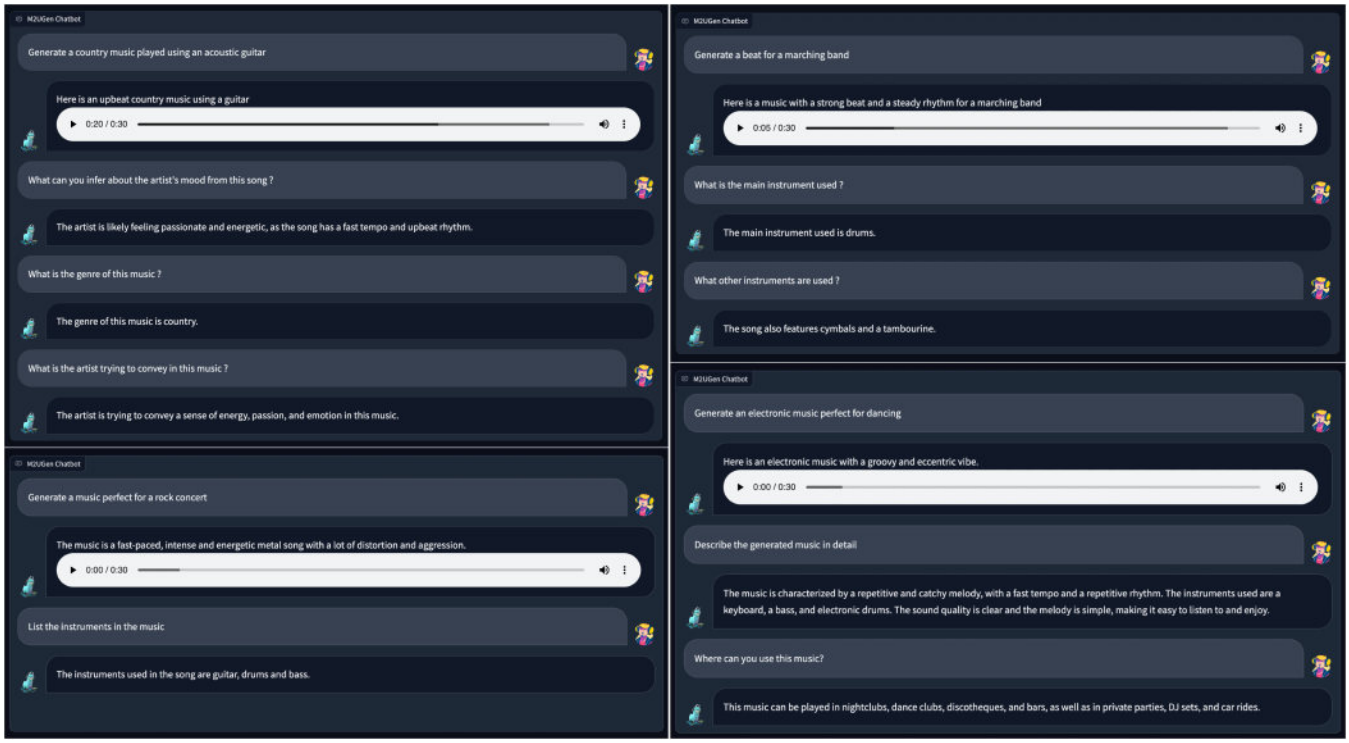


Figure 9: **Text-To-Music Generation and Understanding:** The MuMu-LLaMA model is capable of generating music from text prompts and also answering questions regarding the generated music.

