AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

Anonymous ACL submission

Abstract

We introduce AnyGPT, an any-to-any multimodal language model that utilizes discrete representations for the unified processing of various modalities, including speech, text, images, and music. AnyGPT can be trained stably without any alterations to the current large language model (LLM) architecture or training paradigms. Instead, it relies exclusively on data-level preprocessing, facilitating the seamless integration of new modalities into LLMs, akin to the incorporation of new languages. We build a multimodal text-centric dataset for multimodal alignment pre-training. Utilizing generative models, we synthesize the first large-scale any-to-any multimodal instruction dataset. It consists of 108k samples of multi-turn conversations that intricately interweave various modalities, thus equipping the model to handle arbitrary combinations of multimodal inputs and outputs. Experimental results demonstrate that AnyGPT is capable of facilitating any-to-any multimodal conversation while achieving performance comparable to specialized models across all modalities, proving that discrete representations can effectively and conveniently unify multiple modalities within a language model. Demos are shown in https://any-gpt.github.io/AnyGPT.github.io.

1 Introduction

LLMs have exhibited remarkable proficiency in comprehending and generating human language. Nevertheless, their capabilities are confined to textual processing. The real-world environment is inherently multimodal, with organisms perceiving and exchanging information through diverse channels, including vision, language, sound, and touch.

A promising objective in developing multimodal systems is to augment LLMs with the capacity for multimodal perception. The dominant methodology involves the integration of multimodal encoders with the language model, thus empowering it to process information across various modalities and utilize its sophisticated text-processing abilities to produce coherent responses. However, this strategy is limited to text generation and does not encompass multimodal output.

Pioneering efforts such as Emu (Sun et al., 2023b), SEED-LLaMA (Ge et al., 2023b) and SpeechGPT (Zhang et al., 2023a) have made significant strides by enabling multimodal understanding and generation within language models. Yet, these models incorporate only a single non-textual modality, such as images or audio. While aligning text with one additional modality is relatively straightforward, integrating multiple modalities ($N \geq 3$) within a single framework—and achieving bidirectional alignment among them—poses a more formidable challenge.

Existing explorations in any-to-any multimodal generation have encountered obstacles: some (Tang et al., 2023b) lacked a robust core language model, which impeded the system's reasoning and decision-making capabilities; Others, such as NExT-GPT (Wu et al., 2023), CoDi-2 (Tang et al., 2023a), and Unified-IO2 (Lu et al., 2023), have employed separately pre-trained encoders and decoders. This approach results in representational inconsistencies between the inputs and outputs of the LLMs, which in turn complicates both training and inference processes. Moreover, stabilizing training with such diverse modalities necessitates substantial modifications to existing models and techniques.

To overcome these challenges, we introduce AnyGPT, an any-to-any multimodal language model that employs discrete representations for unified processing. AnyGPT is equipped with multimodal tokenizers that compress raw multimodal data, such as images and audio, into a sequence of discrete semantic tokens. These discrete representations enable the core LLM to unify tasks such as perception, understanding, reasoning, and genera-

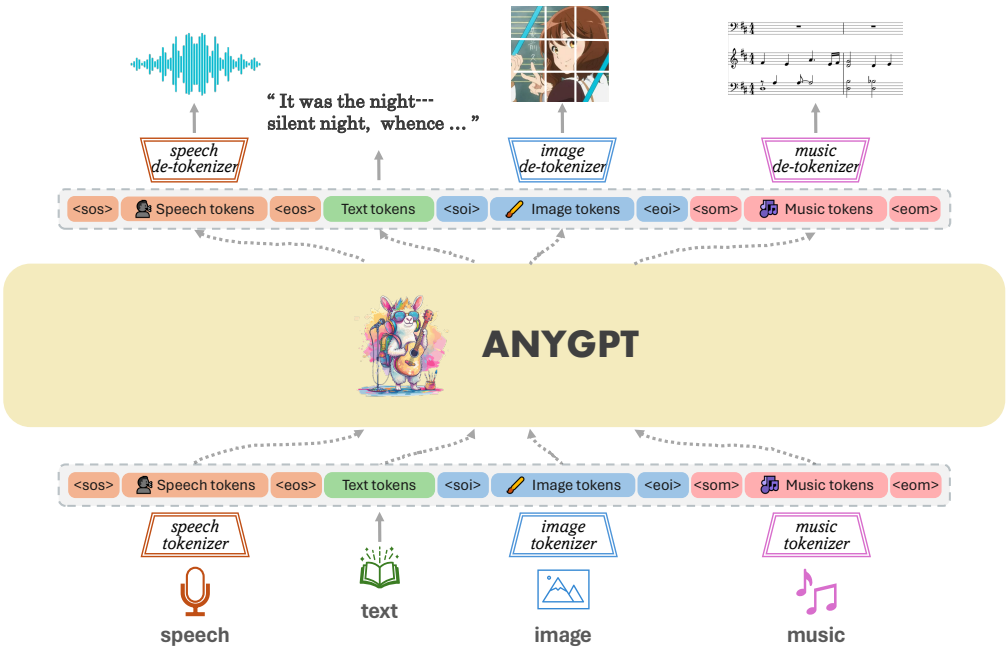


Figure 1: An overview of the AnyGPT model architecture. All modalities are tokenized into discrete tokens, upon which the LLM performs multimodal understanding and generation autoregressively. Only data pre-processing and post-processing are required, with the model's architecture and training objectives remaining unaltered.

tion in an autoregressive manner at the semantic level. Subsequently, de-tokenizers convert the discrete representations back into the original modal representations at the perceptual level. Thanks to discrete representation, which filters out high-frequency, modality-specific perceptual information while preserving essential low-frequency semantic information (Ge et al., 2023a; Borsos et al., 2023a; Rombach et al., 2022), we can train our model stably without any alterations to the existing LLM architecture or training paradigms. Instead, our approach relies solely on data-level preprocessing. This allows for the seamless integration of new modalities into LLMs, akin to the addition of new languages, and permits the direct application of existing LLM tools, which enhances the efficiency of both the training and inference stages.

Furthermore, to mitigate the scarcity of multimodal alignment data encompassing all modalities, we build a text-centric multimodal alignment dataset for pre-training. Our goal is to use text as a bridge, by aligning other modalities with text, to achieve mutual alignment among all modalities, since natural language is the most refined modality of semantic representation and is present in the majority of multimodal alignment datasets. To endow the model with the capability to comprehend and generate content interwoven with multiple modalities, we employ advanced generative models to synthesize a multimodal instruction dataset, AnyInstruct-108k. This dataset, comprising 108k samples of multi-turn conversations, enables AnyGPT to handle arbitrary combinations of multimodal inputs and outputs.

Experimental results demonstrate that AnyGPT achieves zero-shot performance comparable to that of specialized models across various modalities. Furthermore, extensive case studies corroborate AnyGPT's remarkable ability to facilitate any-to-any multimodal dialogue, substantiating the feasibility of using discrete representations to unify multiple modalities.

Our contributions include the following:

- We propose AnyGPT, a token-based any-to-any multimodal language model which can understand and generate various modalities, including speech, text, images, and music.
- One key challenge is the lack of multimodal interleaved instruction-following data. We develop a pipeline using generative models to build AnyInstruct-108k, a dataset comprising 108k multi-turn dialogues with interleaved multimodal elements.
- We demonstrate discrete representations can effectively unify multiple modalities within a language model.

2 Related Work

2.1 Multimodal Large Language Models

To enable cross-modal perception in LLM, a common approach is to connect pre-trained encoders of other modalities as adaptors. However, these models are often limited to text generation.

To empower LLMs with multimodal generation capabilities, Tang et al. (2023b) introduces a frozen text-to-image diffusion model and learns the mapping between the LLM's embeddings and the diffusion model. Sun et al. (2023a) utilizes continuous embeddings to represent the image, calculating either a loss for the next token prediction or the next visual embedding regression. In contrast, SEED-LLaMA (Ge et al., 2023b) trains an image discretization tokenizer to encode the original image into discrete tokens. Through a unified next token prediction task, it achieves unified image understanding and generation. Similarly, in the field of speech, SpeechGPT (Zhang et al., 2023a) enables LLMs to have inherent cross-modal conversation capabilities through discrete speech representation.

To achieve multimodal generation across various modalities on LLMs, NExT-GPT (Wu et al., 2023) utilizes existing high-performance encoders and decoders, connected by a small number of projection layer parameters. However, NExT-GPT does not train the LLM, which may result in suboptimal performance. Moreover, its representation of multimodal input and output lacks a unified form, which poses challenges in unified training and inference.

2.2 Multimodal Discretization

To create a unified multimodal language model, a common approach is to use discretization. A typical method is VQ-VAE (van den Oord et al., 2017). This involves maximizing the restoration of the original representation from the compressed tokens. Some studies (D'efossez et al., 2022; Zeghidour et al., 2021) incorporate residual quantization mechanisms to further enhance fidelity.

In addition to VQ-VAE based tokenizers, some tokenizers focus on extracting high-level semantic representations. Ge et al. (2023b) discretizes the image into semantic-level. The SpeechTokenizer (Zhang et al., 2023b), based on the RVQ-VAE structure, enables the first layer of tokens to retain the semantic information of speech, and the remaining layers to supplement residual information information, achieving a disentanglement of semantic and acoustic information.

