DeepResonance: Enhancing Multimodal Music Understanding via Music-centric Multi-way Instruction Tuning

Anonymous ACL submission

Abstract

Recent advancements in music large language models (LLMs) have significantly improved music understanding tasks, which involve the model's ability to analyze and interpret various musical elements. These improvements primarily focused on integrating both music and text inputs. However, the potential of incorporating additional modalities such as images, videos and textual music features to enhance music understanding remains unexplored. To bridge this gap, we propose DeepResonance, a multimodal music understanding LLM fine-tuned via multi-way instruction tuning with multi-way aligned music, text, image, and video data. To this end, we construct Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T, three 4-way training and evaluation datasets designed to enable DeepResonance to integrate both visual and textual music feature content. We also introduce multi-sampled ImageBind embeddings and a pre-alignment Transformer to enhance modality fusion prior to input into text LLMs, tailoring DeepResonance for multi-way instruction tuning. Our model achieves state-of-the-art performances across six music understanding tasks, highlighting the benefits of the auxiliary modalities and the structural superiority of DeepResonance. We plan to open-source the models and the newly constructed datasets.

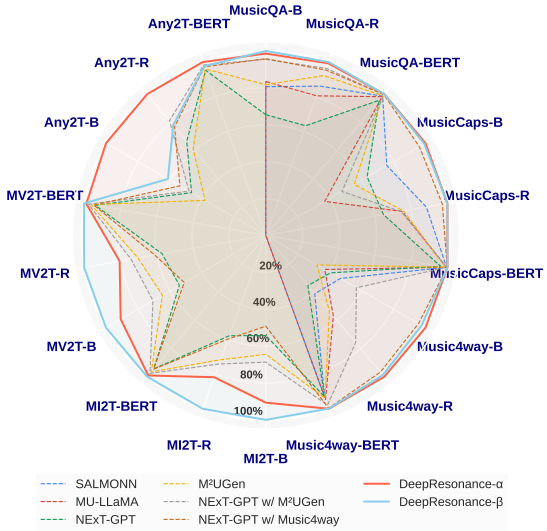

Figure 1: Performance overview of DeepResonance and related models. "Music4way," "MI2T," "MV2T," and "Any2T" refer to the Music4way-MusicCaps, Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T datasets. The metrics "B," "R," and "BERT" represent BLEU-1, ROUGE-L, and BERTScore. All values have been normalized based on maximum.

1 Introduction

> *"Music gives a soul to the universe, wings to the mind, flight to the imagination, and life to everything."*
>
> — Plato

Different modalities are often interwoven. In the context of music, humans typically experience music alongside complementary textual and visual signals, such as lyrics, the composition of a piece, live performances, or the arrangement of instruments in a band. These additional modalities significantly influence how humans perceive and understand music. Therefore, it is reasonable to assume that training a model to incorporate other modalities—such as images, videos, and textual music features can enhance its performance on music understanding tasks (Manco et al., 2021; Gardner et al., 2024; Agostinelli et al., 2023; Liu et al., 2024).

In this work, we introduce the concept of multimodal music understanding, where multiple modality signals are leveraged to enhance the perception of music. We implement this concept through multi-way instruction tuning, inspired by code-switched (Song et al., 2022) and multi-way multilingual translation (Fan et al., 2021) that extend translation models beyond bilingual pairings. For multimodal music understanding, we establish re-

lationships that go beyond the conventional pairing of music and text modalities commonly seen in existing music large language models (LLMs) (Doh et al., 2023; Liu et al., 2024; Hussain et al., 2023; Gardner et al., 2024; Zhao et al., 2024).

We present **DeepResonance**, a multimodal music understanding LLM trained on music-centric multi-way data that integrates the music, text, image, and video modalities. Specifically, we construct two 4-way training datasets, **Music4way-MI2T** and **Music4way-MV2T**, where the source data comprises music, images/videos, and textual descriptions and instructions. The target data includes textual descriptions enriched with multimodal information, such as music, images/videos, and low-level music features, including tempo, chords, key, and downbeats. Using these datasets, we develop DeepResonance based on the NExT-GPT (Wu et al., 2024) architecture. To enhance multimodal music understanding tasks, we propose two key modifications to the backbone model. The first involves **multi-sampled ImageBind embeddings**, which are designed to retain richer information of music, image, and video modalities from the ImageBind encoders (Girdhar et al., 2023), thereby fostering deeper interaction with the music modality and improving multimodal music understanding. The second is a **pre-alignment Transformer**, a module aimed at pre-adapting different modalities to each other before feeding them into the text LLM module. This component is particularly effective in simultaneously handling multimodal inputs of music, text, images, and videos. These innovations collectively advance the model's ability to integrate and process diverse multimodal signals for enhanced music understanding.

We evaluate DeepResonance on three conventional music understanding tasks (i.e., music + text (instruction)$\rightarrow$ text) and three multimodal music understanding tasks (i.e., music + image/video + text (instruction) $\rightarrow$ text (multimodal-enriched)). The former includes two existing benchmarks, MusicQA (Liu et al., 2024) and MusicCaps (Agostinelli et al., 2023), along with our constructed Music4way-MusicCaps, which cover captioning and question-answering tasks for music understanding. The latter evaluation for multimodal music understanding uses the test splits of Music4way-MI2T and Music4way-MV2T, as well as the newly introduced **Music4way-Any2T** to assess the model's robustness.

As shown in Fig. 1, DeepResonance models with different configurations (α and β) consistently outperform related models across all six downstream tasks in supervised settings, demonstrating the effectiveness of our proposed multi-way datasets and model architecture components for music understanding tasks. Additionally, we conduct zero-shot evaluations to assess the model's generalization to unseen datasets and perform ablation studies to evaluate the impact of each proposed component in different downstream task settings. The contributions of this work can be summarized as follows:

- We introduce the Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T datasets, enabling music, text, image, and video integration for music understanding.
- We propose multi-sampled ImageBind embeddings and a pre-alignment Transformer to enhance multimodal fusion for music LLMs.
- Our DeepResonance outperforms existing music LLMs across six downstream tasks, showing the effectiveness of our proposed datasets and models, which will be open-sourced.

2 Related Work

Music Understanding is an emerging topic that builds upon the foundational research efforts in music information retrieval (MIR), which traditionally focused on low-level music feature recognition tasks, such as identifying tempo, chords, keys, and instruments (Faraldo et al., 2016; Pauwels et al., 2019; Gururani et al., 2019; Schreiber et al., 2020). Early work in this area was centered on basic tagging tasks, such as determining the genre or version of a piece of music (Tzanetakis, 2001; Won et al., 2021; Yesiler et al., 2021). Over time, the focus shifted to high-level understanding tasks that require a more comprehensive interpretation of the content, sentiment, and insights conveyed by music. These tasks include captioning, reasoning, question answering, and tool using (Manco et al., 2021; Gardner et al., 2024; Agostinelli et al., 2023; Liu et al., 2024; Deng et al., 2024; Zhao et al., 2024).

Multimodal Instruction Tuning and Music Foundation Models: Recently, multimodal pre-training has successfully bridged image, audio, and video modalities to text LLMs (Tang et al., 2024a; Wu et al., 2024; Gong et al., 2024; Tang et al., 2024b) through multimodal instruction tuning (Liu et al., 2023; Zhao et al., 2023) or universal multimodal

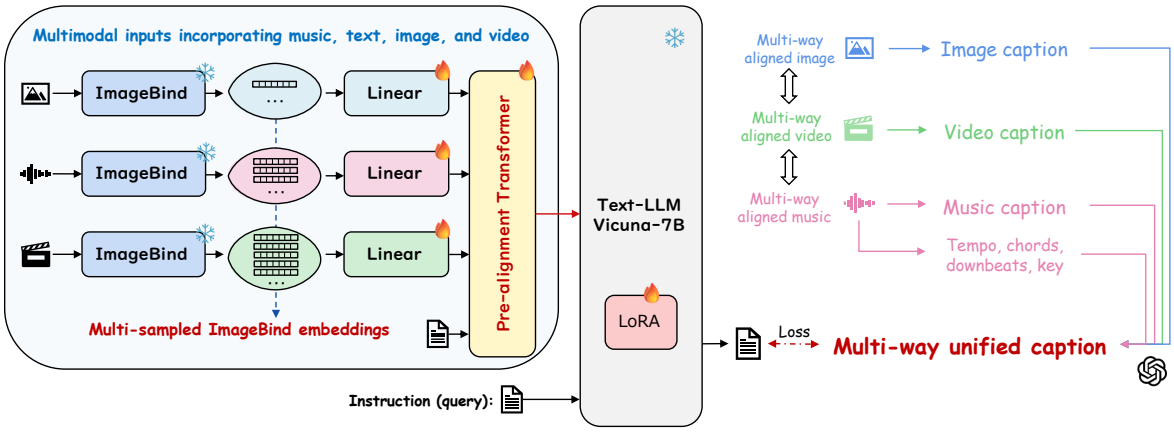

Figure 2: **Overview of DeepResonance.** It integrates the "multi-way unified caption" as a target, along with "multi-sampled ImageBind embeddings" and a "pre-alignment Transformer" as novel architectural components.

embedding space encoders (Girdhar et al., 2023; Zhu et al., 2024). However, few studies have focused on the music modality. MU-LLaMA (Liu et al., 2024) was among the first to instruction-tune LLaMA models (Touvron et al., 2023) for the music domain, while LLark (Gardner et al., 2024) extended music LLMs to support a wide range of tasks, including captioning, reasoning, and low-level music feature recognition. M^{2}UGen (Hussain et al., 2023) introduced music generation modules built on MU-LLaMA, leveraging newly constructed music-centric datasets for instruction fine-tuning. OpenMU (Zhao et al., 2024) unified existing music understanding datasets, curating a comprehensive benchmark for evaluating music + text $\rightarrow$ text tasks. Other models, such as MusCaps (Manco et al., 2021), LP-MusicCaps (Doh et al., 2023), and MusiLingo (Deng et al., 2024), were designed for task-specific purposes.