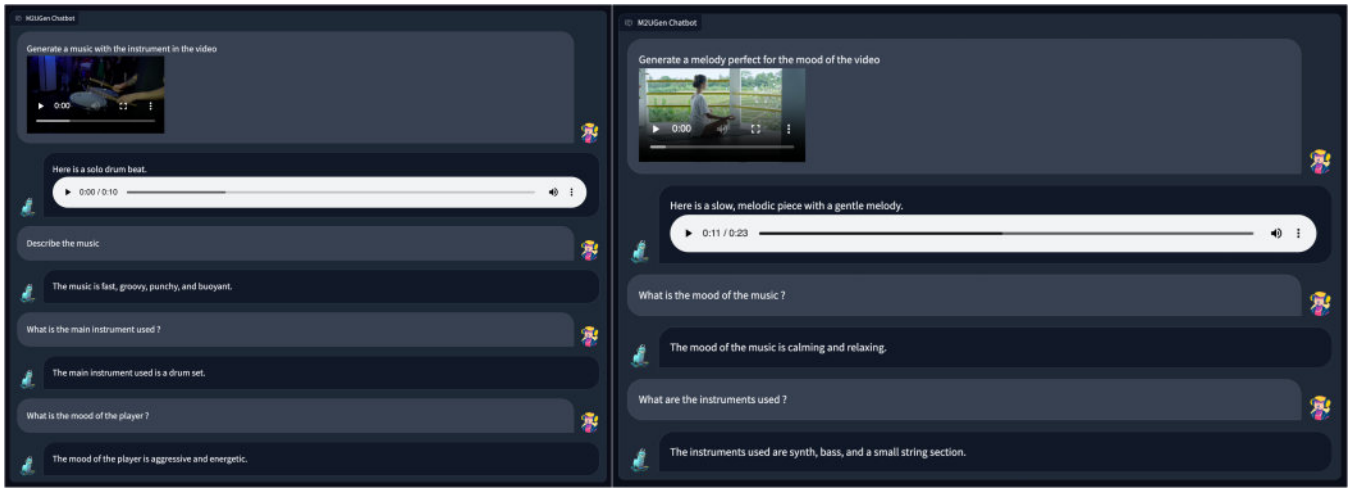


Figure 10: **Video-To-Music Generation and Understanding:** The MuMu-LLaMA model is capable of generating music for videos and also answering questions regarding the generated music.

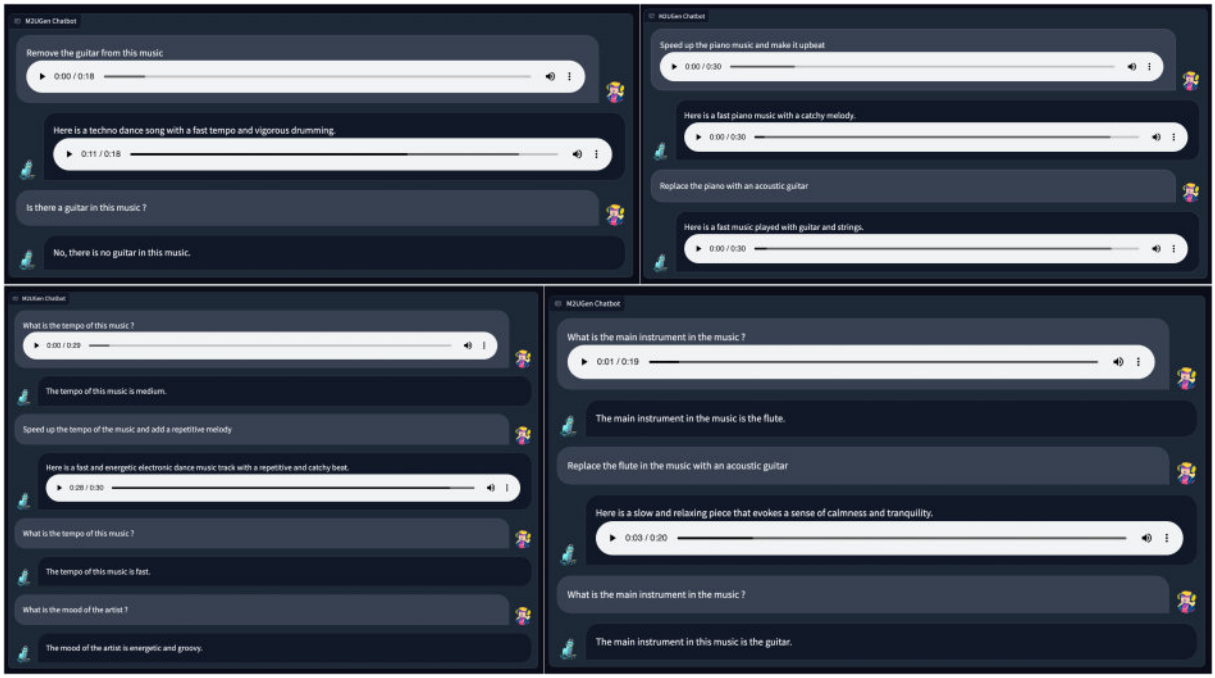


Figure 11: **Music Editing and Understanding:** The MuMu-LLaMA model is capable of editing input music based on natural language prompts and also answering questions regarding the generated music.