3 AnyGPT

Our interest lies in facilitating the generation of any modality to any modality with LLMs. To realize this, we propose a comprehensive framework that can be uniformly trained. As illustrated in Figure 1, this framework is composed of three main components: (1) multimodal tokenizers, (2) a multimodal language model serving as the backbone, and (3) multimodal de-tokenizers. The tokenizers transform continuous non-text modalities into discrete tokens, which are subsequently arranged into a multimodal interleaved sequence. Then the sequences are trained by the language model using the next token prediction training objective. During the inference process, multimodal tokens are decoded back into their original representations by the associated de-tokenizers. To enrich the quality of generation, multimodal enhancement modules can be deployed to post-process the generated results, including applications like voice cloning or image super-resolution. In the following section, we will introduce the details of each module.

3.1 Tokenization

Modality	Image	Speech	Music
Vocab Size	8192	1024	4096
Tokens per Sample	32 / per image	50 / s	200 / s
RVQ	✘	✔	✔
Input Size	224*224	variable duration	5s

Table 1: Details of tokenization of different modalities.

Image Tokenizer We utilize the SEED tokenizer (Ge et al., 2023a) for image tokenization. The SEED tokenizer consists of several components, including a ViT encoder (Dosovitskiy et al., 2021), Causal Q-Former, VQ Codebook (van den Oord et al., 2017), multi-layer perceptron (MLP), and a UNet decoder (Ronneberger et al., 2015). SEED takes a 224×224 RGB image as input, and the ViT encoder encodes the image into 16×16 patches, then the Causal Q-Former converts the patch features into 32 causal embeddings. A codebook with 8192 entries discretizes the embeddings into a sequence of quantized codes. An MLP is employed to decode the visual codes into a generation embedding, which is aligned with the latent space of the pre-trained unCLIP Stable Diffusion(unCLIP-SD) (Rombach et al., 2022).

Finally, the UNet decoder is used to restore the generation embedding to the original image.

Speech Tokenizer The tokenizer for speech we utilize is SpeechTokenizer (Zhang et al., 2023b), adopting an encoder-decoder architecture with residual vector quantization (RVQ). The SpeechTokenizer compresses single-channel audio sequences into a discretized matrix using eight hierarchical quantizers, each with 1,024 entries, and achieves a frame rate of 50 Hz. The first quantizer layer captures semantic content, while layers 2 to 8 encode paralinguistic details. A 10-second audio is thus transformed into a 500×8 matrix, splitting into semantic and acoustic tokens. We adopt a SpeechTokenizer variant pre-trained on the Commonvoice (Ardila et al., 2020) and Librispeech (Panayotov et al., 2015) datasets.

In AnyGPT, the Large Language Model (LLM) is employed to model the semantic tokens, while a voice cloning model supplements the remaining paralinguistic information. As a result, the size of the voice vocabulary in the LLM is equivalent to the size of one codebook, which is 1024. Further details will be discussed on in Section 3.3.

Music Tokenizer Although speech and music share similar data formats, their substantial content differences lead us to treat them as distinct modalities, each equipped with its own tokenizer. For music, we employ Encodec (D'efossez et al., 2022), a convolutional auto-encoder with a latent space quantized using Residual Vector Quantization (RVQ), as the music tokenizer. We use an available off-the-shelf variant of the Encodec[1] pre-trained on 20k pieces of music tracks. This variant processes 32 kHz monophonic audio, and achieves a frame rate of 50 Hz. The embeddings generated are quantized using an RVQ with four quantizers, each with a codebook size of 2048, resulting in a combined music vocabulary size of 8192.

We encode 5 seconds music into 250 latent frames, ultimately generating a 250×4 codes matrix. To enable the language model predict entire music clip, we flatten the 4-layer music codes into a causal sequence in a frame-by-frame manner. The language model begins by predicting the initial four tokens of the first frame and continues in a similar fashion for the subsequent frames.

[1] https://huggingface.co/facebook/encodec_32khz

3.2 Language Model Backbone

Expanding vocabulary To incorporate multimodal discrete representations into pre-trained LLMs, we expand the vocabulary with new modality-specific tokens, and consequently extend the corresponding embeddings and prediction layer, the newly incorporated parameters are initialized randomly. The tokens from all modalities combine to form a new vocabulary, where each modality is trained within the language model to align in a shared representational space. The size of this enhanced vocabulary, denoted by V, is the summation of the vocabulary sizes across all modalities, that is, $V = \sum_{i=1}^{n} V_i$, where V_i signifies the vocabulary size of the i-th modality.

Unified Multimodal Language Model Equipped with the modality-specific tokenizers, we can compress multimodal data into discrete token sequences, which can be trained by the language model using the next token prediction loss. This naturally enables the core LLM to unify tasks such as perception, understanding, reasoning, and generation in an autoregressive manner.

We employ the LLaMA-2 (Touvron et al., 2023) 7B as the backbone, which is pre-trained on 2 TB of text tokens. Apart from reshaping the embedding matrix and prediction layer, the rest of the language model remains unaltered.

3.3 Multimodal Generation

The generation of high-quality multimodal data, including high-definition images, and high-fidelity audio, presents a substantial challenge. These data typically necessitate a large number of bits for accurate representation, resulting in long sequences which is particularly demanding for language models, as the computational complexity increases exponentially with the length of the sequence.

To tackle this, we adopt a two-stage framework for high-fidelity generation, comprising semantic information modeling and perceptual information modeling. First, the language model is tasked with generating content that has undergone fusion and alignment at the semantic level. Then, non-autoregressive models convert multimodal semantic tokens into high-fidelity multimodal content at the perceptual level, striking a balance between performance and efficiency.

Specifically, we employ SEED tokens, aligned with the diffusion latent space, for visual language modeling. Semantic-level SEED tokens

are decoded into high-quality images by a Diffusion Model, which is renowned for its superior generation capabilities. For speech, we utilize SoundStorm (Borsos et al., 2023b), a non-autoregressive Masked Language Model, trained to generate SpeechTokenizer's acoustic tokens from semantic tokens. We train a variant of Soundstorm, which is trained using the SpeechTokenizer on the Multilingual LibriSpeech(MLS) dataset (Pratap et al., 2020). Subsequently, the SpeechTokenizer's decoder transforms all speech tokens into raw audio data. This approach enables AnyGPT replicate the voice of any speaker using a 3-second speech prompt, while significantly reducing the length of the voice sequence for LLM. For music, we employ Encodec tokens to filter out high-frequency details beyond human perception, and then use the Encodec decoder to reconstruct these tokens into high-fidelity audio data.

4 Multimodal Data

4.1 Pre-training Data

To enable the generation from any modality to any other, it is crucial to have data that is well-aligned across these modalities. Unfortunately, such data is notably scarce. To address this challenge, we build a text-centric bi-modal alignment dataset. Here, text is employed as a vital intermediary to bridge the gap between various modalities. By aligning different modalities with the textual modality within a language model, we aim to achieve mutual alignment amongst all modalities.

The representational forms and types of information vary greatly across different modalities, To facilitate a standardized comparison of data volumes across various modalities, we have adopted a quantification approach based on token counts. Figure 2 presents all the data used in pre-training and their respective proportions. A certain level of oversampling is applied to modalities with comparatively lower data quantities, to attain a balanced representation of diverse data types within a single batch. More details are in Appendix A.1.

Image & Text We utilized image-text pairs from LAION-2B (Schuhmann et al., 2022), LAION-COCO (lai, b), LAION-Aesthetics (lai, a) and JouneyDB (Pan et al., 2023). LAION-2B provides images paired with noisy alt-texts sourced from the web, while LAION-COCO represents a 600M subset of this, captioned by BLIP. We refined these datasets by filtering for text quality, image aspect ratio, and clip score, etc., yielding a high-quality corpus of 300M pairs. To enhance the overall image generation fidelity, we supplement our data with the high-quality LAION-Aesthetics subset and the synthetic dataset JourneyDB from Midjourney.

We also incorporate image-text interleaved data to adapt the model to an interleaved mode. We deploy the Multimodal-C4 (MMC4) dataset (Zhu et al., 2023), an enhanced version of the text-only C4 (Raffel et al., 2020). Specifically, we utilize the MMC4-core split, consisting of 7.3M documents.

Figure 2: Pre-training data distribution, segmented by token counts, with the inner section indicating the modality, the middle detailing data types, and the outer specifying individual datasets.

Speech & Text We collect several large-scale English Automatic Speech Recognition (ASR) datasets, including Gigaspeech (Chen et al., 2021), Common Voice (Ardila et al., 2020), and Multilingual LibriSpeech(MLS) (Pratap et al., 2020). These datasets are sourced respectively from online platforms, volunteer crowdsourcing, and audiobooks, collectively constituting a corpus of 57,000 hours of speech-text pairs, encompassing a wide variety of speakers, domains, and recording environments.