Our work differs from the previous studies on music understanding and music foundation models in three aspects: (1) We are the first to apply instruction-tuning on music foundation models using multi-way data that integrates multiple modalities, shifting the paradigm from the traditional music + text $\rightarrow$ text approach to the music + image/video + text $\rightarrow$ text (multimodal-enriched) framework. (2) We are the first to access the generalization of music LLMs to multi-way multimodal inputs using our newly curated Music4way-Any2T dataset and zero-shot evaluation on out-of-domain benchmarks. (3) We propose multi-sampled ImageBind embeddings and pre-alignment Transformer that significantly impact multimodal music understanding tasks, delivering state-of-the-art music LLMs optimized for different downstream tasks.

3 Multi-way Instruction Tuning

We focus on the integration of multi-way information into each data for music understanding tasks, which conventionally contains only music and text.

3.1 Music-centric Multi-way Datasets Construction

M^{2}UGen (Hussain et al., 2023) pioneered multi-modal dataset construction for instruction-tuning music LLMs, creating music-centric pairs with text, images, videos, and other music for captioning, editing, and generation tasks. Drawing inspiration from studies on code-switched and multi-way translation (Song et al., 2022; Fan et al., 2021), which break away from English-centric bilingual pairing relationships, we propose extending traditional music LLM fine-tuning by incorporating multi-way relationships among music and other modalities, including image, video, and text. Building upon the music-centric paired dataset construction pipeline of M^{2}UGen, we expand it with multi-way relationships to create multimodal-enriched training data.

Music4way: Building upon the AudioSet (Gemmeke et al., 2017) music clips filtered by M^{2}UGen[1], we curate a multi-way aligned dataset, comprising a total of 172.57 hours of music.[2] Each video-music pair from AudioSet is processed as shown in Fig. 3 (left). First, we randomly extract a video frame to serve as the aligned image for the video and music. Next, BLIP (Li et al., 2022), VideoMAE (Tong et al., 2022), and MU-LLaMA (Liu et al., 2024) are used to generate captions for the image, video,

[1] https://github.com/shansongliu/M2UGen

[2] Note that in M^{2}UGen, each music clip is paired with only one other modality during training.

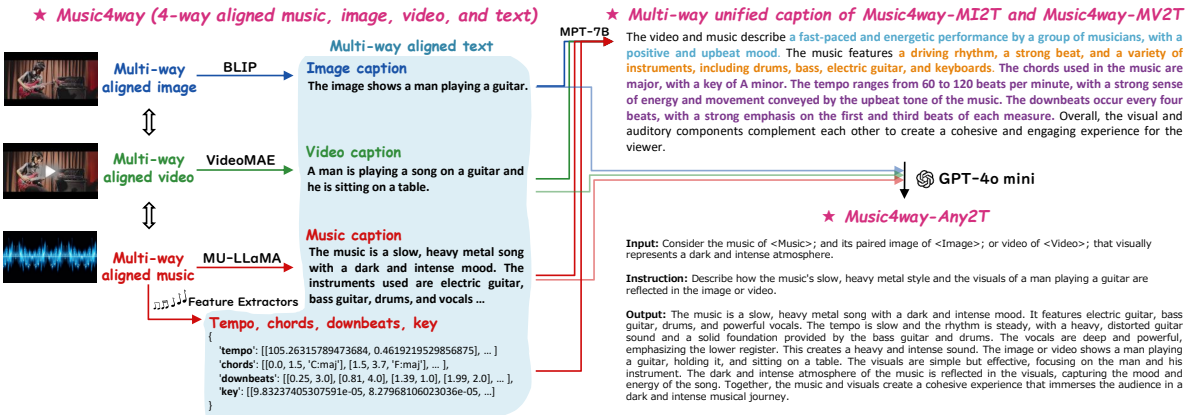

Figure 3: **Multi-way Instruction Tuning Data Construction.** Based on AudioSet, we construct Music4way (M+T→T, I+T→T, V+T→T), Music4way-MI2T (M+I+T→T), Music4way-MV2T (M+V+T→T), and Music4way-Any2T (M+I/V+T→T) for instruction tuning and evaluation. (M: music; I: image; V: video; T: text)

and music, respectively. Following LLark (Gardner et al., 2024), we employ `madmom` (Böck et al., 2016) to extract low-level music features, including tempo, chords, downbeats, and key, representing these features in textual form. This process creates a 4-way alignment for each music-video pair, establishing relationships between music, image, video, and text. The text modality conveys information through captions or feature values derived from the other three modalities. The resulting Music4way dataset includes 59,128 training samples (164.24 hours) and 3,000 evaluation samples (8.33 hours).

3.2 Multi-way Instruction Tuning Data Construction

Building on the constructed Music4way dataset, we design instruction tuning data tailored for training and evaluating music LLMs. As fine-tuning an LLM requires data to be formatted strictly as source-target pairs, we develop two pipelines to transform the multi-way relationships among the four modalities in Music4way into source-target pairs, which ensure seamless integration of multi-modal information for LLM instruction tuning.

Music4way-MI2T and Music4way-MV2T: Inspired by VAST (Chen et al., 2023), which established multimodal connections between vision, audio, and subtitles to generate omni-modal captions for improving visual-text tasks, we extend this approach to the music domain, exploring its potential benefits for music understanding tasks. As shown in Fig. 3, we prompt the open-source MPT-7B (Team, 2023) model to combine captions from images, videos, and music, as well as textual music features, creating a multi-way unified caption. These captions are then paired with music and image/video inputs to form the source-target pairs required for fine-tuning LLMs. For the task of music + image to multi-way unified caption, we construct the Music4way-MI2T dataset. Similarly, for music + video inputs, we construct the Music4way-MV2T dataset. These datasets are designed to train the music understanding model to infer visual content and musical details (e.g., instruments, sentiments, and low-level features) directly from raw music and image/video files. The templates we utilized for prompting MPT-7B to generate the multi-way unified captions are provided in the Appx. C. As shown in Fig. 2, these datasets are then used to fine-tune music understanding models, with inputs comprising music and image/video, a fixed instruction template (see Appx. C), and outputs of multi-way unified captions. Additionally, we create corresponding test sets from the test split of Music4way to benchmark performance on multimodal music understanding tasks. (See Sec. 5.3)

Music4way-Any2T: To evaluate the robustness of multimodal music LLMs to diverse text inputs and queries, we introduce the Music4way-Any2T dataset. As shown in Fig. 3, this dataset features flexible inputs, allowing each modality to appear in any position. It includes instructions and outputs presented as diverse question-answer pairs, covering various aspects (e.g., visual or musical content) of the multi-way aligned data. To construct structured data with input, instruction, and output fields, we use GPT-4o mini[3], prompted with all textual information from the Music4way dataset alongside the multi-way unified caption. Detailed prompting

[3] https://platform.openai.com/docs/models

templates and data examples can be found in the Appx. D. The Music4way-Any2T dataset is used to benchmark the robustness and generalization capabilities of multimodal music understanding models, with results presented in Sec. 5.3.

4 Model Architecture Tailored for Multimodal Music Understanding

In this section, we introduce **DeepResonance**, our model designed to adapt general any-to-text LLMs for multimodal music understanding tasks, leveraging the multi-way datasets introduced in Sec. 3.2. We construct DeepResonance based on NExT-GPT (Wu et al., 2024), a general any-to-any multimodal LLM. This framework integrates an ImageBind encoder (Girdhar et al., 2023) to process inputs from the image, video, and audio modalities, a Vicuna-7B (Chiang et al., 2023) model as the LLM backbone, and linear adaptors to bridge ImageBind to the Vicuna model. The vanilla version of DeepResonance, like NExT-GPT, models the text sequence generation task as follows:

$$\mathcal{P}(w_n|\mathbb{X}_m, \mathbb{X}_v, \mathbb{X}_i, \mathbb{X}_t, \mathbb{Q}, \mathbb{W}_{1:n-1}) = \mathcal{LLM}(\mathcal{A}_m(\mathbf{e}_m), \mathcal{A}_v(\mathbf{e}_v), \mathcal{A}_i(\mathbf{e}_i), \{\mathbf{e}_t\}, \{\mathbf{e}_q\}, \{\mathbf{e}_w\}_{1:n-1}) \quad (1)$$

where $\mathbb{W} = \{w_1, w_2, \ldots, w_n\}$ represents the text sequence to be generated, and $\mathbb{X}_m$, $\mathbb{X}_v$, $\mathbb{X}_i$ and $\mathbb{X}_t$ denote the patch-level multimodal and text tokens for music, video, image, and text, respectively. $\mathbb{Q}$ indicates the query (i.e., instruction) input to the model. $\mathcal{LLM}$ and $\mathcal{A}_*$ ($* \in \{m, v, i\}$) represent the Vicuna-7B LLM and the linear adaptors for each modality. $\mathbf{e}_*$ ($* \in \{m, v, i\}$) denotes the embedding of music, video, and image produced by ImageBind, while $\mathbf{e}_\#$ ($\# \in \{t, q, w\}$) are the LLM's text embeddings of input, query, and output.

However, the pooled single embedding from ImageBind may fail to capture the detailed information required to effectively interact with other modalities in downstream music understanding tasks, particularly for music and video modalities. This limitation arises because such multimodal encoders prioritize coarse-grained representations for cross-modal retrieval tasks (Balaji et al., 2022). Additionally, as NExT-GPT relies solely on modality-specific adaptors to map each modality into LLM's embedding space, it may not effectively model interactions between modalities. These interactions were never pre-trained, and the LLM itself only employs uni-directional attention (Zhou et al., 2024).

Dataset	Used for	In→out modality
COCO	Train stage 1	I+T→T
Music4way	Train stage 2	I+T→T
Music4way	Train stages 1 & 2	V+T→T
Music4way	Train stages 1 & 2	M+T→T
Alpaca	Train stage 2	T→T
MusicQA	Train stage 2	M+T→T
MusicCaps	Train stage 2	M+T→T
Music4way-MI2T	Train stage 2	M+I+T→T
Music4way-MV2T	Train stage 2	M+V+T→T

Table 1: Overview of training data. M: Music; I: Image; V: Video; T: Text.