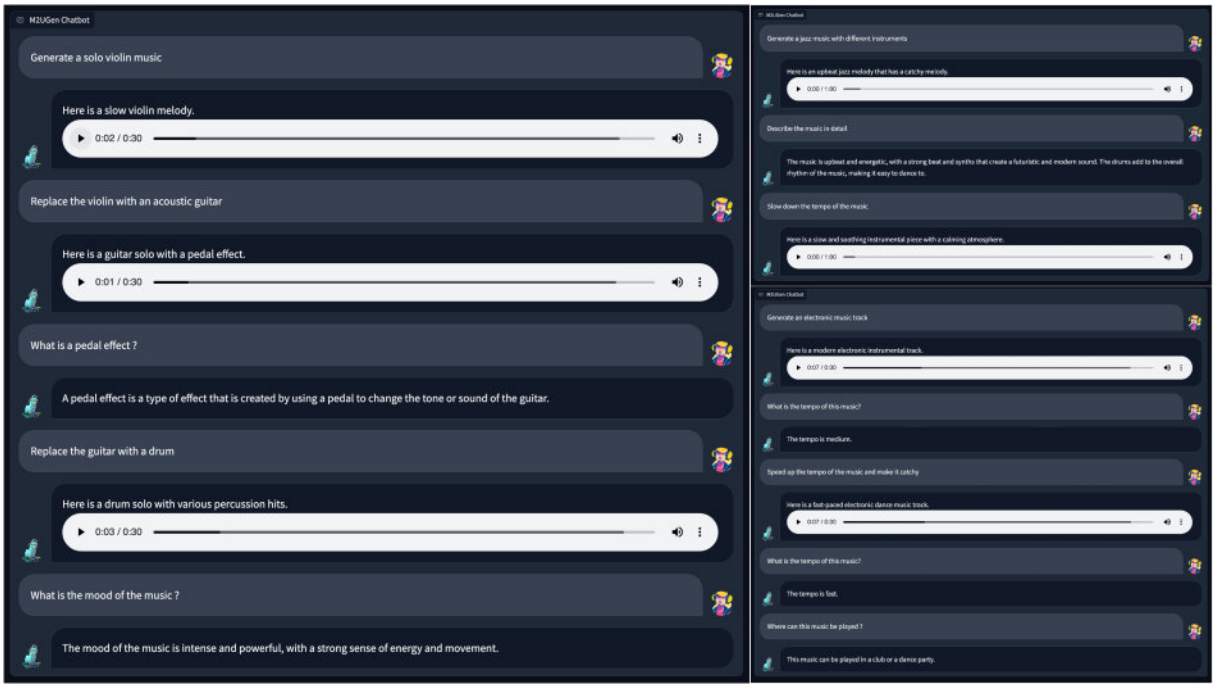


Figure 12: **Text-To-Music Generation and Understanding + Music Editing:** The MuMu-LLaMA model is capable of generating music from text prompts, answering questions regarding the generated music and also editing the generated music using Natural Language prompts.