Music&Text We embark on an extensive data collection process by crawling over one million music videos from the Internet. The core step involves matching the titles of these videos with corresponding songs using the Spotify API. Subsequently, we harvest a comprehensive set of metadata for each music audio, including video titles, descriptions, keywords, playlist names, and Spotify lyrics. This metadata is formatted into JSON and fed into GPT-4 (Achiam et al., 2023) for processing. GPT-4's role is pivotal as an intelligent caption generator; it

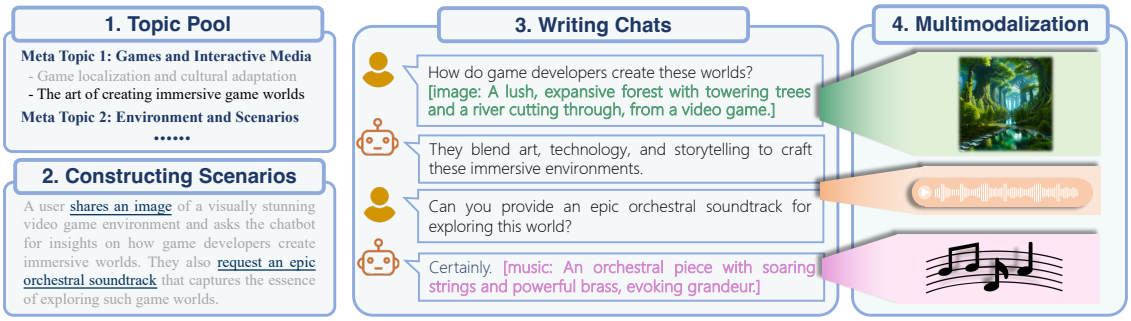


Figure 3: The construction process of the multimodal interleaved instruction dataset AnyInstruct is divided into two stages: Generation of text-based conversations incorporating multimodal elements and Text-to-Multimodality Conversion. The first stage generates topics, scenarios, and textual dialogues, while the second stage produces the final multimodal dialogues.

utilizes the noisy metadata to extract meaningful information and succinctly summarize it into coherent sentences. This approach allows us to generate high-quality text captions for a large amount of music audio, effectively minimizing the occurrence of hallucinations in the dataset.

Training Sample Construction. To train the Language Model (LM), we employ various templates to construct multimodal sentences, which the LM then processes autoregressively. Further training details can be found in Appendix A.2. Additionally, We observe significant variation in sentence lengths across different modalities and datasets. To enhance training efficiency, samples from the same dataset are concatenated into a long sequence, adhering to the model's maximum sequence length. Consequently, each token in the sequence contributes to the loss.

4.2 Multimodal Interleaved Instruction Data Construction

Effective human-machine interaction should permit the exchange of information in a variety of interleaved modalities. However, the increasing number of modalities in conversation significantly complicates the data collection process. To our knowledge, there is currently no large-scale instruction dataset involving more than two modalities. This poses a significant limitation on the development of a comprehensive model capable of managing dialogues with multiple, intertwined modalities.

To overcome this limitation, we draw inspiration from the most recent research on data synthesis (Wang et al., 2022; Wu et al., 2023), and build a dataset comprised of 108k multi-turn conversation samples with generative models. With careful curation, each synthetic conversation integrates multiple modalities—text, speech, images, and music—in an interleaved manner. Specifically, our data synthesis process is carried out in two stages, as illustrated in Figure 3.

Generation of text-based conversations incorporating multimodal elements. In this phase, we employ GPT-4 to generate a series of text-based conversations. Notably, we incorporate non-text modality in the form of their textual descriptions within these conversations. To ensure high-quality data at scale, we divide this stage into four steps. **(1)** Initially, we brainstorm 100 meta topics to cover a broad spectrum of scenarios related to audio-visual elements. **(2)** Next, we employ GPT-4 to expand these meta-topics into 20,000 specific topics. **(3)** Subsequently, we prompt LLM to generate specific dialogue scenarios based on these topics. Acknowledging the intrinsic constraints of a text-based LLM in generating multimodal elements, we prepare several demonstrations that encompass as many modality combinations as possible. While generating scenarios, a subset is sampled from this demonstration pool, serving as examples for the LLM. This approach guides the model to effectively synthesize varied and contextually appropriate conversational scenarios. **(4)** Finally, we utilize GPT-4 to generate multi-turn conversations derived from scenarios. In these synthesized dialogues, multimodal elements, including images and music, are depicted through detailed textual representations. We curate a diverse range of conversation examples, similar to scenario generation, to prompt the model into creating dialogues with the widest possible variety of modalities. As a result, we compiled a substantial corpus of multimodal conversa-

tional data in solely textual format.

Text-to-Multimodality Conversion. In this phase, we employ advanced generative models to convert textual descriptions into multimodal elements. We use OpenAI's DALL-E-3 (Betker et al., 2023) for image generation, MusicGen (Copet et al., 2023) for music composition, and Microsoft Azure's text-to-speech API (Microsoft) for speech synthesis from user's instructions and model's text responses.

After filtering, we obtain a dataset of 108k high-quality multimodal dialogues, featuring a variety of multimodal combinations. This dataset includes around 205k images, 503k voice recordings, and 113k music tracks. Additionally, we enhanced our dataset by extracting dialogues from existing text-only instruction datasets well-suited for spoken narration. This results in 100k voice dialogues through the employment of text-to-speech models.

The two-stage approach efficiently collected a diverse array of high-quality multimodal conversations at scale. Appendix D provides the prompts used during the data synthesis process.

5 Experiment

5.1 Evaluation

We evaluate the fundamental capabilities of the pre-trained base AnyGPT (Section 3), covering multimodal understanding and generation tasks for all modalities. This evaluation aimed to test the alignment between different modalities during the pre-training process. Specifically, we test both text-to-X and X-to-text tasks for each modality, where X is image, music, and speech separately.

To simulate real-world scenarios, all evaluations are conducted in a *zero-shot* mode. This means that AnyGPT will be *not fine-tuned nor pre-trained* on downstream training samples during the evaluation process. This challenging evaluation setting requires the model to generalize to an unknown test distribution, showcasing the generalist abilities of AnyGPT across different modalities. The evaluation results demonstrate that AnyGPT, as a generalist multimodal language model, achieves commendable performance on various multimodal understanding and generation tasks.

5.1.1 Image

Image Understanding We assess the image comprehension capabilities of AnyGPT on the image captioning task. The comparison results are presented in Table 2. We utilize the MS-COCO 2014 captioning benchmark (Lin et al., 2014) and adopt the *Karpathy split testset* following previous studies (Li et al., 2023; Tang et al., 2023b).

Method	CIDEr ↑
Flamingo (9B) (Alayrac et al., 2022)	79.4
Flamingo (80B)	84.3
Emu (14B) (Sun et al., 2023b)	112.4
DreamLLM (8B) (Dong et al., 2023)	115.4
InstructBLIP (14B) (Dai et al., 2023)	102.2
SEED-LLaMA (8B) (Ge et al., 2023b)	**123.6**
AnyGPT (8B)	107.5

Table 2: Comparison results on image captioning task. Results in grey indicate models have trained on training samples.

Method	CLIP_{score} ↑
GILL (Koh et al., 2023)	0.67
Emu	0.66
SEED-LLaMA	**0.69**
AnyGPT	0.65

Table 3: Comparison results on text-to-image generation task. We adopt MS-COCO captions to generate images and calculate the CLIP similarity score (CLIP_{score}) for evaluation.

Image Generation The results of the text-to-image generation task are presented in Table 3. To ensure consistency with previous research (Koh et al., 2023; Ge et al., 2023b; Sun et al., 2023a), we randomly select 30k images from the MS-COCO validation set and use CLIP_{score} as the evaluation criterion. This metric computes a similarity score between the generated image and its corresponding caption from a real image, based on CLIP-ViT-L (Radford et al., 2021).

5.1.2 Speech

Method	WER ↓	SIM ↑
Ground Truth	*1.9*	*0.93*
VALL-E (Wang et al., 2023)	7.9	0.75
USLM (Zhang et al., 2023b)	**6.5**	**0.84**
AnyGPT	8.5	0.77

Table 4: Comparison results on TTS task.

ASR We evaluate the performance of AnyGPT on the Automatic Speech Recognition (ASR) task by calculating the Word Error Rate (WER)

on the test-clean subsets of the LibriSpeech dataset (Panayotov et al., 2015). We use Wav2vec 2.0 and Whisper Large V2 as baselines. Wav2vec 2.0 is pre-trained with 60,000 hours of speech and fine-tuned on LibriSpeech, while Whisper Large V2 is evaluated in a zero-shot setting but is trained with 680,000 hours of speech. The results are shown in Table 5.

Method	WER ↓
Human-level (Amodei et al., 2016)	5.8
Wav2vec 2.0 (Baevski et al., 2020)	2.7
Whisper Large V2 (Radford et al., 2022)	**2.7**
AnyGPT	8.5

Table 5: Comparison results on ASR task. We use Word Error Rate (WER) as the metric.

TTS We conduct a zero-shot Text-to-Speech (TTS) evaluation on the VCTK dataset. The results are presented in Table 4. We evaluate the TTS systems with speaker similarity and Word Error Rate (WER), where WER is focused on speech quality. More experimental details can be found in Appendix C.

5.1.3 Music

Method	CLAP_{score} ↑
Music understanding	
<music, real caption>	0.16
<music, generated caption>	0.11
Music generation	
Riffusion (Forsgren and Martiros, 2022)	0.19
Mousai (Schneider et al., 2023)	**0.23**
AnyGPT	0.14

Table 6: Comparison results for music understanding and generation tasks. A metric scoring the alignment between music and textual captions is reported (CLAP_{score}). For music captioning, the CLAP_{score} of <music, real caption> pair and <music, generated caption> pair are computed for comparison.

we evaluate AnyGPT's performance on the MusicCaps benchmark (Agostinelli et al., 2023) for both music understanding and generation tasks. We utilize the CLAP_{score} (Wu et al., 2022; Huang et al., 2023) score as the objective metric, which measures the similarity between the generated music and a textual description.