To address these challenges, we propose two modules for music LLMs: **multi-sampled ImageBind embeddings** and **pre-alignment Transformer**, as shown in Fig. 2. The former leverages embeddings from multiple clips sampled by ImageBind without pooling, while the latter introduces a Transformer to globally integrate and align information across modalities before feeding it into the LLM. Formally, the proposed model is defined as

$$\mathcal{P}(w_n|\mathbb{X}_m, \mathbb{X}_v, \mathbb{X}_i, \mathbb{X}_t, \mathbb{Q}, \mathbb{W}_{1:n-1}) = \mathcal{LLM}(\mathcal{T}(\mathcal{A}_m(\{\mathbf{e}_m\}_{1:\mathrm{N}_m}), \mathcal{A}_v(\{\mathbf{e}_v\}_{1:\mathrm{N}_v}), \mathcal{A}_i(\{\mathbf{e}_i\}_{1:\mathrm{N}_i}), \{\mathbf{e}_t\}), \{\mathbf{e}_q\}, \{\mathbf{e}_w\}_{1:n-1}). \quad (2)$$

Here, $\mathcal{T}$ represents the pre-alignment Transformer, and $\mathbf{e}_*$ ($* \in \{m, v, i\}$) from Eq. 1 is reformulated as $\{\mathbf{e}_*\}_{1:\mathrm{N}_*}$ to incorporate multi-sampled ImageBind embeddings for each modality, where N_* denotes the number of sampled clips for a given modality. With these components, the multi-sampled ImageBind embeddings preserve finer-grained information for each modality, which is expected to enhance music understanding tasks (Sec. 5). The pre-alignment Transformer uses bidirectional attention to encode cross-modal dependencies, enhancing the interactions between music and other modalities, ultimately aiming to improve multimodal music understanding tasks (Sec. 5.3).

5 Experiments and Results

Following the training strategy of NExT-GPT and M^{2}UGen, we train DeepResonance in two stages. Table 1 summarizes all datasets used for training, with training details provided in Appx. B. Subsequently, we evaluate DeepResonance across three music understanding tasks and three multimodal music understanding tasks in supervised settings. Additionally, we assess the model's zero-shot performance on out-of-domain datasets and conduct

Model	MusicQA			MusicCaps			Music4way-MusicCaps		
	B-1	R-L	BERTS	B-1	R-L	BERTS	B-1	R-L	BERTS
SALMONN (Tang et al., 2024a)	†28.7	†35.4	†90.3	†19.7	†19.1	†86.9	19.1	20.0	87.0
MU-LLaMA (Liu et al., 2024)	†29.7	†33.1	†89.9	*†9.6	*†16.2	*†86.8	15.1	27.6	88.3
OpenMU (Zhao et al., 2024)	†24.5	†25.5	†88.6	†23.9	†19.4	†86.6	–	–	–
MusiLingo sft. w/ MusicCaps (Deng et al., 2024)	–	–	–	–	†**21.7**	†86.8	–	–	–
NExT-GPT (Wu et al., 2024)	23.3	26.0	87.6	16.5	14.0	84.0	16.6	17.2	86.7
M^{2}UGen (Hussain et al., 2023)	†29.1	†37.9	†90.5	*†14.4	*†16.4	*†86.5	*†13.1	*†26.0	*†87.6
NExT-GPT w/ M^{2}UGen	†34.0	†39.6	†91.2	12.4	16.1	86.9	*†23.2	*†36.7	*†91.3
NExT-GPT w/ Music4way	†34.1	†39.2	†91.3	†25.0	†21.0	†87.2	†39.1	†46.8	†93.0
DeepResonance-α (ours)	†**35.1**	†**40.8**	†**91.6**	†**26.0**	†**21.6**	†**87.3**	†**40.9**	†**48.4**	†**93.3**
DeepResonance-β (ours)	†**35.6**	†**41.1**	†**91.6**	†**25.8**	†**21.6**	†**87.3**	†**39.9**	†**47.8**	†**93.2**

Table 2: **Results on MusicQA, MusicCaps, and Music4way-MusicCaps.** The top two performances are highlighted in **bold**. "*" denotes the test data was included in the corresponding model's training set. "†" indicates that the test set's train split is included in the corresponding model's training data, representing a supervised setting. "B-1", "R-L", and "BERTS" denote BLEU-1, ROUGE-L, and BERTScore, respectively.

an ablation study to demonstrate the effectiveness of each proposed component. Results are reported using BLEU (Papineni et al., 2002),[4] ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020).

5.1 Baselines and Ours

Below are the baseline models and our models that we compare in this section.

SALMONN (Tang et al., 2024a): A robust baseline model for audio understanding. It leverages a variety of audio datasets for training, including AudioCaps (Kim et al., 2019), WaveCaps (Mei et al., 2024), MusicNet (Thickstun et al., 2017), etc.

MU-LLaMA (Liu et al., 2024): The first LLM instruction-tuned for music understanding with the MusicCaps and MusicQA as fine-tuning data.

NExT-GPT (Wu et al., 2024): The first any-to-any multimodal LLM trained on multimodal fine-tuning data, serving as the backbone for ours.

M^{2}UGen (Hussain et al., 2023): The first any-to-any multimodal LLM tailored to the music domain, trained on newly curated data derived from the music split of AudioSet (Gemmeke et al., 2017).

NExT-GPT w/ M^{2}UGen: We train a NExT-GPT model on the same data as M^{2}UGen, excluding MusicCaps, as its test split was inadvertently included in M^{2}UGen's training data.

NExT-GPT w/ Music4way: We train a NExT-GPT model using the Music4way datasets (see Sec. 3.1), a more extensive and 4-way aligned dataset compared to M^{2}UGen's training data. We exclude Music4way-MI2T and Music4way-MV2T (Table 1), aiming to evaluate the benefits of them.

DeepResonance (Ours): The models introduced in Sec. 4, built upon NExT-GPT and trained with datasets as shown in Table 1. We introduce two DeepResonance variants: DeepResonance-α, trained without the pre-alignment Transformer, and DeepResonance-β, which includes it. The following sections examine their effectiveness, as the pre-alignment Transformer is specifically designed for multimodal music understanding tasks (Sec. 5.3).

We also include results from a task-specific music understanding model, **MusiLingo** (Deng et al., 2024), for comparison. We also report the performance of **OpenMU** (Zhao et al., 2024), a benchmark model for multiple music understanding datasets for supervised comparison. In zero-shot evaluations, OpenMU and MusiLingo serve as upper bounds for supervised performance, as they were directly trained on those benchmarks.

5.2 Music Understanding Tasks

Table 2 reports the performance of all models introduced in Sec.5.1 on music understanding tasks (music + text → text) using MusicQA(Liu et al., 2024), MusicCaps (Agostinelli et al., 2023), and our constructed Music4way-MusicCaps[5].

First, we observe that DeepResonance outperforms baseline models on three datasets, demonstrating the effectiveness of our proposed training datasets and model architecture. Second, we find that applying the same training data to the NExT-GPT framework yields a better performance than using the M^{2}UGen framework, suggesting that NExT-GPT is a more suitable backbone model. Third, by expanding M^{2}UGen's training data with our constructed Music4way dataset, we observe an improvement in performance, further validating

[4]Following prior studies, we primarily report BLEU-1, while BLEU-1 and BLEU are shown in Appx. E, F, G, and H.

[5]A music captioning subset of the Music4way test split.

Model	Music4way-MI2T			Music4way-MV2T			Music4way-Any2T		
	B-1	R-L	BERTS	B-1	R-L	BERTS	B-1	R-L	BERTS
NExT-GPT (Wu et al., 2024)	26.7	21.3	85.2	26.5	21.0	84.8	25.4	23.4	86.6
M^{2}UGen (Hussain et al., 2023)	*31.7	*26.4	*87.1	*31.7	*25.9	*86.8	*20.8	*21.5	*87.3
NExT-GPT w/ M^{2}UGen	*33.8	*27.3	*88.1	*34.6	*27.3	*88.1	*26.3	*28.5	*89.4
NExT-GPT w/ Music4way	24.2	22.0	85.4	25.0	22.5	85.5	29.4	27.0	88.4
DeepResonance-α (ours)	**44.6**	**30.1**	**89.4**	**44.5**	**29.6**	**89.2**	**54.7**	**35.1**	**90.9**
DeepResonance-β (ours)	**49.2**	**36.8**	**90.2**	**49.0**	**36.8**	**90.3**	**33.5**	**27.4**	**88.7**

Table 3: **Results on Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T.** The top two performances are highlighted in **bold**. "*" denotes the test data was included in the corresponding model's training set.

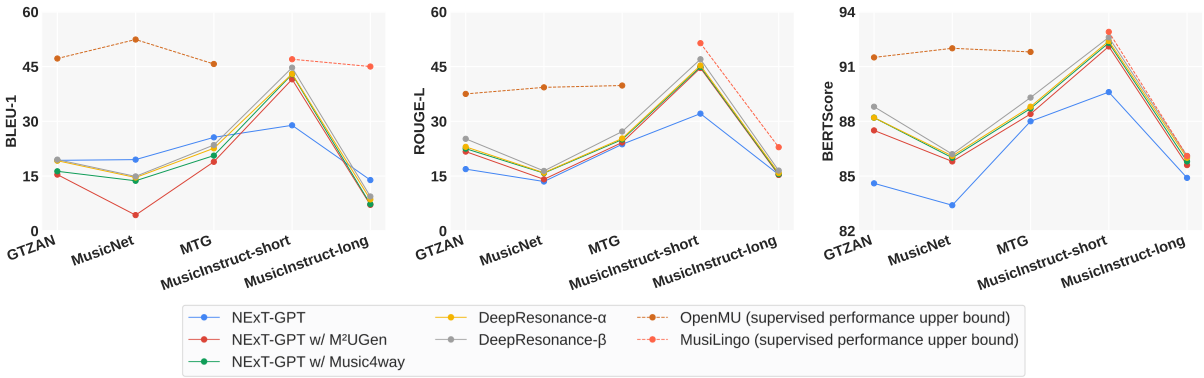

Figure 4: Zero-shot evaluation on GTZAN, MusicNet, MTG-Jamendo, MusicInstruct-short, and MusicInstruct-long.

the effectiveness of the Music4way dataset. Finally, the comparison between DeepResonance-α and DeepResonance-β shows a comparable performance, indicating that the pre-alignment Transformer is not crucial for music + text $\rightarrow$ text tasks. This aligns with expectations, as the pre-alignment Transformer is designed for multimodal music understanding, whereas these tasks involve only a single modality (music) alongside the text query.

