Computational Visual Media
https://doi.org/10.26599/s41095-0xx-xxxx-x

Review Article

A Survey on Cross-Modal Interaction Between Music and Multimodal Data

Sifei Li[*1,2], **Mining Tan**[*1,2], **Feier Shen**[2,1], **Minyan Luo**[1,2], **Zijiao Yin**[2], **Fan Tang**[3], **Weiming Dong**[1,2](✉), **and Changsheng Xu**[1,2]

© The Author(s)

arXiv:2504.12796v2 [cs.MM] 21 Feb 2026

Abstract

Multimodal learning has driven innovation across various industries, particularly in the field of music. By deeply exploiting the interaction with visual media and enhancing immersion, it not only lowers the entry barriers to music but also increases the appeal of both music and visual media. This survey aims to provide a comprehensive review of multimodal tasks related to music, outlining how music contributes to multimodal learning and offering insights for researchers seeking to expand the boundaries of computational music. Unlike visual media, which are often semantically and visually intuitive, music primarily interacts with humans through auditory perception, making its data representation inherently less intuitive. Therefore, this paper first introduces common representations of music and provides an overview of music datasets. Subsequently, we categorize cross-modal interactions between music and multimodal data into: music-driven cross-modal interactions, music-oriented cross-modal interactions, and bidirectional music cross-modal interactions. For each category, we systematically trace the development of relevant sub-tasks, analyze existing limitations, and discuss emerging trends. Furthermore, we provide a comprehensive summary of the datasets and evaluation metrics used in multimodal tasks related to music, offering benchmark references for future research. Finally, we discuss the current challenges in cross-modal interactions involving music and propose potential directions for future research.

Keywords AI music, Multimodal learning, Music generation, Music understanding

1 Introduction

As an auditory art form, music not only provides pure auditory pleasure but also unlocks innovative experiences through interactions with other modalities of data. In the fields of film and animation, music enhances visual content to create emotionally impactful scenes. In songs, music combines with language, where vocals and melodies work together to convey emotions and evoke resonance. In dance performances, music guides movement, allowing dancers to convey deep emotional narratives through physical expression. This multimodal fusion expands the expressive power of music and makes it more accessible to a wider audience, supporting its spread and popularity. Within a multimodal interaction framework, music serves as a bridge between different modalities, creating immersive and emotionally engaging experiences. It plays an indispensable role in cultural expression and artistic innovation. Furthermore, this multimodal synergy opens new avenues for artificial intelligence applications in music, driving transformations in music composition, performance, and appreciation, and paving the way for novel forms of musical creativity and interaction.

Artificial intelligence (AI)-driven music and multimodal interaction is an emerging interdisciplinary field focused on integrating music with other data modalities for innovative expression. By leveraging advanced AI technologies, this interaction enables the analysis, generation, and synchronization of music with other modalities, creating richer user experiences. AI-based music and multimodal interaction can be categorized into three types:

(1) *Music-driven cross-modal interaction:* Music serves as input to generate outputs in other modalities (e.g., visuals, text). This type of interaction relies on models with a deep understanding of musical content to generate multimodal output that helps interpret or complement

1 MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. E-mail: S. Li*, lisifei2022@ia.ac.cn; M. Tan*, tanmining2024@ia.ac.cn; M. Luo, luominyan2025@ia.ac.cn; W. Dong, weiming.dong@ia.ac.cn; C. Xu, csxu@nlpr.ia.ac.cn. *Equal contribution.
2 School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190, China. E-mail: F. Shen, shenfeier21@mails.ucas.ac.cn; Z. Jiao, yinzijiao22@mails.ucas.ac.cn.
3 Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China. E-mail: tfan.108@gmail.com.

music, thereby enriching its expression and enhancing user engagement.

(2) *Music-oriented cross-modal interaction:* Non-musical inputs are used to retrieve or generate music, enabling personalized and context-aware composition.