For the evaluation of music captioning, we found that existing objective metrics may be limited in expressing the performance in the music captioning task. The diversity and subjectivity of music lead to varying opinions from individuals. Only specific music genres and instruments possess distinctive characteristics that can be easily recognized. While recent studies (Gardner et al., 2023) have explored this issue, it remains a challenging problem to address. To ensure an objective evaluation, we compute CLAP_{score} of <music, real caption> pairs and <music, generated caption> pairs for comparison. These scores are averaged across the entire test set.

5.2 Example Demonstrations

After fine-tuning on the AnyInstruct-108k dataset, AnyGPT demonstrates the capability and potential in any-to-any multimodal dialogue. We provide compelling conversation examples of AnyGPT in Appendix E. These examples showcase AnyGPT is capable of comprehending and reasoning contents across various modalities in any combination. Specifically, AnyGPT can comprehend instructions interwoven with multiple modalities, including text, voice, images, and music, and can adeptly select the appropriate multimodal combination for its reply. The two-stage framework of semantic-acoustic hierarchical modeling empowers AnyGPT to generate voice responses that matches the timbre and emotion of a 3-second speech prompt. For additional examples and to experience the speech and music content, we highly recommend visiting the demo page.

6 Conclusion

In this work, we introduced AnyGPT, an any-to-any multimodal language model that utilizes discrete representations for the unified processing of various modalities, including speech, text, images, and music. Discrete multimodal representations facilitate a seamless integration of new modalities—comparable to incorporating a foreign language—without necessitating alterations to the existing LLM architecture or training paradigms. To equip the model to handle arbitrary combinations of multimodal inputs and outputs, we synthesize the first large-scale any-to-any multimodal instruction dataset, AnyInstruct-108k, consisting of multi-turn conversations that intricately interweave various modalities. Experimental results indicate that AnyGPT achieves promising results in various cross-modal tasks and demonstrates that discrete representations can effectively and conveniently unify multiple modalities within a unified large language model.

Limitations and Future Work

Any-to-Any Multimodal LLM Benchmark The domain of any-to-any multimodal large language models (LLMs) is an emerging field of research. However, the lack of a dedicated benchmark to evaluate the models' capabilities across multiple dimensions, as well as to mitigate potential risks, presents a considerable challenge. Consequently, the development of a comprehensive benchmark is imperative.

Enhancing LLMs Although the multimodal LLMs with discrete representations can be trained stably, a higher loss is observed compared to unimodal training, preventing optimal performance in each modality. Potential strategies to improve multimodal fusion could involve scaling LLMs and tokenizers or adopting a Mixture-Of-Experts (MOE) architecture to better manage diverse data and optimize performance.

Better Tokenizer In multimodal LLMs employing discrete representations, the tokenizer's quality sets a ceiling for the model's comprehension and generative potential. Enhancing the tokenizer can be approached from various angles, including the adoption of superior codebook training methods, the development of more cohesive multimodal representations, and the application of information disentanglement across various modalities.".

Longer Context Multimodal content, such as images and audio, often spans extensive sequences. AnyGPT, for instance, limits music modeling to 5 seconds, significantly restricting the practical usefulness of its audio output. Moreover, for any-to-any multimodal dialogue, an extended context allow for a higher number of conversational exchanges, thereby enriching the interaction's depth and complexity.

References

a. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/.

b. Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse 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. *ArXiv preprint*, abs/2301.11325.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023a. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *ArXiv preprint*, abs/2305.09636.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel

Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3670–3674. ISCA.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *ArXiv preprint*, abs/2306.05284.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, abs/2305.06500.

Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *ArXiv preprint*, abs/2210.13438.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jian-Yuan Sun, Hongyu Zhou, Hao-Ran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *ArXiv preprint*, abs/2309.11499.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Seth* Forsgren and Hayk* Martiros. 2022. Riffusion - Stable diffusion for real-time music generation.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. 2023. Llark: A multimodal instruction-following language model for music.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023a. Planting a seed of vision in large language model. *ArXiv preprint*, abs/2307.08041.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023b. Making llama see and draw with seed tokenizer. *ArXiv preprint*, abs/2310.01218.

Rongjie Huang, Jia-Bin Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiaoyue Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *ArXiv preprint*, abs/2301.12661.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models. *ArXiv preprint*, abs/2305.17216.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

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 *European Conference on Computer Vision*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *ArXiv preprint*, abs/2312.17172.

Microsoft. Microsoft azure text-to-speech api. `https://azure.microsoft.com/en-us/products/ai-services/ai-speech`.

Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. 2023. Journeydb: A benchmark for generative image understanding. *ArXiv preprint*, abs/2307.00716.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2757–2761. ISCA.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv preprint*, abs/2212.04356.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-music generation with long-context latent diffusion. *ArXiv preprint*, abs/2301.11757.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023a. Generative multimodal models are in-context learners. *ArXiv preprint*, abs/2312.13286.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative pretraining in multimodality. *ArXiv preprint*, abs/2307.05222.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023a. Codi-2: In-context, interleaved, and interactive any-to-any generation. *ArXiv preprint*, abs/2311.18775.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023b. Any-to-any generation via composable diffusion. *ArXiv preprint*, abs/2305.11846.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315.

Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *ArXiv preprint*, abs/2301.02111.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *ArXiv preprint*, abs/2212.10560.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *ArXiv preprint*, abs/2309.05519.

Yusong Wu, K. Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Conference on Empirical Methods in Natural Language Processing*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechtokenizer: Unified speech tokenizer for speech large language models. *ArXiv preprint*, abs/2308.16692.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *ArXiv preprint*, abs/2304.06939.

A pretraining

A.1 Data

Modality	Dataset	Description	Sample Rate
Interleaved Image-Text	MMC4-core-ff	101M image-interleaved documents collected from Common Crawl. We use the mmc4-core split which is consist of 7.3M documents.	0.05
Image-Text	Laion-2B	2B image-text pairs from web.	0.3
	LAION-COCO	600M image-text pairs, where the caption is generated by BLIP.	
	JourneyDB	4429K Midjourney images, with image caption.	
	LAION-Aesthetics	Several collections of subsets from LAION 5B with high visual quality.	
Speech-Text	Multilingual Librispeech	Processing audiobooks read from LibriVox, we used a 44,000-hour subset of English.	0.13
	CommonVoice	Microphone recordings from internet volunteers, of which we used a 3000-hour subset of English.	0.27
	GigaSpeech	10,000 hours of English voice data sourced from audiobooks, podcasts, and YouTube videos.	
Music-Text	Youtube-Music-1M	100M music-text pairs from Youtube.	0.25
	MusicGen-Synthesis	20k music-text pairs extracted from the AnyInstruct-108k dataset, synthesized by MusicGen.	

Table 7: Details of data used in pre-training stage.

A.2 pre-training

We employ various templates to construct multimodal sentences, ensuring a diverse spectrum within our pre-training data. Each non-text modality content is identified by special tokens placed at both the beginning and end. Typically, the paired data comprises a non-text modality (X) - such as images,

	Pre-training Stage	Fine-tuning Stage
Gradient clipping (Global-norm)	1.0	1.0
Batch size	480	64
Max length	4500	4500
Training steps	81000	5000
Learning rate scheduler	cosine	cosine
Peak learning rate	6e-5	2e-5
Warmup ratio	0.03	0.03
Optimizer	Adam	Adam
GPU	A100	A100

Table 8: Training hyperparameters used in experiments.

speech, or music - and its corresponding text, which could be a caption or transcription. We prompt OpenAI GPT-4 to generate hundreds of bidirectional instructions, specifically X-to-text or text-to-X such as "Please generate an image based on the provided text." Given a token sequence (S) and related text (T), we randomly pick a generation direction alongside an instruction (I) from our pre-established pool, forming a triplet (I, S, T). This triplet is then incorporated into a sequence using the template
[Human]: $\{I\}.\{S\}$**<eoh>. [AnyGPT]:** $\{T\}$**<eos>.** or its variant
[Human]: $\{I\}$. This is input:$\{T\}$**<eoh>. [AnyGPT]:** $\{S\}$**<eos>.**, depending on the generation direction.

For interleaved multimodal data, like a web document with interspersed images and text, we directly replace non-text content with the corresponding tokens sequence as they naturally form sentences.

As most of the image and music data are sourced from the web, there is a certain level of noise that can affect the quality of multimodal generation. Consequently, after the initial pre-training, we selectively utilized high-quality datasets—JourneyDB and LAION-Aesthetics for text-to-image generation, and LAION-COCO for image captioning. For music data, we incorporated the AnyInstruct-108k dataset. The remaining data were kept unchanged, and we continued to pre-train the model for an additional 4000 steps.

We report the detailed training hyperparameters of AnyGPT in Tab 8.