5.3 Multimodal Music Understanding Tasks

Table 3 lists the results on our proposed multimodal music understanding tasks (music + image/video + text $\rightarrow$ text), including Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T. We compare the performance of DeepResonance with baseline models capable of processing multiple modalities—music, text, image, and video—such as the M^{2}UGen and NExT-GPT-based models.

First, we observe that Music4way-MI2T and Music4way-MV2T represent novel downstream tasks for which no existing baseline models are inherently equipped. Through supervised fine-tuning with our curated training data for these datasets, the DeepResonance models gain the ability to generate unified multimodal captions successfully. Second, on Music4way-Any2T, which features flexible inputs and open-ended question-answer pairs, the baseline models perform poorly. Their generalization remains weaker than DeepResonance models, highlighting their limitations in handling diverse input patterns. Third, comparing DeepResonance-α and DeepResonance-β, we find that the latter demonstrates a superior performance on Music4way-MI2T and Music4way-MV2T. This indicates that the pre-alignment Transformer, with its additional parameters, effectively integrates multiple modalities, thereby improving the supervised performance in structured multimodal music understanding tasks. However, DeepResonance-β exhibits reduced robustness on Music4way-Any2T, which highlights a trade-off, as the model's increased complexity may hinder its adaptability with limited instruction tuning data.

5.4 Zero-shot Evaluation

Fig. 4 present the zero-shot performance of DeepResonance and baselines on music understanding benchmarks, including GTZAN (Tzanetakis, 2001), MusicNet (Thickstun et al., 2017), MTG-Jamendo (Bogdanov et al., 2019), MusicInstruct-short, and MusicInstruct-long (Deng et al., 2024). We adopt the benchmark settings outlined in OpenMU-bench (Zhao et al., 2024), combining captioning and reasoning test sets where separate splits exist. All NExT-GPT-based models without

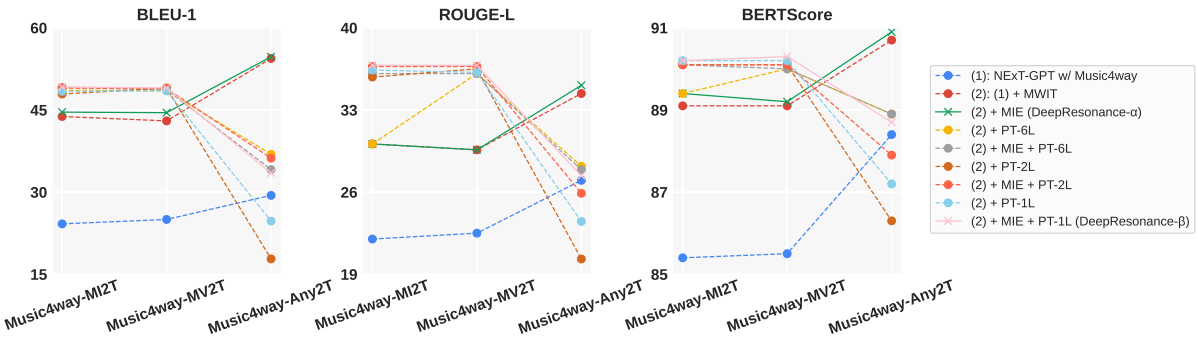

Figure 5: Ablation study on Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T. "PT-*L" indicates the number of layers used in the pre-alignment Transformer. See Appx. H for result details.

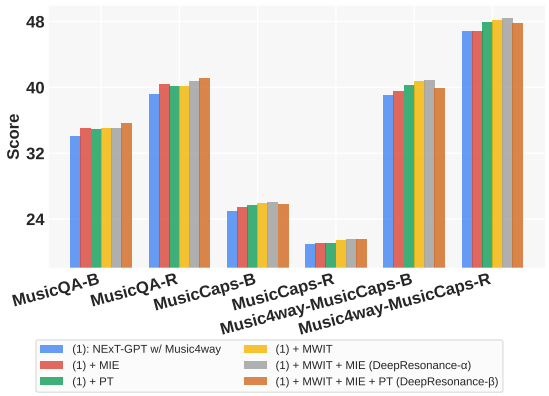

Figure 6: Ablation study on MusicQA, MusicCaps, and Music4way-MusicCaps. See Appx. H for result details.

exposure to test data during training are compared, to ensure the zero-shot configurations. Detailed results are provided in the Appx. G.

First, DeepResonance-β achieves the best performance across all five benchmarks in terms of ROUGE-L and BERTScore, demonstrating the effectiveness of the proposed training data and model architecture. Second, DeepResonance-β outperforms DeepResonance-α, highlighting the pre-alignment Transformer's effectiveness in improving inference on unseen data. Referring back to Sec. 5.3, we recommend DeepResonance-α for tasks with flexible inputs and DeepResonance-β for other scenarios. Finally, DeepResonance's zero-shot performance approaches the supervised upper bounds of MusiLingo on MusicInstruct in terms of BERTScore, demonstrating the out-of-domain generalization capabilities of DeepResonance.

5.5 Ablation Study

Figs. 5 and 6 present the results of ablation studies evaluating the effectiveness of key components in the proposed methods, including Music4way-MI2T and Music4way-MV2T instruction tuning data (**MWIT**), multi-sampled ImageBind embeddings (**MIE**), and the pre-alignment Transformer (**PT**). For music + text $\rightarrow$ text tasks (Fig. 6), we observe that MWIT, MIE, and PT each contribute positively to performance. However, when combined, PT does not consistently complement the other two components across all three benchmarks, with MWIT + MIE (DeepResonance-α) yielding the consistent improvements. For multimodal music understanding tasks (music + image/video + text $\rightarrow$ text), we compare settings with and without MIE and PT in Fig. 5, as MWIT serves as the in-domain data for these tasks. Integrating MIE and PT enhances performance, with PT proving most effective when limited to a single Transformer layer. This highlights the effectiveness of MIE and PT while suggesting that increasing PT's parameters may lead to overfitting on limited instruction tuning data. Moreover, PT with different layer settings negatively impacts performance on Music4way-Any2T (refer to Sec. 5.3).

6 Conclusion

In this study, we introduced DeepResonance, a multimodal music understanding LLM capable of comprehending music through its connections with other modalities, such as image and video. To train and evaluate DeepResonance, we developed new music-centric multi-way instruction-following datasets. In addition, we proposed modules designed to enhance music-centric multimodal instruction tuning. Empirical results highlight the effectiveness of DeepResonance across six music understanding tasks and zero-shot scenarios. Future work will explore more refined instruction tuning datasets to improve the model's generalization capabilities for music understanding tasks.

Limitations

The proposed methods have the following limitations:

(1) The input music training data is mostly limited to clips shorter than 30 seconds, with a significant portion (e.g., AudioSet) being 10 seconds. This may restrict the fine-tuned models' performance on longer music sequences. Additionally, the dataset's image frames are directly extracted from videos, assuming relevance within short clips (10s). For longer videos, selecting the most representative frames, such as cover images, should be explored in future work.

(2) The effectiveness of DeepResonance has been verified using Vicuna (based on LLaMA) as the backbone LLM, but further studies are needed to assess its improvements when applied to other LLM architectures.

(3) The extracted music features (e.g., downbeats, tempo, key, chords) rely on existing MIR algorithms, which may have precision limitations. Incorporating more signal-oriented features such as MFCCs or Chromagrams could be explored in future work.

(4) While the proposed methods perform well on supervised datasets included in the training data, further the enhancing generalization capability to out-of-domain (distribution-shifted) music remains an open challenge.

(5) Generating instruction tuning data with LLMs is a well-established and widely accepted approach, as seen in Self-Instruct (Wang et al., 2023). Our instruction tuning data construction process is relatively simple, ensuring the overall reliability. However, as with any LLM-generated data, biases may exist, and users should be mindful of potential biases.

Ethical Considerations

In this study, we leveraged publicly available datasets (without licensing issues) to create new datasets for multimodal music understanding. The newly generated content consists solely of text produced by LLMs such as GPT-4o mini, with no originally generated music, images, or videos. We fine-tuned the multimodal music understanding model through instruction tuning. While the model has been adapted to a specific domain, it may still generate hallucinations or biased content due to the nature of LLM-based text generation. Users should exercise caution when using the generated content, be aware of the potential risks associated with LLM outputs, and implement content safety checks as a post-processing measure.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. 2023. Musiclm: Generating music from text. *CoRR*, abs/2301.11325.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324.

Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. 2016. madmom: A new python audio and music signal processing library. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1174–1178. ACM.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2024. MusiLingo: Bridging music and text with pre-trained language models for music captioning and query response. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3643–3655, Mexico City, Mexico. Association for Computational Linguistics.

Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, pages 409–416.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Ángel Faraldo, Emilia Gómez, Sergi Jordà, and Perfecto Herrera. 2016. Key estimation in electronic dance music. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*, pages 335–347. Springer.

Joshua Patrick Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. 2024. Llark: A multimodal instruction-following language model for music. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Siddharth Gururani, Mohit Sharma, and Alexander Lerch. 2019. An attention mechanism for musical instrument recognition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 83–90.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M^2ugen: Multi-modal music understanding and generation with the power of large language models. *CoRR*, abs/2311.11255.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 286–290. IEEE.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2021. Muscaps: Generating captions for music audio. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wav-

caps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:3339–3354.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Johan Pauwels, Ken O'Hanlon, Emilia Gómez, and Mark B. Sandler. 2019. 20 years of automatic chord recognition from audio. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 54–63.

Hendrik Schreiber, Julián Urbano, and Meinard Müller. 2020. Music tempo estimation: Are we done yet? *Trans. Int. Soc. Music. Inf. Retr.*, 3(1):111.

Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2022. switch-glat: Multilingual parallel machine translation via code-switch decoder. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024a. SALMONN: towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024b. Codi-2: In-context, interleaved, and interactive any-to-any generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27415–27424. IEEE.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

John Thickstun, Zaïd Harchaoui, and Sham M. Kakade. 2017. Learning features of music from scratch. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

George Tzanetakis. 2001. Automatic musical genre classification of audio signals. In *ISMIR 2001, 2nd International Symposium on Music Information Retrieval, Indiana University, Bloomington, Indiana, USA, October 15-17, 2001, Proceedings*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Minz Won, Sergio Oramas, Oriol Nieto, Fabien Gouyon, and Xavier Serra. 2021. Multimodal metric learning for tag-based music retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 591–595. IEEE.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Furkan Yesiler, Guillaume Doras, Rachel M. Bittner, Christopher J. Tralie, and Joan Serrà. 2021. Audio-based musical version identification: Elements and challenges. *IEEE Signal Process. Mag.*, 38(6):115–136.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mengjie Zhao, Zhi Zhong, Zhuoyuan Mao, Shiqi Yang, Wei-Hsiang Liao, Shusuke Takahashi, Hiromi Wakaki, and Yuki Mitsufuji. 2024. Openmu: Your swiss army knife for music understanding. *CoRR*, abs/2410.15573.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *CoRR*, abs/2307.08581.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *CoRR*, abs/2408.11039.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Discussion on Multi-way Alignment

The 4-way alignment introduced in this work, which connects music, text, image, and video, is facilitated by pairing music with video and video with image. Each modality is further linked to text through captioning or feature extraction. Therefore, the 4-way relationship is constructed from several 2-way mappings, with any pair among four modalities being closely correlated as they stem from a single original music-video pair. Future work may develop finer-grained multi-way alignment for music understanding tasks. We encourage further discussion and research on how to establish improved multi-way relationships across different modalities.

B Training and Evaluation Details

Following the training strategy of NExT-GPT and M^{2}UGen, we train DeepResonance in two stages. In the first stage, we fine-tune only the parameters of the linear adaptors and the proposed pre-alignment Transformer. This stage focuses on captioning tasks for music, image, and video modalities, utilizing images from COCO (Lin et al., 2014) and music and video clips from the constructed Music4way dataset.[6] In the second stage, we fine-tune the linear adaptors and the pre-alignment Transformer while simultaneously performing LoRA-based fine-tuning (Hu et al., 2022) on Vicuna. This stage incorporates instruction tuning tasks using datasets including Alpaca (Taori et al., 2023), MusicCaps (Agostinelli et al., 2023), and MusicQA (Liu et al., 2024)[7], along with our constructed Music4way, Music4way-MI2T, and Music4way-MV2T datasets. A summary of all datasets used for training and evaluation is provided in Table 1. As depicted in Fig. 2, instructions are fed directly into the text-LLM, bypassing the adaptors and the pre-alignment Transformer, as they do not require interaction with the input information.

We fine-tune for 5 and 2 epochs in the first and second stages, respectively, utilizing a learning rate of $1\mathrm{e}{-4}$ and a batch size of 16. Training is conducted on 8 NVIDIA A100 GPUs (40GB each). For LoRA, the rank and alpha are both set to 32, following NExT-GPT. We train the pre-alignment Transformer with various layer configurations (see Sec. 5.5) and find that a single Transformer layer achieves the best performance. Regarding the trainable parameters, the linear adaptors, LoRA, pre-alignment Transformer, and LLaMA embedding layers contain 4M, 33M, 157M, and 262M parameters, respectively, comprising 5.6% of the total model parameters. Regarding the training budget, stage 1 took 25.2 hours, while stage 2 took 20.3 hours for DeepResonance-α. For DeepResonance-β, the training times were 27.0 hours for stage 1 and 22.7 hours for stage 2.

For evaluation, we report the mean results from three inference runs and include BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020), following the setup of M^{2}UGen (Hussain et al., 2023). We report BLEU-1, BLEU, ROUGE-L precision, ROUGE-L recall, ROUGE-L F1, BERTScore precision, BERTScore recall, and BERTScore F1 details in Appx. E, F, G, and H.

C Construction Details and Data Examples of Music4way-MI2T and Music4way-MV2T

Table 5 presents the specific templates used to prompt the MPT-7B model (Team, 2023) to generate unified captions based on music, video, and image captions, along with low-level music features. Table 6 provides data examples from the Music4way-MI2T and Music4way-MV2T datasets we constructed for multi-way instruction tuning.

[6] We use COCO instead of Music4way for the image captioning task in the first stage, as empirical results indicate that using the larger image-text dataset, COCO, yields better performance.

[7] We use the "fine-tune" split of the MusicQA dataset and exclude the "train" split to avoid overlap with the test split of MusicCaps.

Dataset	#Instance_train	#Instance_test	Used for	In→out modality	Task
COCO	82,783	–	Train stage 1	I+T→T	Image captioning
Music4way	59,128	–	Train stage 2	I+T→T	Image captioning
Music4way	59,128	–	Train stages 1 & 2	V+T→T	Video captioning
Music4way (Music4way-MusicCaps)	59,128	3,000	Train stages 1 & 2, test	M+T→T	Music captioning
Alpaca	51,974	–	Train stage 2	T→T	Text question-answering
MusicQA	70,011	5,040	Train stage 2, test	M+T→T	Music captioning and question-answering
MusicCaps	2,640	2,839	Train stage 2, test	M+T→T	Music captioning
Music4way-MI2T	59,128	3,000	Train stage 2, test	M+I+T→T	Multimodal captioning
Music4way-MV2T	59,128	3,000	Train stage 2, test	M+V+T→T	Multimodal captioning
Music4way-Any2T	59,128	3,000	Test	M+I/V+T→T	Multimodal question-answering
GTZAN	–	1,406	Test	M+T→T	Music captioning and question-answering
MusicNet	–	140	Test	M+T→T	Music captioning
MusicInstruct-long	–	16,658	Test	M+T→T	Music captioning
MusicInstruct-short	–	13,935	Test	M+T→T	Music captioning
MTG	–	25,452	Test	M+T→T	Music captioning and question-answering

Table 4: Overview of datasets used for training and evaluation. M: music; I: image; V: video; T: text.

D Construction Details and Data Examples of Music4way-Any2T

Table 7 displays the templates used to prompt GPT-4o mini[8] to generate structured text comprising input, instruction, and output, based on music, video, and image captions, unified captions, and music features. Table 8 presents data examples from the Music4way-Any2T dataset we constructed to evaluate the robustness and generalization capabilities of music LLMs.

E Detailed Results on Music Understanding Tasks

Tables 9, 10, and 11 report the detailed performance on all metrics of all models introduced in Sec.5.1 on music understanding tasks (music + text → text) using MusicQA(Liu et al., 2024), MusicCaps (Agostinelli et al., 2023), and our constructed Music4way-MusicCaps.

F Detailed Results on Multimodal Music Understanding Tasks

Tables 12, 13, and 14 list the result details on all metrics on our newly proposed multimodal music understanding tasks (music + image/video + text → text), including Music4way-MI2T, Music4way-MV2T, and Music4way-Any2T.

G Detailed Results of Zero-shot Evaluation

The detailed results of the zero-shot evaluation are presented in Tables 15, 16, 17, 18, and 19.

[8] https://platform.openai.com/docs/models#gpt-4o-mini

H Detailed Results of Ablation Study

The detailed results of the ablation study are presented in Tables 20, 21, 22, 23, 24 and 25.

Input	- Video Caption: A man is playing a song on a guitar and he is sitting on a table. - Image Caption: The image shows a man playing a guitar. - Music Caption: The music is a slow, heavy metal song with a dark and intense mood. The instruments used are electric guitar, bass guitar, drums, and vocals ... - Music Features: — Tempo: [[105.26315789473684, 0.4619219529856875], ...] — Chords: [[0.0, 1.5, 'C:maj'], [1.5, 3.7, 'F:maj'], ...] — Downbeats: [[0.25, 3.0], [0.81, 4.0], [1.39, 1.0], [1.99, 2.0], ...] — Key: [[9.83237405307591e-05, 8.27968106023036e-05, ...]
Instruction	Given the above information of video captions, image captions, music captions, and music features, generate a unified description that combines the elements of both the video and the music, taking into account the mood, style, and emotions conveyed by the captions and music features. The description should be cohesive and provide a holistic view of the content, reflecting how the visual and auditory components complement each other. Focus on creating a narrative that integrates the rhythm, harmony, and tonality of the music with the visual elements and storyline of the video.
Output (Unified Captions)	The video and music describe a fast-paced and energetic performance by a group of musicians, with a positive and upbeat mood. The music features a driving rhythm, a strong beat, and a variety of instruments, including drums, bass, electric guitar, and keyboards. The chords used in the music are major, with a key of A minor. The tempo ranges from 60 to 120 beats per minute, with a strong sense of energy and movement conveyed by the upbeat tone of the music. The downbeats occur every four beats, with a strong emphasis on the first and third beats of each measure. Overall, the visual and auditory components complement each other to create a cohesive and engaging experience for the viewer.

Table 5: Prompting MPT-7B to generate unified captions for Music4way-MI2T and Music4way-MV2T.

Input (Music4way-MI2T) Input (Music4way-MV2T)	<Music> <Image> <Music> <Video>
Instruction	Analyze the music by considering both its auditory and visual components. Describe the music in detail, incorporating its tempo, chords, downbeats, and key, while also reflecting on how these musical features align with the video or a key image from the video.
Output (Unified Captions)	The video and music describe a fast-paced and energetic performance by a group of musicians, with a positive and upbeat mood. The music features a driving rhythm, a strong beat, and a variety of instruments, including drums, bass, electric guitar, and keyboards. The chords used in the music are major, with a key of A minor. The tempo ranges from 60 to 120 beats per minute, with a strong sense of energy and movement conveyed by the upbeat tone of the music. The downbeats occur every four beats, with a strong emphasis on the first and third beats of each measure. Overall, the visual and auditory components complement each other to create a cohesive and engaging experience for the viewer.

Table 6: Data examples of Music4way-MI2T and Music4way-MV2T for instruction tuning. "<Music>", "<Image>", and "<Video>" represent the paths to the respective files.