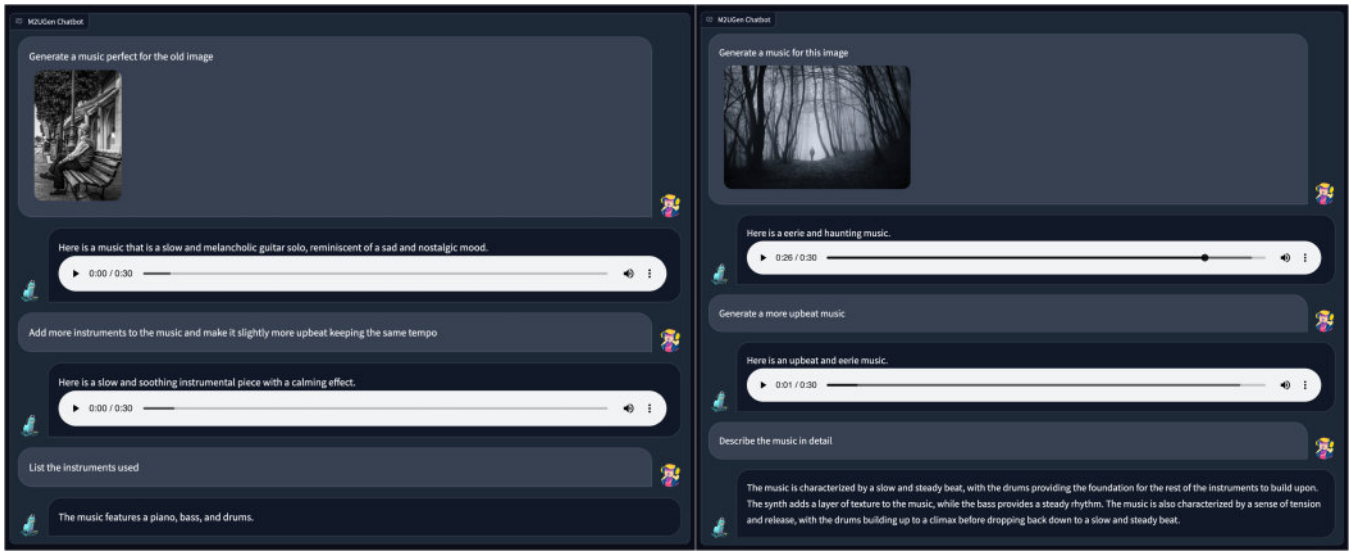


Figure 13: **Image-To-Music Generation and Understanding + Music Editing:** The MuMu-LLaMA model is capable of generating music for images, answering questions regarding the generated music and also editing the generated music using Natural Language prompts.

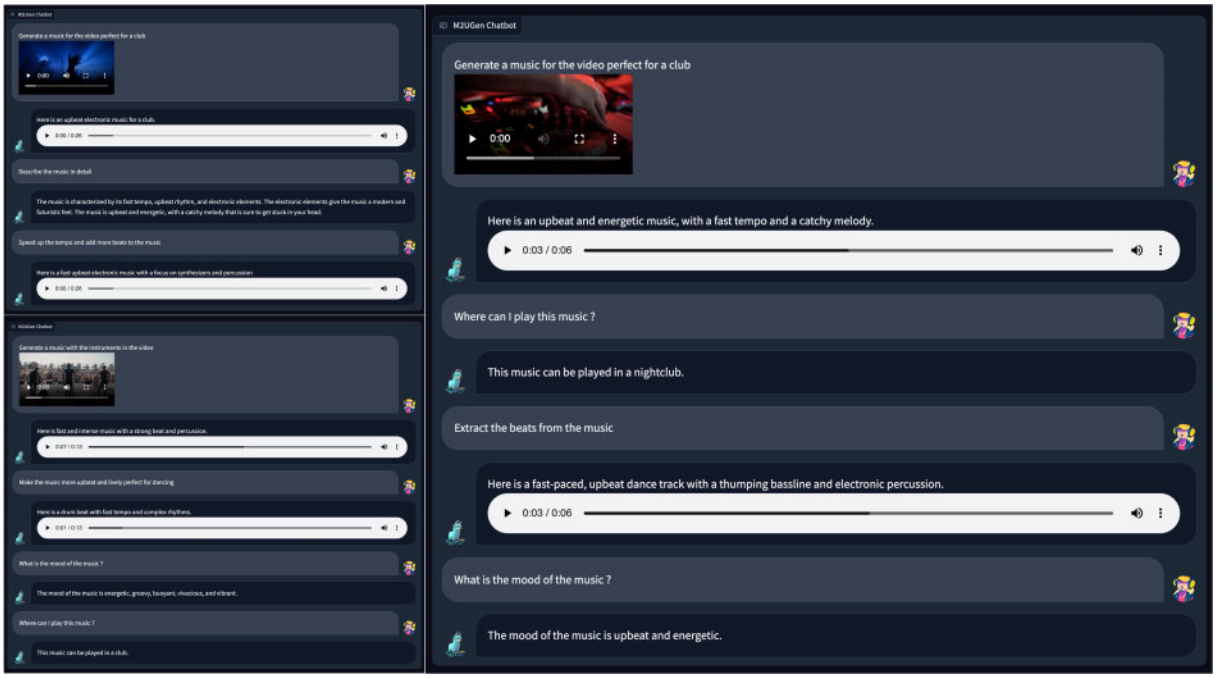


Figure 14: **Video-To-Music Generation and Understanding + Music Editing:** The MuMu-LLaMA model is capable of generating music for videos, answering questions regarding the generated music and also editing the generated music using Natural Language prompts.