B Instruction Tuning

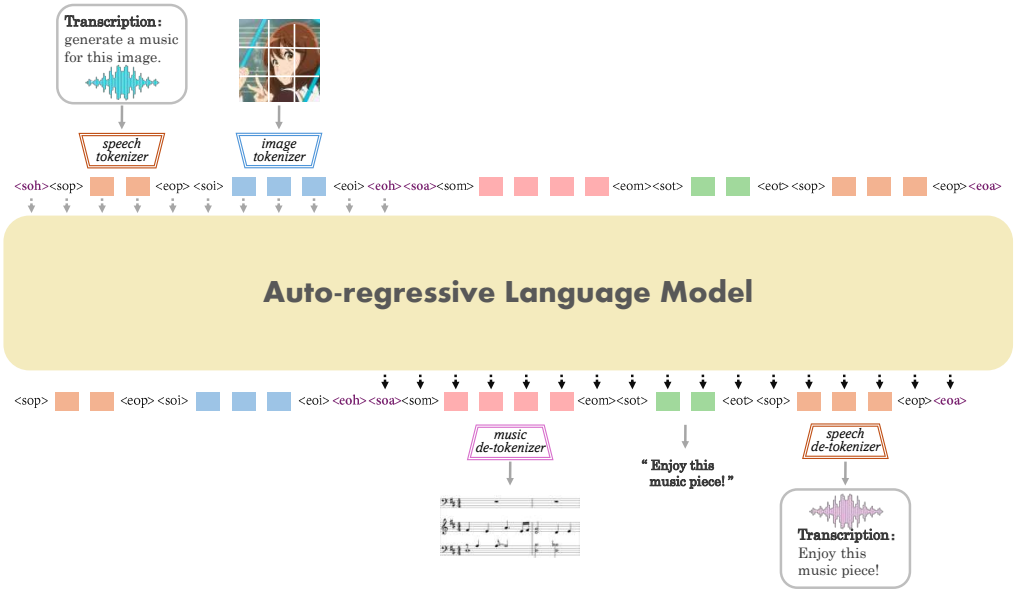


Figure 4: An example of an AnyGPT multimodal dialogue: the input is an image and a voice command to generate music. The output is music that meets the requirements, along with corresponding text and voice responses. All data are processed into discrete tokens and are autoregressively processed by the LLM.

C Evaluation

Generated Modality	Text	Image	Speech	Music
Decoding Strategy	Beam Search	Sampling	Sampling	Sampling
Beam size	5	-	-	-
Top-P	-	0.7	0.7	1.0
Repetition Penalty	1.0	1.0	1.0	1.15

Table 9: Details of generation decoding strategies used in evaluation.

We conduct a zero-shot Text-to-Speech (TTS) evaluation on the VCTK dataset. There is no overlap in speakers between our training data and the VCTK dataset. We randomly select a 3-second clip from each speaker as the vocal prompt along with a separate text as input.

The results can be found in Table 4. We evaluate the TTS systems with speaker similarity and WER. To evaluate the speaker similarity between the generated speech and the prompt speech, we employ WavLM-TDNN[2]. It can generate speaker embeddings for both the generated speech and the prompt

[2] https://github.com/yangdongchao/UniAudio/blob/main/UniAudio/tools/evaluation/compute_similarity_vc.py

speech, then compute the cosine similarity between these embeddings. WER is calculated using the Whisper medium model to transcribe the generated speech, with lower WER indicating higher quality of the synthesized speech.

We compare our model with VALL-E and USLM, both of which employ two autoregressive models for speech modeling. They utilize Encodec and SpeechTokenizer, respectively, as speech tokenizers.

D Prompts for Constructing Multimodal Interleaved Instruction Data

In the first stage of our pipeline to construct multimodal interleaved instruction data (Sec. 4.2) with GPT4. To facilitate reproducibility, we detail our prompts to the language model for brainstorming a topic pool (Fig. 5), constructing chatting scenarios (Fig. 5), and detailing the chat contents (Fig. 7), with multimodal content written as their text descriptions.

Prompt: Please list me 50 **non-academic** conversation topics about `{metatopic}` between an ordinary person and a helpful chatbot. Each topic should be made up of 1-10 words and the conversation contain understanding and generation of images or music.

GPT4: `{50 sampled topics}`

Prompt: continue

GPT4: `{50 more sampled topics}`

.....

Figure 5: Prompts for brainstorming chat topics. We prepare 50 `metatopic` and repeat the conversation for 4 rounds for each of them. This gives 200 topics per metatopic and a total of 10,000 topics in our final topic pool.

Prompt:
You are a creative assistant. I am now asking you to help me brainstorm some chatting scenarios where the user asks the agent for help.
Note that the scenarios should be between ordinary people and a helpful chatbot, and it should not be an academic discussion!
During the conversation, the speakers can use images or music to help convey information (but do not use video!). And the user should ask questions about it if he/she provides an image or a piece of music.
Note that the image should not be a chart.
Note that the images and music should not be the famous masterpieces that may arise copyright issues.

Here are some of my ideas, and please show me more in the same format as mine.

`{demonstrations}`

Here's the topics for you to try: `{topics}`

Now it's your turn. In these scenarios, `{requirements}`.

GPT4:
`{synthetic scenarios of the provided topics, following requirements}`

Figure 6: Prompts for constructing chat scenarios. In each API call, we sample 5 different `{demonstrations}`, with each containing a topic and detailed description of the scenarios. And we sample 10 different `{topics}` for GPT4 to synthesize scenarios. To ensure the diversity of user and chatbot actions, we explicitly sample `{requirements}` from "the user provide images", "the user share music", "the user asks for music", and "the user asks for images". We up weight "the user share music" as we observe that the model tends to omit this requirement.

Prompt:
You are helping me to write conversations about a user talking to a chatbot named AnyGPT.

In the conversations, both the user can provide images or music to help express her/his needs and ideas. And the chatbot MMGPT can also respond to the user with images or music in its utterances.

The images and music in the chat are in the form of image descriptions and music descriptions like [image: description] and [music: description], respectively. The user should provide images and music in this format and the chatbot will respond to the user like this as well.

Note that at most one music appears in one conversation and the description of music should be straightforward, focusing on genres and instruments, and never mention a known music directly.

Before each conversation, I will first show you a scenario and you can start writing about the chat.

Here is an example:
—
`{demonstrations}`
—

Now it's your turn for the next conversation. You only need to answer following the format in which the user and AnyGPT take turns.
The conversation should be consistent with the introduction to the scenario.
Remember that the utterances should be concise, try to use 5-15 words per utterance.
Note that: the user utterance should always be a question or instruction.
In some turns, the user provides an image or a piece of music and asks a question or makes an instruction to AnyGPT relating to the provided image or music.
In other turns, the user requests AnyGPT to generate the required images or music.
Note that: the description of music should focus on genres, style, and instruments. And make the description of images and music within [image:] or [music:] more detailed.
Note that: never directly include a famous person's name in the image descriptions or mention a piece of known music in the music description.
Tips: when the user asks to convert between images and music, AnyGPT should first utter his understanding of the input image or music before generating the requested result.
Keep the dialog in 2 or 3 rounds.
Each dialog should include one music and at most 2 images.

—
In this conversation, `{new_scenario_description}`

GPT4:
`{A synthetic chat according to the scenario description.}`

Figure 7: Prompts for writing chat content. For each API call, we sample 3 demonstrations. Each demonstration contains a scenario description, as the {new_scenario_description}, and the corresponding chat.

E Examples Demonstration

Figure 8: Speech conversations (Voice Clone)