Input	Generate the output for the following input; the style of the input, instruction, and output may vary. - Video Caption: A man is playing a song on a guitar and he is sitting on a table. - Image Caption: The image shows a man playing a guitar. - Music Caption: The music is a slow, heavy metal song with a dark and intense mood. The instruments used are electric guitar, bass guitar, drums, and vocals ... - Music Features: — Tempo: [[105.26315789473684, 0.4619219529856875], ...] — Chords: [[0.0, 1.5, 'C:maj'], [1.5, 3.7, 'F:maj'], ...] — Downbeats: [[0.25, 3.0], [0.81, 4.0], [1.39, 1.0], [1.99, 2.0], ...] — Key: [[9.83237405307591e-05, 8.27968106023036e-05, ...] - Unified Caption: The video and music describe a fast-paced and energetic performance by a group of musicians, with a positive and upbeat mood. The music features a driving rhythm, a strong beat, and a variety of instruments, including drums, bass, ...
Instruction	Help generate input, instruction, and output triplets using the given paired music caption, video caption, image caption, music features, and unified caption. The unified caption includes all three captions as well as music features such as tempo, chords, downbeats, and key. Guidelines: 1. Input: Should be a sentence that includes two or all three modalities: music (mandatory), video, and/or image. Music must be referred to as <Music>. Image must be referred to as <Image>. Video must be referred to as <Video>. 2. Instruction: Should be a text-based question or directive that requires generating a unified output based on the given inputs. It should guide the model to consider both the audio and the visual aspects, explaining how they interact to create a unified experience. 3. Output: Should be a textual response, potentially composed using the information from the music, video, image captions or the unified caption. Example: Input: Consider the music of <Music> and its paired image of <Image> that visually represents a key moment of the music. Instruction: Provide a description of how the music's rhythm, tempo, and tonal qualities are visually represented in the image, combining insights from both the music and the image captions. Output: The music is characterized by a slow tempo with a calm, serene melody, primarily featuring piano and soft strings. The image complements this mood, showing a peaceful sunset over a calm ocean. The soft, gentle waves in the image mirror the steady downbeats of the music, while the warm color palette in the image reflects the key of the music, which is in A major, creating a sense of tranquility. Together, they evoke a feeling of peace and reflection.
Output	Input: Consider the music of <Music>; and its paired image of <Image>; or video of <Video>; that visually represents a dark and intense atmosphere. Instruction: Describe how the music's slow, heavy metal style and the visuals of a man playing a guitar are reflected in the image or video. Output: The music is a slow, heavy metal song with a dark and intense mood. It features electric guitar, bass guitar, drums, and powerful vocals. The tempo is slow and the rhythm is steady, with a heavy, distorted guitar sound and a solid foundation provided by the bass guitar and drums. The vocals are deep and powerful, emphasizing the lower register. This creates a heavy and intense sound. The image or video shows a man playing a guitar, holding it, and sitting on a table. The visuals are simple but effective, focusing on the man and his instrument. The dark and intense atmosphere of the music is reflected in the visuals, capturing the mood and energy of the song. Together, the music and visuals create a cohesive experience that immerses the audience in a dark and intense musical journey.

Table 7: Prompting GPT-4o mini to generate triplets of input, instruction, and output for Music4way-Any2T. The instruction represents the system description used when prompting GPT-4o mini.

Example 1	
Input	The music from <Music> plays alongside a captivating scene of <Video> where a person is sitting and showcasing their musical talent with a string instrument, which is also reflected in <Image>.
Instruction	Analyze how the mood and emotion conveyed through the music interact with the visual elements presented in the video and image, focusing on the cohesive atmosphere they create together.
Output	The music is characterized by a slow tempo and intense instrumental arrangement, prominently featuring the guitar which serves as the main tune. The video depicts a person deeply focused while playing a string instrument, highlighting their connection with the music. This is mirrored in the image of the man playing the guitar, showcasing the intimacy of this musical expression. The harmonious blend of the electric guitar enhances the emotional depth, adding layers to the music that evoke feelings of intensity. The consistent downbeats and the key of D minor resonate with the serious tone of the visuals. Together, the visual components and the auditory experience create a mesmerising atmosphere, pulling the viewer into a world where music and emotion intertwine seamlessly.
Example 2	
Input	Consider the music of <Music>, which is a slow and mellow instrumental with a melancholic mood, alongside the paired video of <Video> where a man is playing a song on a guitar.
Instruction	Explain how the guitar performance in the video enhances the melancholic atmosphere of the music, and describe how both elements work together to create an emotional experience for the viewer.
Output	The music, characterized by its slow tempo and melancholic mood, beautifully complements the visuals of the man playing the guitar in the video. As he strums the instrument on the table, the gentle, steady rhythm of the music aligns with his movements, emphasizing the emotional weight of each chord he plays. The use of major and minor chords, particularly in the F# major scale, enhances the bittersweet feelings conveyed by the music. The visuals of the man fully engrossed in his performance highlight the introspective nature of the piece, creating a profound connection between the audio and the visual. Together, they evoke a powerful sense of sadness and reflection, inviting the viewer to immerse themselves in this poignant moment.
Example 3	
Input	Consider the music of <Music> that is fast-paced and energetic, while viewing the video of <Video> showing a young man playing the guitar, alongside the image of <Image> depicting the same scene.
Instruction	Describe how the fast-paced rhythm and energetic tone of the music interact with the visuals of the video and image, focusing on the emotions and energy conveyed by both the audio and visual components.
Output	The fast-paced music, characterized by a strong beat and an upbeat tempo, perfectly complements the visuals of the young man playing guitar in both the video and the image. The upbeat rhythm creates a vibrant atmosphere, as the energetic strumming of the guitar aligns with the driving drums and pulsating bass. The 4/4 time signature of the music further emphasizes the lively performance, with each downbeat coinciding with the man's movements. The use of E major and A major chords adds a bright and happy tonal quality that is visually represented in both the video and image through the man's engaging expression and dynamic posture, conveying a sense of excitement and intensity that radiates through the screen. Together, the music and visuals fuse seamlessly, creating an exhilarating experience that captures the essence of a lively performance.

Table 8: More examples of Music4way-Any2T for evaluation. "<Music>", "<Image>", and "<Video>" represent the paths to the respective files.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
SALMONN (Tang et al., 2024a)†	28.7	13.1	45.7	37.2	35.4	90.8	89.9	90.3
MU-LLaMA (Liu et al., 2024)†	29.7	10.9	30.8	**46.4**	33.1	89.8	90.0	89.9
OpenMU (Zhao et al., 2024)†	24.5	6.1	20.8	**44.8**	25.5	86.9	90.5	88.6
NExT-GPT (Wu et al., 2024)	23.3	7.6	26.2	38.4	26.0	86.7	88.5	87.6
M^{2}UGen (Hussain et al., 2023)†	29.1	13.1	**52.7**	35.0	37.9	91.7	89.3	90.5
NExT-GPT w/ M^{2}UGen†	34.0	14.3	49.1	38.3	39.6	92.1	90.5	91.2
NExT-GPT w/ Music4way†	34.1	13.8	48.2	37.9	39.2	92.1	90.5	91.3
DeepResonance-α (ours)†	**35.1**	**15.1**	51.0	38.6	**40.8**	**92.5**	**90.7**	**91.6**
DeepResonance-β (ours)†	**35.6**	**15.3**	**51.3**	39.0	**41.1**	**92.5**	**90.8**	**91.6**

Table 9: **Results on MusicQA.** The top two performances are highlighted in **bold**. "†" indicates that the test set's train split is included in the corresponding model's training data.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
SALMONN (Tang et al., 2024a)†	19.7	0.9	23.6	21.5	19.1	**87.6**	86.3	86.9
MU-LLaMA (Liu et al., 2024)†	*9.6	*0.2	*32.0	*12.0	*16.2	*88.7	*85.1	*86.8
OpenMU (Zhao et al., 2024)†	23.9	1.2	18.8	**22.8**	19.4	86.1	**87.2**	86.6
MusiLingo sft. w/ MusicCaps (Deng et al., 2024)†	–	–	–	–	**21.7**	–	–	86.8
NExT-GPT (Wu et al., 2024)	16.5	0.1	15.3	16.4	14.0	84.3	83.8	84.0
M^{2}UGen (Hussain et al., 2023)†	*14.4	*0.4	*26.1	*14.5	*16.4	*87.8	*85.4	*86.5
NExT-GPT w/ M^{2}UGen	12.4	0.1	**28.2**	13.2	16.1	**88.6**	85.3	86.9
NExT-GPT w/ Music4way†	25.0	2.7	23.0	20.6	21.0	**87.6**	87.0	87.2
DeepResonance-α (ours)†	**26.0**	**3.0**	23.4	**21.8**	**21.6**	**87.6**	**87.1**	**87.3**
DeepResonance-β (ours)†	**25.8**	**2.8**	**23.6**	21.4	**21.6**	87.5	**87.1**	**87.3**

Table 10: **Results on MusicCaps.** The top two performances are highlighted in **bold**. "*" denotes results that should be interpreted with caution, as the test data was included in the corresponding model's training set. "†" indicates that the test set's train split is included in the corresponding model's training data.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
SALMONN (Tang et al., 2024a)	19.1	0.8	15.7	39.8	20.0	85.5	88.6	87.0
MU-LLaMA (Liu et al., 2024)	15.1	0.4	52.1	20.9	27.6	90.1	86.6	88.3
NExT-GPT (Wu et al., 2024)	16.6	0.2	12.4	37.0	17.2	85.3	88.2	86.7
M^{2}UGen (Hussain et al., 2023)†	*13.1	*0.0	*52.8	*19.5	*26.0	*89.6	*85.8	*87.6
NExT-GPT w/ M^{2}UGen†	*23.2	*6.7	*60.8	*29.7	*36.7	*93.1	*89.6	*91.3
NExT-GPT w/ Music4way†	39.1	18.4	55.5	46.9	46.8	91.7	**92.5**	93.0
DeepResonance-α (ours)†	**40.9**	**19.9**	**57.8**	**47.5**	**48.4**	**94.0**	**92.6**	**93.3**
DeepResonance-β (ours)†	**39.9**	**19.3**	**57.3**	**47.3**	**47.8**	**93.9**	**92.5**	**93.2**