(3) *Bidirectional music cross-modal interaction:* Music functions as both input and output, supporting editing, understanding, and generation across modalities, or even a more adaptive music agent. This enables deep integration and fusion between music and multimodal data, fostering new possibilities for AI-driven music composition, performance, and interactive experience.

Music exhibits a unique level of complexity, presenting significant challenges for AI in music and multimodal data interaction. Unlike visual data, which contains intuitive spatial and color features that are easier to quantify, music involves intricate interactions of melody, rhythm, and harmony with nearly infinite variations. Its emotional and artistic intent is abstract, making it difficult to define with numerical values or text. Unlike text, which has a fixed semantic structure and grammar, music communicates through continuous sequences, where meaning arises from the whole rather than individual notes, often conveying ambiguous emotions. Music is not only a multi-track signal but also typically requires a longer duration to form a complete composition. Furthermore, compared to speech, music demands higher sampling rates and spans longer durations, resulting in significantly increased data sequences. Moreover, music relies on long-term structures (e.g., phrases) rather than short-term linguistic units, which necessitates maintaining long-term structural coherence and managing the complex interplay of polyphonic harmonies. Meanwhile, the subjective nature of musical interpretation makes aligning it with other modalities particularly challenging.

In the early stages of AI applications in music, symbolic music was more extensively studied than music in audio form [1, 2]. However, its reliance on music theory limited access to researchers without a musical background. Early applications were also constrained by computational limitations, focusing primarily on single-modal music tasks and neglecting cross-modal interactions. Recent advances in computational power have enabled greater exploration of audio music and multimodal interaction, though challenges remain. Music can generally be categorized into instrumental music and songs. In instrumental music, the absence of vocal tracks simplifies modeling to some extent. However, high-quality multimodal instrumental music data remains relatively scarce on the internet, imposing additional requirements on data collection. Songs build upon instrumental music by incorporating vocal tracks. Since songs contain linguistic information, modeling songs requires additional attention to speech-related features. Modeling songs as a whole is difficult, so many studies decompose the problem by analyzing vocals and instrumentals separately. However, developing a unified framework for both instrumental and vocal components is efficient and matches user expectations, making it a significant area of research with notable breakthroughs. Despite the availability of multimodal song data online, strict copyright restrictions hinder data construction and open sharing, posing further challenges for this field.

Our survey explores the potential and challenges of cross-modal interactions between music and multimodal data, outlining the progression from music-driven to music-oriented and bidirectional cross-modal interactions. While previous studies have primarily focused on AI applications in music, particularly AI music generation methods [3–7], recent works include reviews of foundation models in music [8], their impact on music understanding [9], and advances in video-to-music generation [10, 11]. In contrast to prior research, this paper adopts a unique perspective by exploring the development of AI in music through the lens of cross-modal interaction between music and multimodal data. We believe that multimodal music serves as a crucial pathway toward broader accessibility to music in the future and as an important means of facilitating cultural dissemination. Currently, this field is experiencing a phase of vigorous growth and holds significant research value. This review aims to clarify the research trajectory of music-multimodal data interactions, highlight key challenges and provide insights into future trends.

The structure of this review is outlined in Figure 1. Sec. 2 introduces music data representations, including waveforms, spectrograms, and symbolic formats, along with related datasets. Sec. 3 discusses music-driven cross-modal interaction tasks and methods, while Sec. 4 focuses on music-oriented cross-modal interaction tasks and methods. Sec. 5 explores bidirectional cross-modal interaction tasks and methods. Sec. 6 summarizes multimodal music datasets, evaluation metrics, and their relevance to specific tasks. Finally, Sec. 7 reviews the current state of cross-modal interactions, highlights existing challenges, and suggests future research directions.