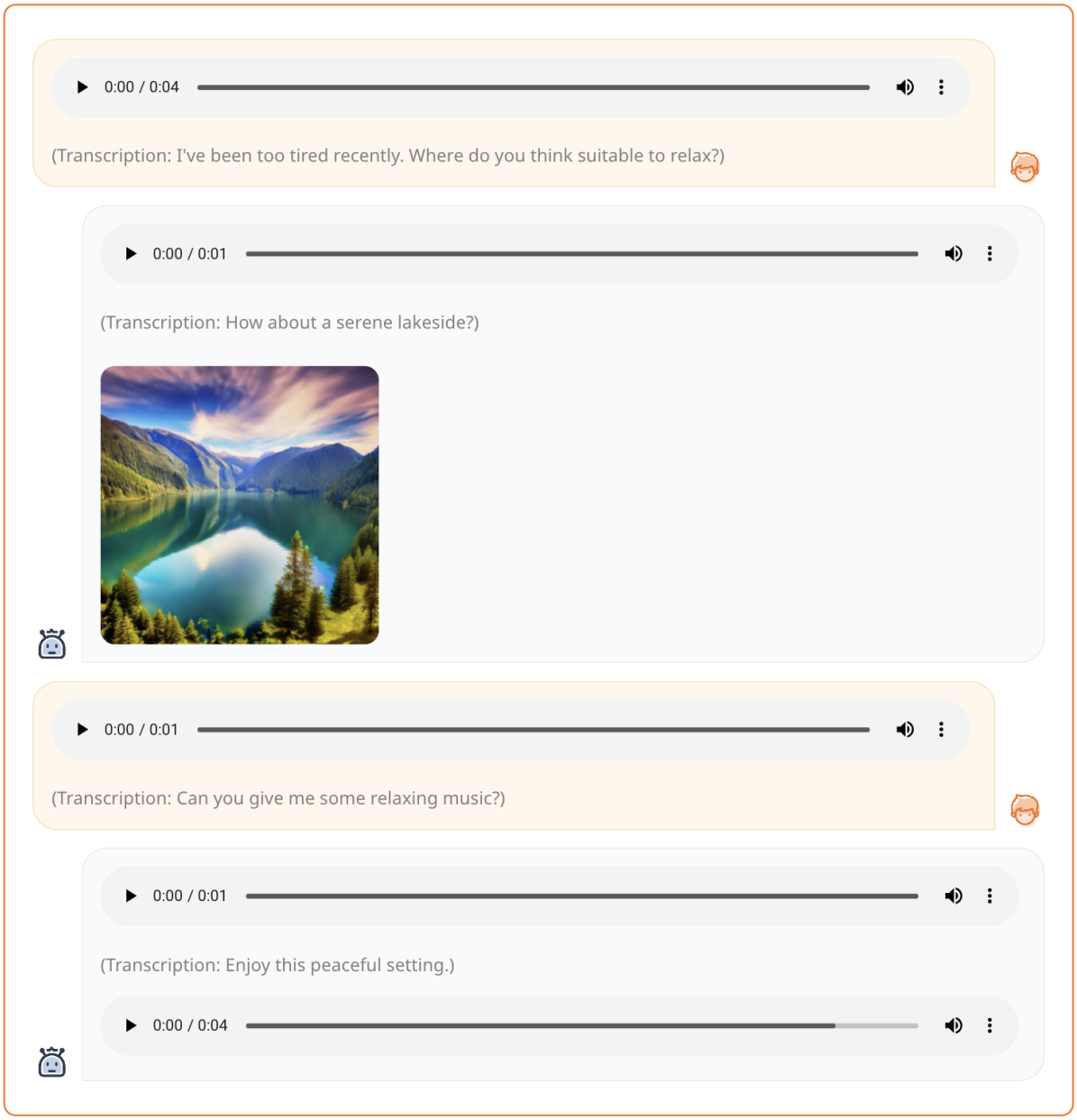


Figure 9: Speech Instruction + Image → Text + Music + Speech Response

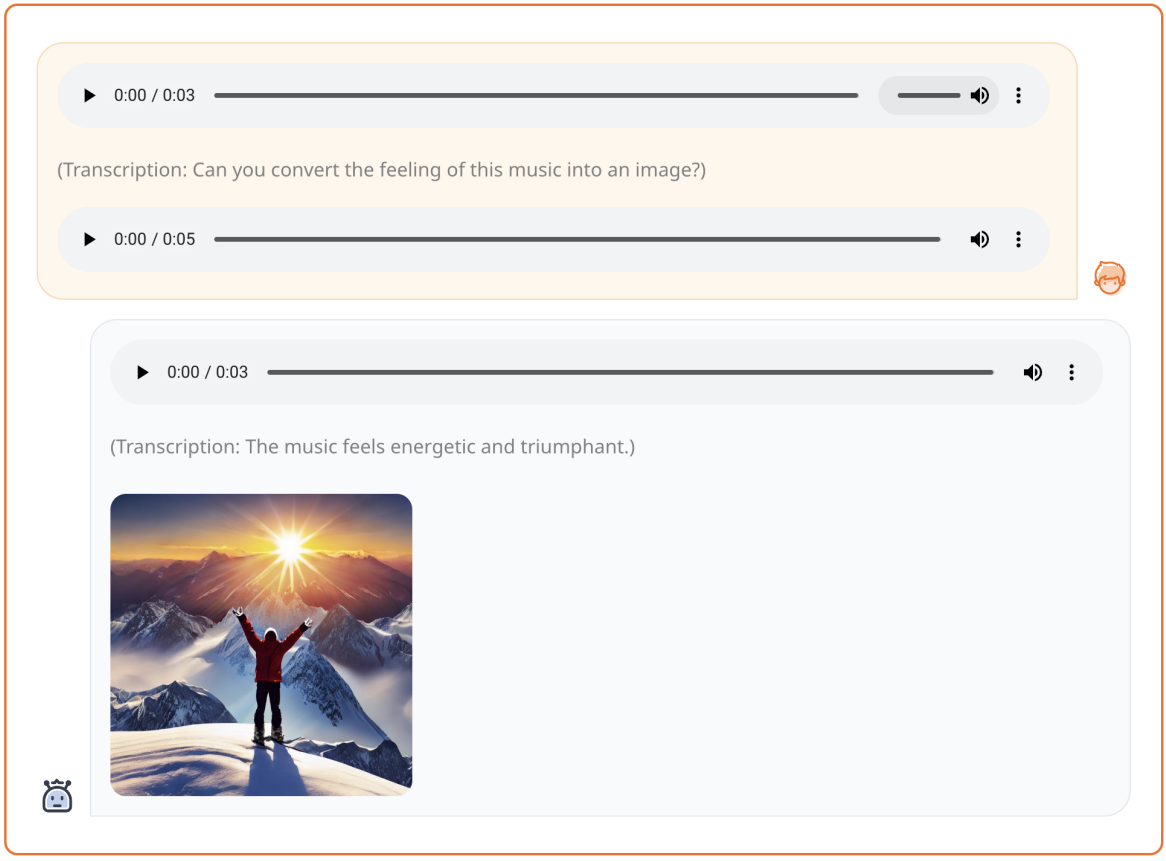


Figure 10: Speech Instruction + Music → text + Music + Speech Response

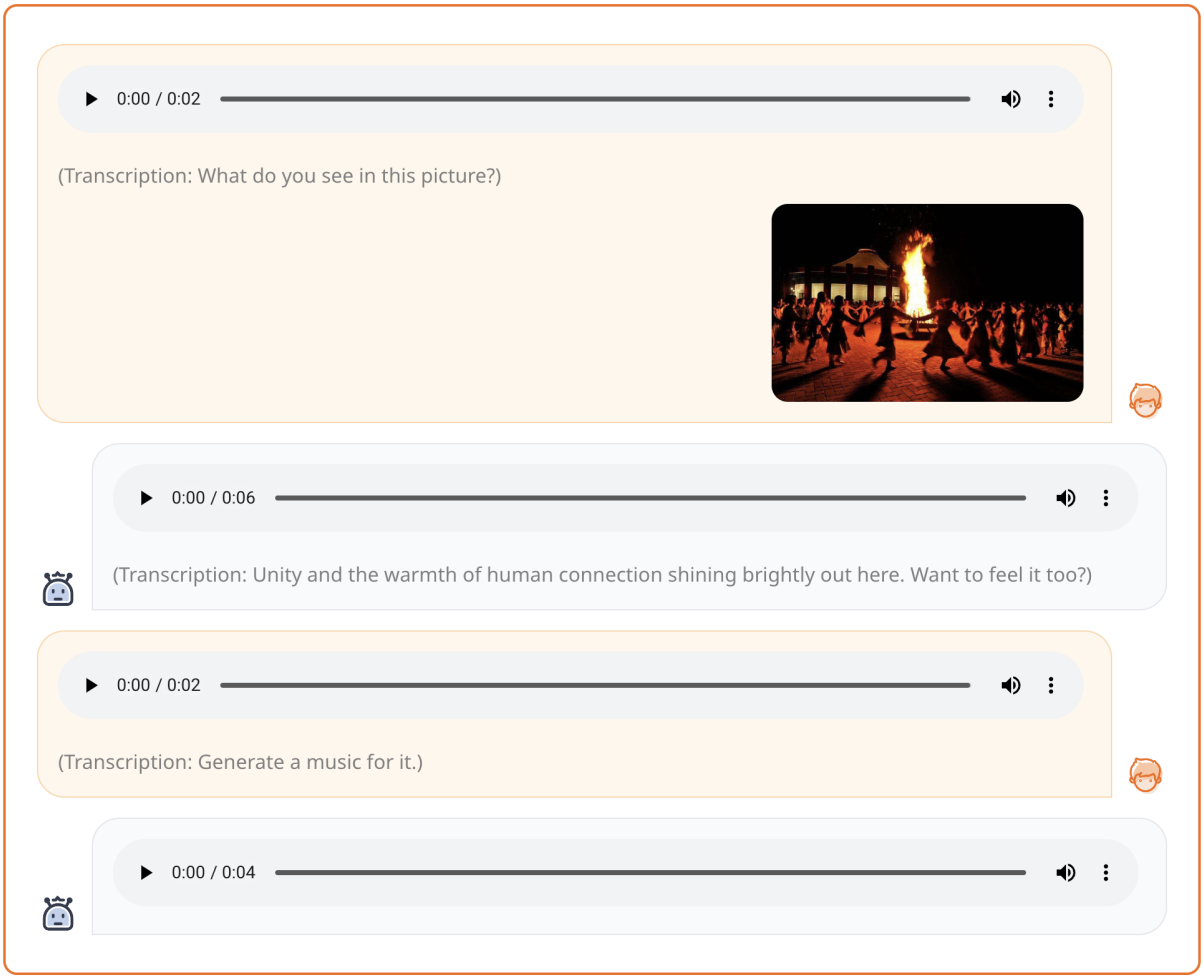