Table 11: **Results on Music4way-MusicCaps.** The top two performances are highlighted in **bold**. "*" denotes results that should be interpreted with caution, as the test data was included in the corresponding model's training set. "†" indicates that the test set's train split is included in the corresponding model's training data.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	26.7	1.2	26.2	19.7	21.3	85.5	85.0	85.2
M^{2}UGen (Hussain et al., 2023)	*31.7	*4.1	*29.5	*26.5	*26.4	*87.7	*86.6	*87.1
NExT-GPT w/ M^{2}UGen	*33.8	*4.1	*35.0	*23.9	*27.3	*89.0	*87.3	*88.1
NExT-GPT w/ Music4way	24.2	1.7	25.1	23.6	22.0	85.7	85.3	85.4
DeepResonance-α (ours)	**44.6**	**9.2**	**27.8**	**33.8**	**30.1**	**89.1**	**89.6**	**89.4**
DeepResonance-β (ours)	**49.2**	**17.2**	**36.8**	**38.3**	**36.8**	**90.1**	**90.3**	**90.2**

Table 12: **Results on Music4way-MI2T.** The top two performances are highlighted in **bold**. "*" denotes results that should be interpreted with caution, as the test data was included in the corresponding model's training set.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	26.5	1.1	25.6	19.6	21.0	85.1	84.6	84.8
M^{2}UGen (Hussain et al., 2023)	*31.7	*4.3	*27.9	*26.5	*25.9	*87.2	*86.4	*86.8
NExT-GPT w/ M^{2}UGen	*34.6	*4.0	*34.8	*24.2	*27.3	*88.9	*87.4	*88.1
NExT-GPT w/ Music4way	25.0	1.8	25.4	24.2	22.5	85.7	85.4	85.5
DeepResonance-α (ours)	**44.5**	**8.9**	**28.2**	**32.7**	**29.6**	**89.0**	**89.4**	**89.2**
DeepResonance-β (ours)	**49.0**	**17.2**	**36.7**	**38.4**	**36.8**	**90.3**	**90.3**	**90.3**

Table 13: **Results on Music4way-MV2T.** The top two performances are highlighted in **bold**. "*" denotes results that should be interpreted with caution, as the test data was included in the corresponding model's training set.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	25.4	3.4	33.5	19.8	23.4	87.4	85.9	86.6
M^{2}UGen (Hussain et al., 2023)	*20.8	*2.5	*34.4	*17.1	*21.5	*88.9	*85.8	*87.3
NExT-GPT w/ M^{2}UGen	*26.3	*4.8	*41.7	*22.2	*28.5	*91.1	*87.7	*89.4
NExT-GPT w/ Music4way	29.4	2.8	**35.9**	22.7	27.0	89.7	87.1	88.4
DeepResonance-α (ours)	**54.7**	**14.2**	**35.9**	**34.6**	**35.1**	**91.2**	**90.6**	**90.9**
DeepResonance-β (ours)	**33.5**	**3.9**	34.0	**23.9**	**27.4**	**90.0**	**87.5**	**88.7**

Table 14: **Results on Music4way-Any2T.** The top two performances are highlighted in **bold**. "*" denotes results that should be interpreted with caution, as the test data was included in the corresponding model's training set.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	**19.3**	1.0	22.2	16.7	16.9	85.0	84.2	84.6
NExT-GPT w/ M^{2}UGen	15.4	2.9	**38.6**	16.6	21.7	88.9	86.3	87.5
NExT-GPT w/ Music4way	16.3	2.6	37.2	17.2	22.5	**89.6**	86.9	**88.2**
DeepResonance-α (ours)	19.2	**3.1**	35.8	**18.7**	**23.0**	89.4	**87.0**	**88.2**
DeepResonance-β (ours)	**19.5**	**4.5**	**39.2**	**20.1**	**25.2**	**90.1**	**87.5**	**88.8**
OpenMU (supervised performance upper bound)	*47.2*	*18.5*	*35.7*	*41.2*	*37.5*	*91.1*	*91.9*	*91.5*

Table 15: **Results on GTZAN.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	**19.5**	0.0	23.6	10.1	13.5	83.7	83.2	83.4
NExT-GPT w/ M^{2}UGen	4.3	**0.2**	**42.0**	9.4	14.1	**88.0**	84.8	85.8
NExT-GPT w/ Music4way	13.7	**0.1**	24.8	12.2	15.8	86.9	**85.2**	86.0
DeepResonance-α (ours)	14.6	**0.1**	24.4	**12.5**	**15.9**	**87.0**	**85.2**	**86.1**
DeepResonance-β (ours)	**14.9**	**0.1**	**25.1**	**12.7**	**16.4**	**87.0**	**85.6**	**86.2**
OpenMU (supervised performance upper bound)	*52.4*	*22.0*	*35.8*	*44.4*	*39.3*	*91.5*	*92.4*	*92.0*

Table 16: **Results on MusicNet.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	**13.9**	0.1	23.7	**12.2**	15.4	85.7	**84.2**	84.9
NExT-GPT w/ M^{2}UGen	7.1	0.1	33.0	10.2	15.3	87.6	83.6	85.6
NExT-GPT w/ Music4way	7.3	**0.2**	32.7	10.4	15.5	87.8	83.8	85.8
DeepResonance-α (ours)	8.6	**0.2**	**35.9**	10.5	**15.8**	**88.4**	83.8	**86.0**
DeepResonance-β (ours)	**9.4**	**0.2**	**33.6**	**11.4**	**16.5**	**88.4**	**84.0**	**86.1**
MusiLingo (supervised performance upper bound)	*45.0*	–	–	–	*22.9*	–	–	*86.1*

Table 17: **Results on MusicInstruct-long.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	28.9	10.3	28.6	**50.1**	32.1	88.5	90.9	89.6
NExT-GPT w/ M^{2}UGen	41.5	15.5	47.3	46.8	44.6	92.5	91.8	92.1
NExT-GPT w/ Music4way	42.9	16.4	47.8	46.5	44.9	92.7	91.9	92.3
DeepResonance-α (ours)	**43.0**	**16.8**	**48.0**	47.3	**45.3**	**92.8**	**92.1**	**92.4**
DeepResonance-β (ours)	**44.7**	**18.5**	**50.4**	**47.9**	**47.0**	**93.1**	**92.2**	**92.6**
MusiLingo (supervised performance upper bound)	*47.0*	–	–	–	*51.4*	–	–	*92.9*

Table 18: **Results on MusicInstruct-short.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
NExT-GPT (Wu et al., 2024)	**25.6**	**5.1**	22.6	**31.2**	23.7	87.6	**88.4**	88.0
NExT-GPT w/ M^{2}UGen	18.9	3.3	**38.4**	19.7	24.2	89.6	87.2	88.4
NExT-GPT w/ Music4way	20.6	3.0	36.7	20.4	25.0	89.7	87.7	88.7
DeepResonance-α (ours)	22.6	3.7	35.4	21.6	**25.3**	**89.9**	87.9	**88.8**
DeepResonance-β (ours)	**23.5**	**4.7**	**37.8**	**23.0**	**27.2**	**90.4**	**88.2**	**89.3**
OpenMU (supervised performance upper bound)	*45.7*	*19.2*	*36.7*	*46.3*	*39.8*	*91.2*	*92.4*	*91.8*

Table 19: **Results on MTG.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
(1): **NExT-GPT w/ Music4way**	34.1	13.8	48.2	37.9	39.2	92.1	90.5	91.3
(1) + MIE	35.0	14.5	50.1	38.3	40.4	92.3	90.6	91.4
(1) + PT (6-layer)	**35.8**	**15.3**	49.0	**39.2**	40.7	92.1	90.6	91.3
(1) + MIE + PT (6-layer)	35.5	14.8	46.4	39.0	39.4	91.4	90.5	90.9
(1) + PT (2-layer)	34.5	14.8	50.9	38.3	40.5	92.4	90.6	91.5
(1) + MIE + PT (2-layer)	34.3	14.3	50.2	38.0	40.1	92.4	90.6	91.4
(1) + PT (1-layer)	34.9	14.6	49.6	38.3	40.2	92.3	**90.8**	91.5
(1) + MIE + PT (1-layer)	**35.6**	15.3	50.6	**39.5**	**41.1**	92.4	**90.8**	91.5
(2): (1) + Music4way-MI2T & Music4way-MV2T	35.0	14.5	49.5	38.5	40.2	92.3	90.7	91.5
(2) + MIE (**DeepResonance-**α)	35.1	15.1	**51.0**	38.6	40.8	**92.5**	90.7	**91.6**
(2) + PT (6-layer)	34.4	14.8	49.0	38.0	40.0	92.2	90.4	91.3
(2) + MIE + PT (6-layer)	35.3	14.6	46.6	38.7	39.4	91.6	90.4	91.0
(2) + PT (2-layer)	34.2	14.7	**51.0**	38.0	40.4	**92.5**	90.6	91.5
(2) + MIE + PT (2-layer)	34.5	14.6	50.8	38.1	40.4	**92.5**	90.6	91.5
(2) + PT (1-layer)	35.5	15.0	50.3	**39.2**	40.9	92.4	**90.8**	**91.6**
(2) + MIE + PT (1-layer) (**DeepResonance-**β)	**35.6**	**15.3**	**51.3**	39.0	**41.1**	**92.5**	**90.8**	**91.6**