2 Data Representation and Datasets

Musical data representation bridges human music perception with computational models and plays a critical role in how effectively music can be processed, generated, and understood by machines. Humans perceive music through complex

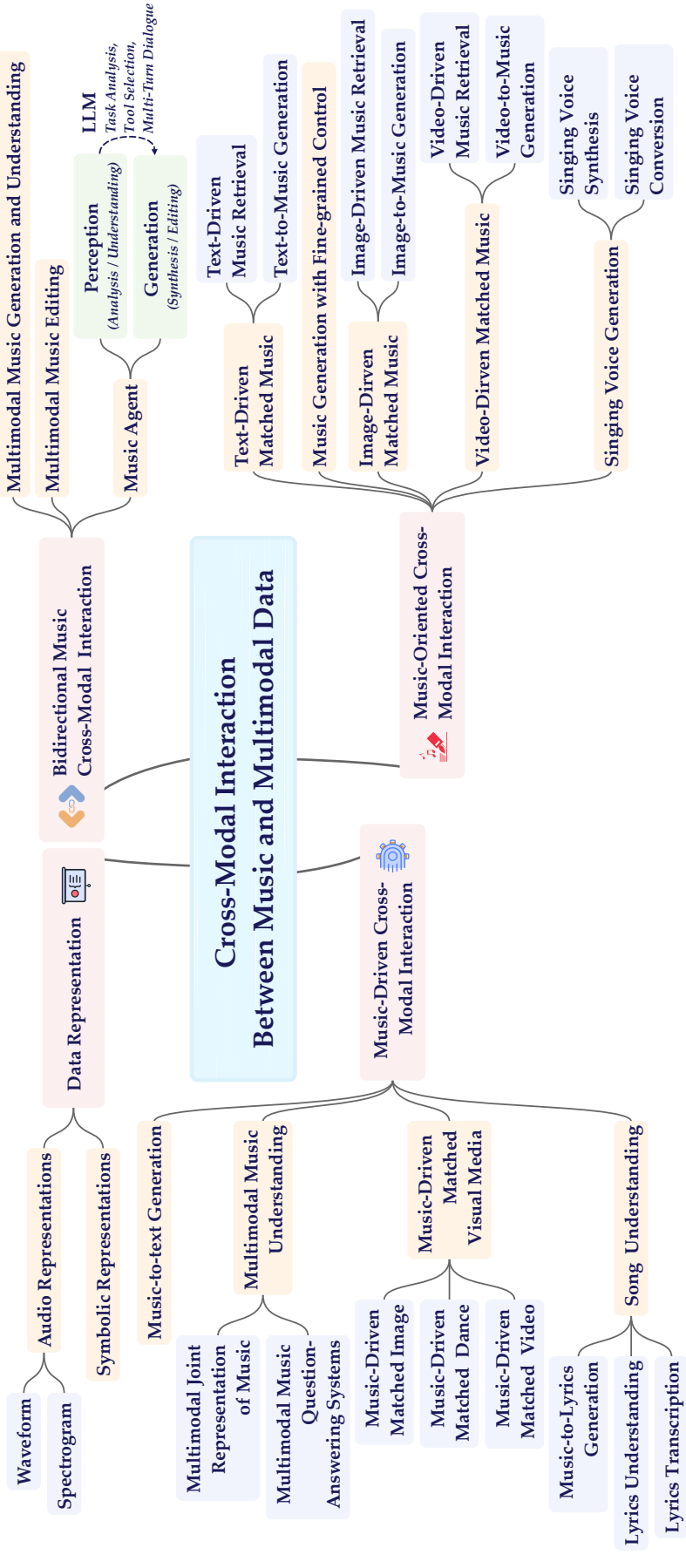
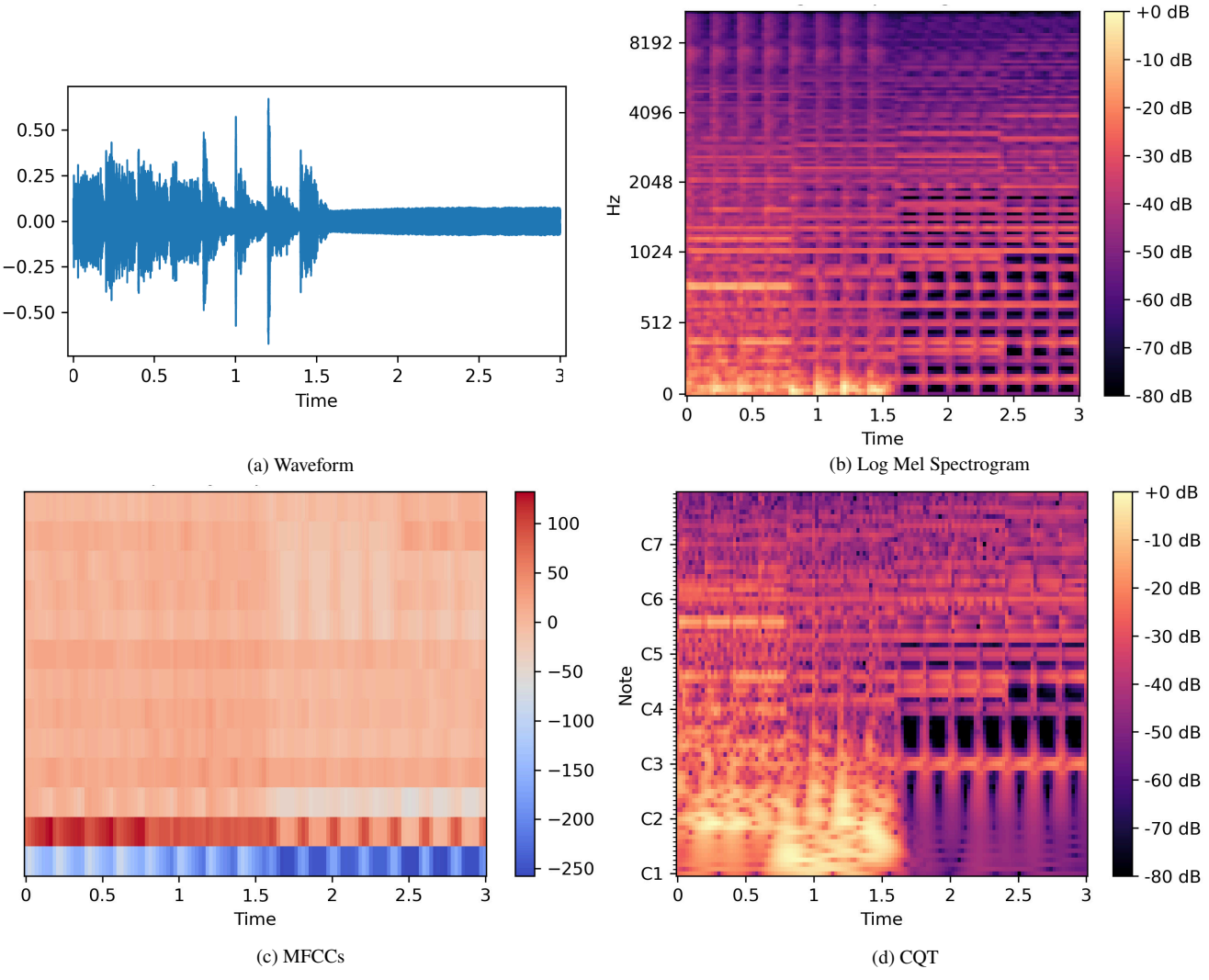