Figure 11: Speech Instruction + Image → text + Music + Speech Response

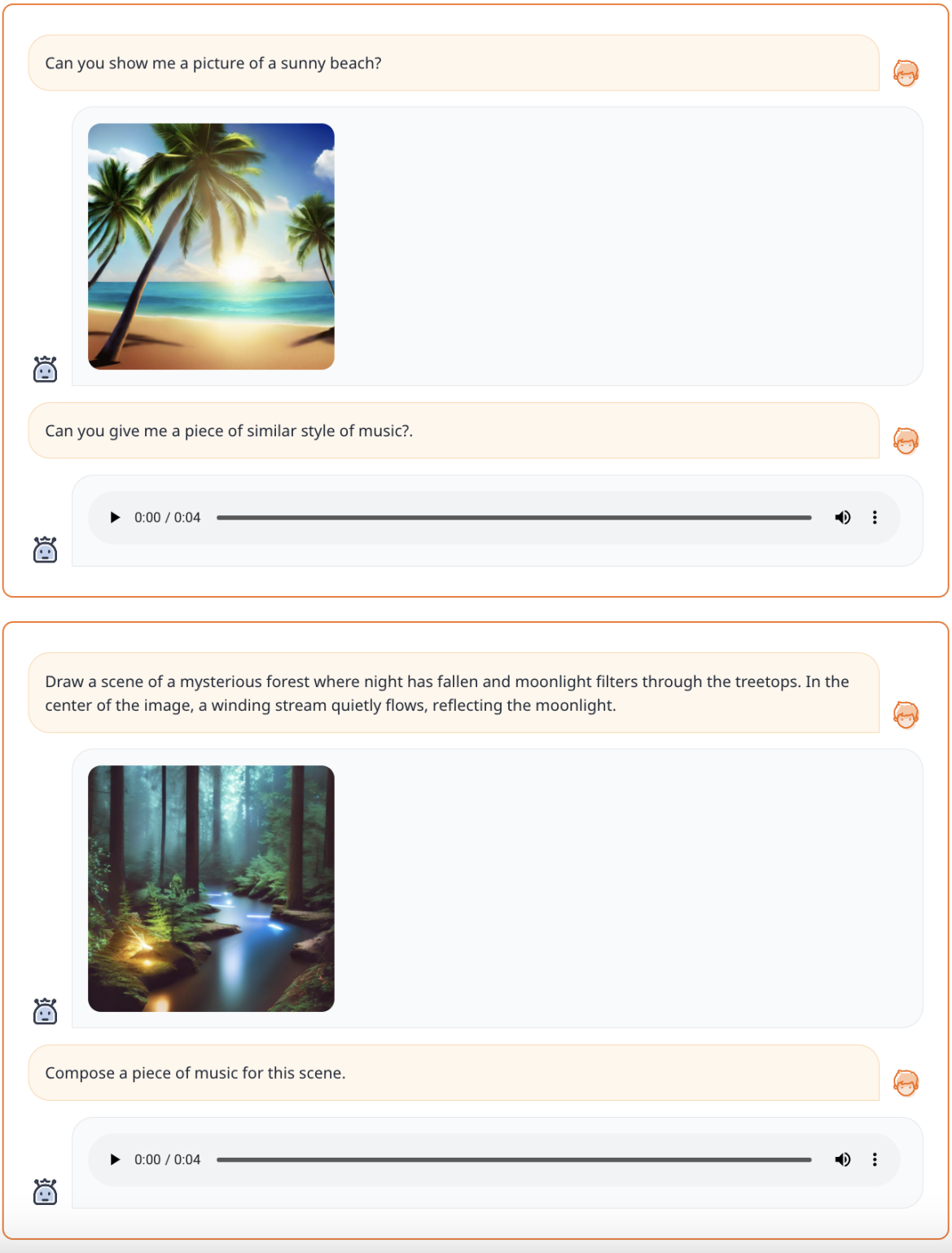


Figure 12: Text → Image + Music

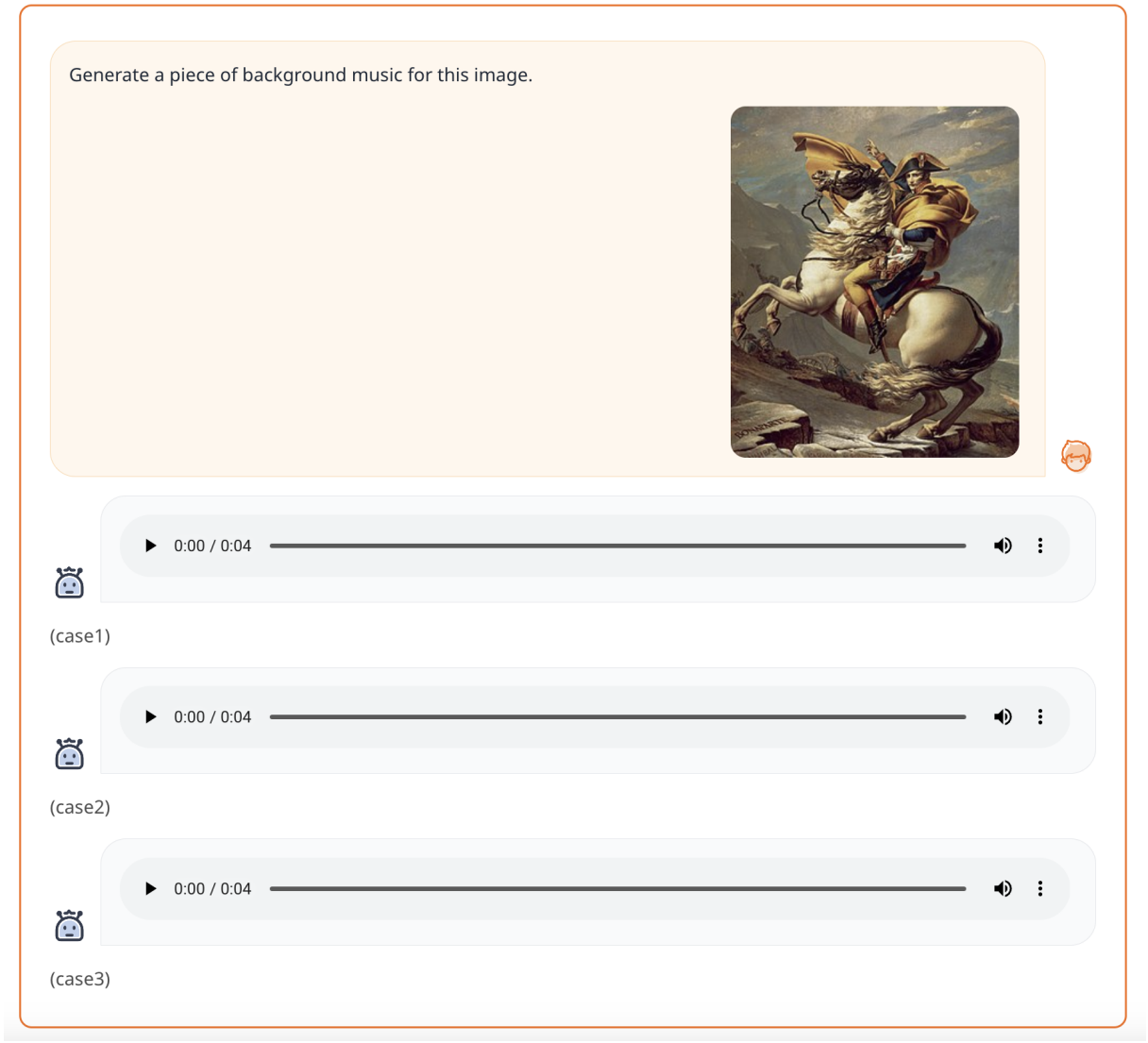


Figure 13: Text + Image → Music

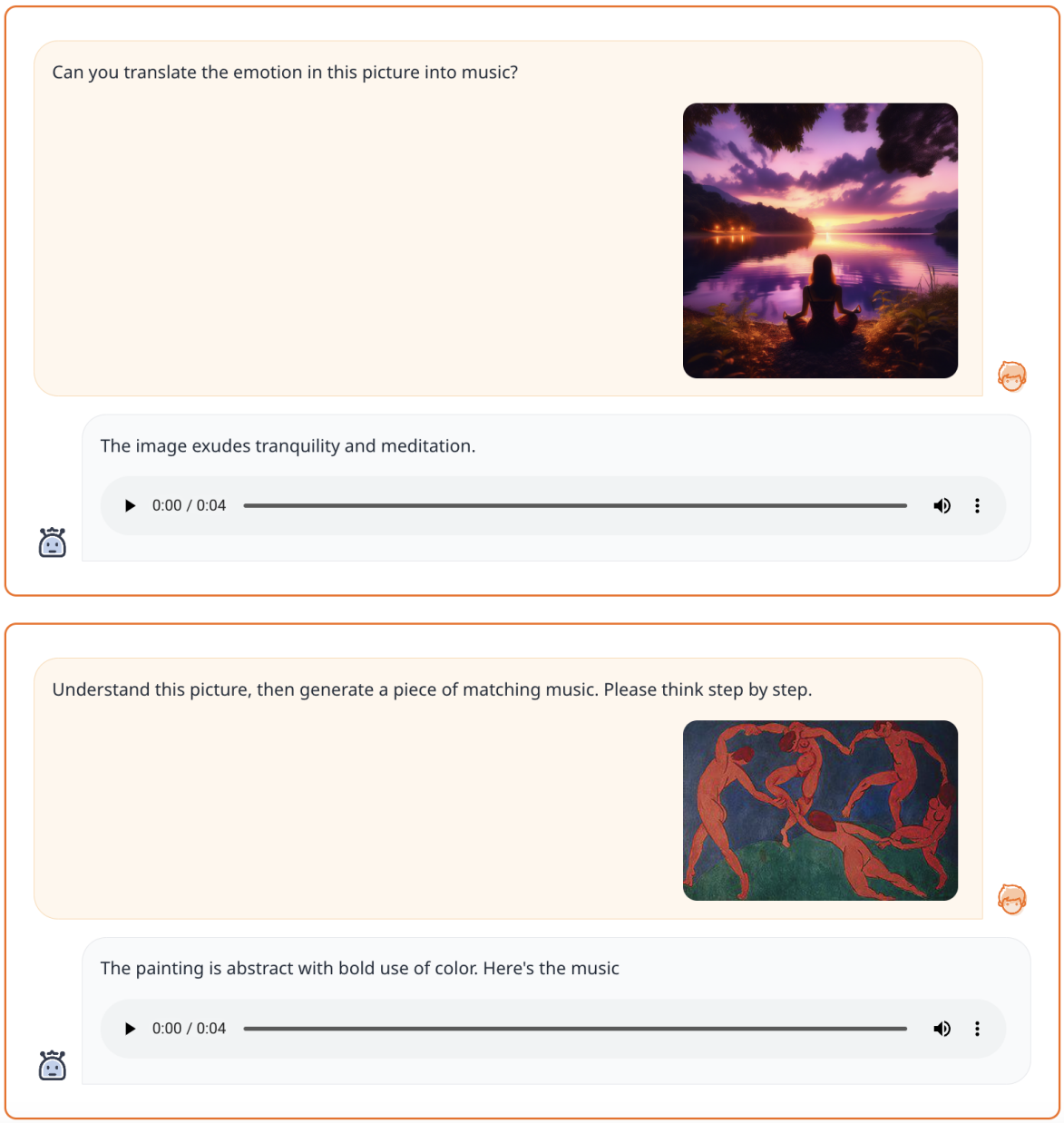