Table 20: **Ablation study on MusicQA.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
(1): **NExT-GPT w/ Music4way**	25.0	2.7	23.0	20.6	21.0	**87.6**	87.0	87.2
(1) + MIE	25.4	2.6	22.8	21.3	21.1	87.5	87.0	87.2
(1) + PT (6-layer)	24.3	2.5	21.3	20.5	20.0	87.1	86.6	86.8
(1) + MIE + PT (6-layer)	21.9	1.7	21.1	18.5	18.6	86.9	86.1	86.5
(1) + PT (2-layer)	24.7	2.4	21.8	20.6	20.2	87.3	86.9	87.1
(1) + MIE + PT (2-layer)	25.5	2.5	22.2	21.2	20.8	87.4	87.0	87.2
(1) + PT (1-layer)	25.7	**2.8**	22.7	21.4	21.1	87.5	**87.1**	**87.3**
(1) + MIE + PT (1-layer)	25.8	**2.8**	**23.5**	21.5	**21.6**	87.5	87.0	**87.3**
(2): (1) + Music4way-MI2T & Music4way-MV2T	25.9	**2.8**	23.1	**21.6**	21.4	**87.6**	87.0	**87.3**
(2) + MIE (**DeepResonance-**α)	**26.0**	**3.0**	23.4	**21.8**	**21.6**	**87.6**	**87.1**	**87.3**
(2) + PT (6-layer)	21.6	1.7	21.3	18.0	18.5	86.9	86.1	86.5
(2) + MIE + PT (6-layer)	21.6	1.8	21.2	18.2	18.5	87.0	86.1	86.5
(2) + PT (2-layer)	24.9	2.5	22.0	20.8	20.4	87.3	86.9	87.1
(2) + MIE + PT (2-layer)	25.1	2.4	22.1	21.0	20.5	87.3	86.9	87.1
(2) + PT (1-layer)	**25.9**	**2.8**	**23.5**	21.4	**21.6**	87.5	87.0	**87.3**
(2) + MIE + PT (1-layer) (**DeepResonance-**β)	25.8	**2.8**	**23.6**	21.4	**21.6**	87.5	**87.1**	**87.3**

Table 21: **Ablation study on MusicCaps.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
(1): **NExT-GPT w/ Music4way**	39.1	18.4	55.5	46.9	46.8	91.7	92.5	93.0
(1) + MIE	39.5	18.2	55.8	46.5	46.8	93.7	92.5	93.0
(1) + PT (6-layer)	35.9	14.5	50.2	43.2	42.7	92.8	91.9	92.3
(1) + MIE + PT (6-layer)	33.4	14.1	**58.0**	40.5	45.0	93.6	91.0	92.3
(1) + PT (2-layer)	39.6	18.7	56.7	46.3	47.3	93.9	92.4	93.1
(1) + MIE + PT (2-layer)	39.9	19.1	57.2	46.9	47.7	93.9	92.5	**93.2**
(1) + PT (1-layer)	40.3	19.3	57.6	46.8	47.9	**94.0**	92.5	**93.2**
(1) + MIE + PT (1-layer)	39.9	19.1	56.7	47.1	47.5	93.8	92.5	93.1
(2): (1) + Music4way-MI2T & Music4way-MV2T	**40.7**	**19.8**	57.0	**48.0**	**48.2**	93.9	**92.6**	**93.2**
(2) + MIE (**DeepResonance-**α)	**40.9**	**19.9**	57.8	**47.5**	**48.4**	**94.0**	**92.6**	**93.3**
(2) + PT (6-layer)	33.5	14.2	**58.0**	40.4	45.0	93.6	91.0	92.3
(2) + MIE + PT (6-layer)	33.6	14.2	**58.0**	40.5	45.1	93.6	91.0	92.3
(2) + PT (2-layer)	39.6	18.7	56.7	46.3	47.3	93.9	92.4	93.1
(2) + MIE + PT (2-layer)	39.2	18.8	57.1	46.8	47.4	93.9	92.4	93.1
(2) + PT (1-layer)	40.3	19.5	57.3	47.4	48.0	93.9	92.5	**93.2**
(2) + MIE + PT (1-layer) (**DeepResonance-**β)	39.9	19.3	57.3	47.3	47.8	93.9	92.5	**93.2**

Table 22: **Ablation study on Music4way-MusicCaps.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
(1): **NExT-GPT w/ Music4way**	24.2	1.7	25.1	23.6	22.0	85.7	85.3	85.4
(1) + MIE	23.7	1.7	30.0	18.5	22.1	87.9	85.8	86.8
(1) + PT (6-layer)	20.6	1.3	30.8	17.8	21.0	87.6	85.5	86.5
(1) + MIE + PT (6-layer)	12.0	0.3	25.4	14.6	16.2	85.0	84.1	84.5
(1) + PT (2-layer)	23.0	1.8	32.8	19.9	23.3	87.8	85.9	86.8
(1) + MIE + PT (2-layer)	10.7	0.2	29.2	11.8	16.0	86.9	84.3	85.6
(1) + PT (1-layer)	19.4	1.4	30.5	17.1	20.7	87.6	85.5	86.6
(1) + MIE + PT (1-layer)	15.1	0.6	30.1	14.6	18.8	87.4	84.9	86.1
(2): (1) + Music4way-MI2T & Music4way-MV2T	43.8	8.7	28.0	34.0	30.1	88.8	89.5	89.1
(2) + MIE (**DeepResonance-**α)	44.6	9.2	27.8	33.8	30.1	89.1	89.6	89.4
(2) + PT (6-layer)	48.6	15.5	**37.1**	36.2	36.1	90.0	90.1	90.0
(2) + MIE + PT (6-layer)	**49.1**	15.7	36.3	36.9	36.1	**90.1**	90.1	90.1
(2) + PT (2-layer)	47.9	15.9	36.7	36.5	35.8	**90.1**	90.1	90.1
(2) + MIE + PT (2-layer)	49.0	**17.1**	36.7	**38.3**	**36.7**	90.0	**90.3**	90.1
(2) + PT (1-layer)	48.4	16.6	**37.3**	37.0	36.4	**90.3**	**90.2**	**90.2**
(2) + MIE + PT (1-layer) (**DeepResonance-**β)	**49.2**	**17.2**	36.8	**38.3**	**36.8**	**90.1**	**90.3**	**90.2**

Table 23: **Ablation study on Music4way-MI2T.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
(1): **NExT-GPT w/ Music4way**	25.0	1.8	25.4	24.2	22.5	85.7	85.4	85.5
(1) + MIE	14.2	0.6	34.7	14.9	20.1	88.1	84.9	86.4
(1) + PT (6-layer)	17.0	1.1	34.6	16.6	20.8	88.1	85.3	86.7
(1) + MIE + PT (6-layer)	16.3	0.7	28.3	16.1	19.0	86.2	84.9	85.5
(1) + PT (2-layer)	23.3	1.8	31.9	20.6	23.2	87.6	85.8	86.7
(1) + MIE + PT (2-layer)	6.4	0.2	36.2	10.3	15.2	87.5	83.8	85.6
(1) + PT (1-layer)	19.5	1.3	30.8	17.0	20.8	87.7	85.5	86.6
(1) + MIE + PT (1-layer)	10.3	0.6	**37.5**	13.5	18.9	88.3	84.7	86.5
(2): (1) + Music4way-MI2T & Music4way-MV2T	43.0	8.8	28.0	32.7	29.6	88.8	89.4	89.1
(2) + MIE (**DeepResonance-**α)	44.5	8.9	28.2	32.7	29.6	89.0	89.4	89.2
(2) + PT (6-layer)	48.5	15.4	**37.0**	36.2	36.1	89.9	90.1	90.0
(2) + MIE + PT (6-layer)	48.7	15.3	36.9	36.2	35.9	**90.4**	90.0	**90.2**
(2) + PT (2-layer)	**49.0**	16.9	36.6	37.9	36.5	90.0	**90.3**	90.1
(2) + MIE + PT (2-layer)	48.9	**17.0**	36.6	**38.4**	**36.7**	90.0	**90.3**	90.1
(2) + PT (1-layer)	48.5	16.5	**37.0**	36.9	36.2	90.2	90.2	**90.2**
(2) + MIE + PT (1-layer) (**DeepResonance-**β)	**49.0**	**17.2**	36.7	**38.4**	**36.8**	**90.3**	**90.3**	**90.3**

Table 24: **Ablation study on Music4way-MV2T.** The top two performances are highlighted in **bold**.

Model	BLEU-1	BLEU	ROUGE-P	ROUGE-R	ROUGE-F1	BERT-P	BERT-R	BERT-F1
(1): **NExT-GPT w/ Music4way**	29.4	2.8	35.9	22.7	27.0	89.7	87.1	88.4
(1) + MIE	16.2	1.2	39.1	16.7	22.8	89.8	85.9	87.8
(1) + PT (6-layer)	3.7	0.0	10.9	10.2	7.0	70.9	77.2	73.8
(1) + MIE + PT (6-layer)	21.3	1.6	34.2	19.6	23.5	88.3	85.9	87.0
(1) + PT (2-layer)	7.4	0.5	36.9	10.8	14.4	84.3	80.8	82.5
(1) + MIE + PT (2-layer)	1.3	0.0	**41.5**	7.5	12.1	87.3	82.4	84.8
(1) + PT (1-layer)	23.5	1.9	34.7	20.0	24.1	88.9	86.1	87.5
(1) + MIE + PT (1-layer)	1.6	0.0	**43.5**	8.2	13.4	87.9	82.6	85.2
(2): (1) + Music4way-MI2T & Music4way-MV2T	**54.4**	**13.7**	34.3	**34.8**	**34.4**	**90.9**	**90.5**	**90.7**
(2) + MIE (**DeepResonance-**α)	**54.7**	**14.2**	35.9	**34.6**	**35.1**	**91.2**	**90.6**	**90.9**
(2) + PT (6-layer)	36.9	4.6	33.2	25.2	28.2	90.1	87.8	88.9
(2) + MIE + PT (6-layer)	34.1	4.1	34.2	24.3	27.9	90.2	87.6	88.9
(2) + PT (2-layer)	17.8	1.5	36.2	17.1	20.3	88.0	84.9	86.3
(2) + MIE + PT (2-layer)	36.2	3.4	28.6	25.0	25.9	88.6	87.3	87.9
(2) + PT (1-layer)	24.7	2.4	34.0	20.3	23.5	88.6	86.0	87.2
(2) + MIE + PT (1-layer) (**DeepResonance-**β)	33.5	3.9	34.0	23.9	27.4	90.0	87.5	88.7

Table 25: **Ablation study on Music4way-Any2T.** The top two performances are highlighted in **bold**.