Fig. 1 Overview of cross-modal interaction between music and multimodal data.



(a) Waveform

(b) Log Mel Spectrogram

(c) MFCCs

(d) CQT

Fig. 2 Examples of audio music representations

interactions of perception and cognition, forming features such as pitch, rhythm, dynamics, and timbre [12, 13]. Table 1 lists several datasets commonly used to train models for music-related tasks.

2.1 Audio Representations

In music information processing, audio representation is a continuous form of expression that focuses on capturing detailed acoustic features, such as timbre. Examples of audio representations are shown in Figure 2.

The *waveform* is the most direct representation of an audio signal, depicting its variation over time. The horizontal axis represents time, while the vertical axis indicates amplitude. Since waveforms precisely capture the fine details of sound signals, they serve as a fundamental representation widely used in audio processing and analysis. Many deep learning models [38–40] use raw waveforms as input.

The *log-mel spectrogram* is an audio representation that mimics human hearing. It starts with the short-time Fourier transform (STFT), which splits audio into overlapping frames, applies a window function, and uses FFT for time-frequency analysis. The STFT magnitude is then filtered with mel band-pass filters, designed to reflect how humans perceive sound—linear at low frequencies and logarithmic at high ones. Finally, a logarithmic transformation converts the result to a decibel scale. Although the log-mel spectrogram lowers the STFT's spectral resolution and may lose some detail, it is widely used in audio-related tasks [35, 41, 42] due to its computational efficiency and perceptual relevance.

Using *mel-frequency cepstral coefficients (MFCCs)* is a popular feature extraction and compression method in audio processing. By applying the discrete cosine transform (DCT) to the log-mel spectrum, MFCCs reduce dimensionality and inter-band correlation, retaining the most significant cepstral

Table 1 Overview of single-modal music datasets.

Dataset	Representation	Description	Open
MusicNet [14]	audio	330 classical music recordings with 1M+ annotated note-level labels, including timing, instrument, and rhythmic position of each note.	Yes
Million Song Dataset [15]	audio features	Metadata and audio features of 1M songs provided by Echo Nest.	Yes
POP909 [16]	MIDI	909 piano arrangements of popular songs with separate melody, sub-melody and accompaniment tracks, plus tempo, rhythm, key and chord annotations.	Yes
Free Music Archive [17]	audio	106,574 music tracks organized in a hierarchical taxonomy of 161 genres, featuring full-length high-quality audio, pre-computed features, track- and user-level metadata, tags, and free-form text.	Yes
GTZAN [18]	audio	1,000 audio clips (each 30 seconds long), evenly distributed across 10 music genres with 100 samples per genre.	Yes
Lakh MIDI Dataset [19]	MIDI	176,581 MIDI files, 45,129 of which are matched and aligned with the Million Song Dataset.	Yes
MAESTRO [20]	MIDI, audio	Approximately 200 hours of paired audio and MIDI recordings from ten years of International Piano-e-Competition.	Yes
ASAP [21]	MIDI, MusicXML	222 digital musical scores aligned with 1068 performances (more than 92 hours) of Western classical piano music.	Yes
Opencpop [22]	MIDI, audio	100 unique Mandarin songs with phonetic annotations, including utterance, note, phoneme boundaries, and pitch types, totaling 3,756 utterances lasting about 5.2 hours.	Yes
NUS-48E Corpus [23]	audio	48 English pop songs by 12 singers with phoneme annotations.	Yes
IrishMAN [24]	ABC, MIDI, MusicXML	216,284 Irish tunes with automatic annotations using control codes derived from ABC symbols.	Yes
Nottingham Music Dataset	ABC notation	1,200 British and American folk songs.	Yes
Henrik Norbeck's ABC Tunes	ABC notation	3000+ Irish and Swedish folk songs.	Yes
ADL piano dataset [25]	MIDI	11,086 piano pieces extracted from the Lakh MIDI dataset and publicly available sources on the internet.	Yes
Symphony [26]	MIDI	46,359 files with multiple instruments and tracks, mainly orchestral.	Yes
Symbolic Symphony Set	MusicXML	High-quality collection of four complete symphonies (16 movements), containing 285,387 notes with detailed annotations including note events, temporal attributes, musical form, and orchestral texture.	Yes
NES-MDB [27]	MIDI	5,278 NES game background music tracks, featuring four programmable 8-bit register-controlled instrument voices.	Yes
GiantMIDI-Piano [28]	MIDI, audio	10,855 transcribed classical piano works (1,237 hours).	Yes
MetaMIDI Dataset [29]	MIDI, audio	436,631 MIDI files with metadata, matched with 10M+ Spotify music tracks.	Yes
MTG-Jamendo [30]	audio	55,000+ full audio tracks with 195 tags for genre, instrument, and mood/theme categories.	Yes
Music4ALL [31]	audio	109,269 tracks (911 hours) with metadata, tags, genre information, lyrics, etc.	Yes
AudioSet	URL of audio	A large-scale collection of human-labeled 10-second sound clips drawn from YouTube videos, including 1,011,305 music clips, 117,343 musical instrument clips, etc.	Yes
AcousticBrainz	audio features	2,524,739 audio feature files with MusicBrainz metadata.	Yes
Disco-10M [32]	features, URL of audio	15296232 audio feature files covering various genres and artists.	No
Chordified JSB Chorales Dataset [33]	MusicXML	366 JSB chorales preprocessed using automated 'chordification' and labelling techniques.	Yes
GTTM Database	MusicXML	300 classical music melodies with hierarchical music structure analysis.	Yes
EMOPIA [34]	audio, MIDI	1,087 music clips from 387 songs and clip-level emotion labels annotated by four dedicated annotators.	Yes
PopCS [35]	audio	117 Chinese pop songs (total 5.89 hours) with lyrics.	Yes
OpenSinger [36]	audio	50 hours of Mandarin singing voice recorded by 66 singers.	Yes
M4Singer [37]	audio, MIDI	29.77 hours of Mandarin singing voice recorded by 20 professional singers, covering all four singing styles, i.e., soprano, alto, tenor, and bass.	Yes