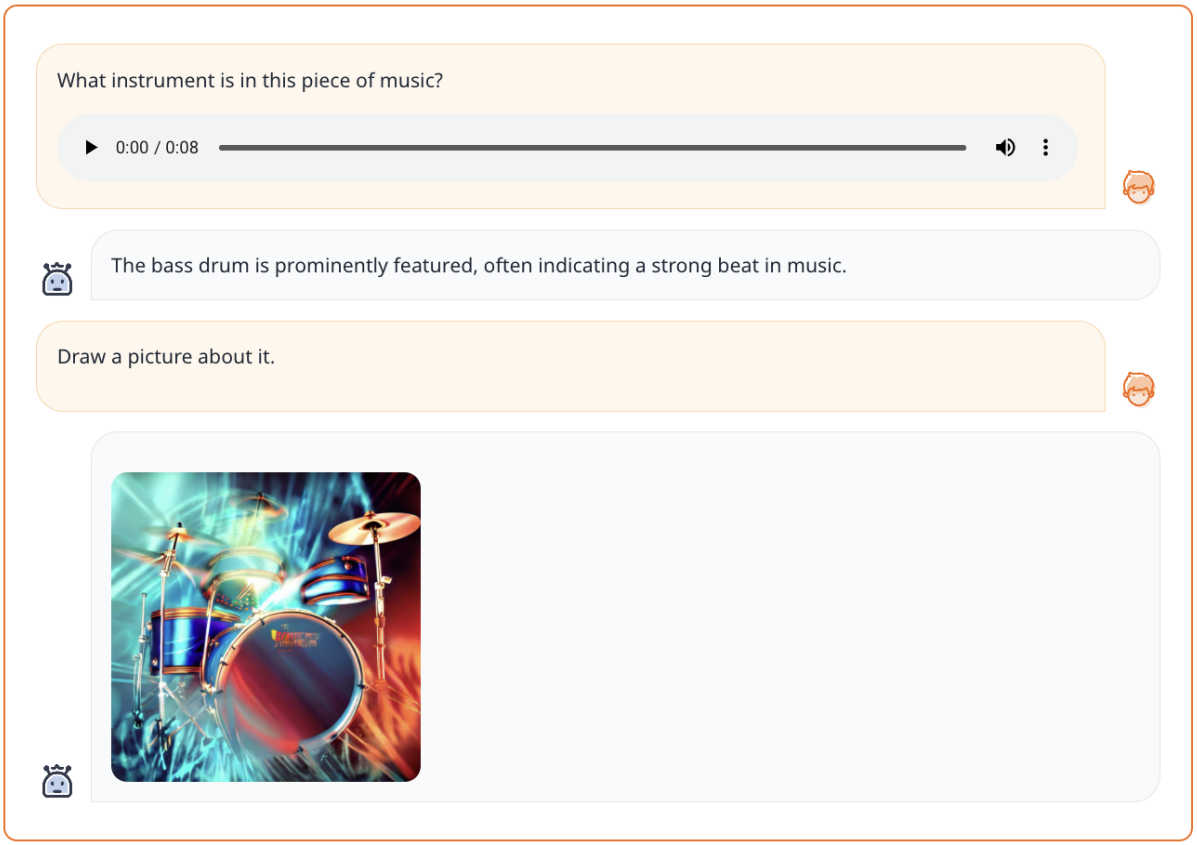


Figure 14: Text + Image → Text + Music



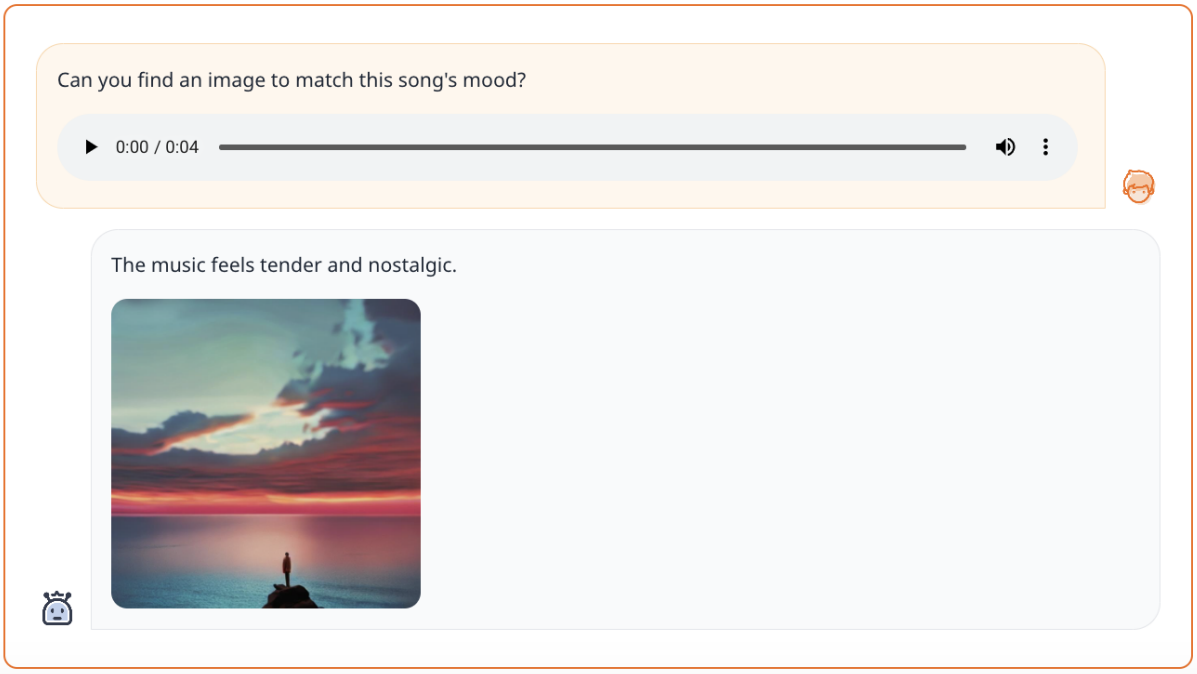


Figure 15: Text + Music → Text + Image

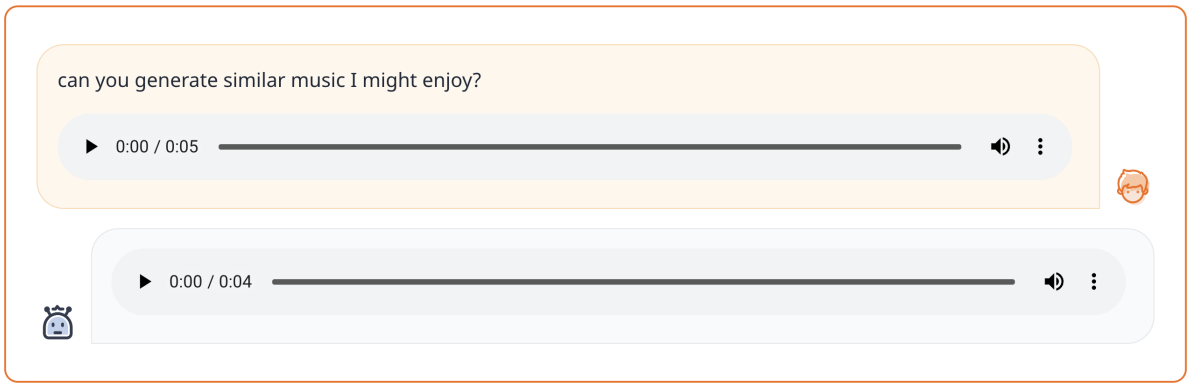


Figure 16: Text + Music → Muisc