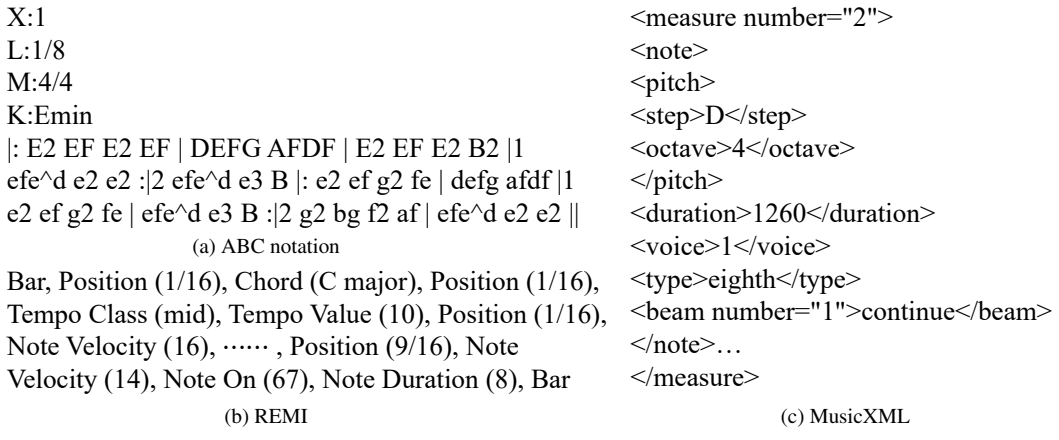


Fig. 3 Examples of symbolic music representations

coefficients. The first 12-13 coefficients typically capture the key spectral information. MFCCs are commonly used in speech processing [43, 44], and can also be applied in music classification [45] and music information retrieval [46].

The *constant-Q transform (CQT)* refers to a filter bank in which center frequencies are exponentially spaced, the filter bandwidths vary, but the ratio of center frequency to bandwidth remains constant (Q). Unlike the Fourier transform, which uses a linear frequency axis, the CQT employs a logarithmic frequency scale that naturally follows the musical scale. It applies longer time windows at lower frequencies for better frequency resolution and shorter windows at higher frequencies for improved time resolution, optimizing musical note frequency extraction and analysis. Although the CQT lags behind the mel spectrogram in terms of perception and reconstruction quality due to its lower frequency scaling, it has proven useful for tasks such as music representation learning [47] and musical style transfer [48, 49].

2.2 Symbolic Representations

Symbolic representation encodes the logical structure of music using discrete symbols like notes, chords, and rhythms, as shown in Figure 3.

MIDI (musical instrument digital interface) is a widely used protocol in electronic music to transmit performance data between instruments, software, and devices. Instead of encoding traditional musical notation, MIDI captures performance details such as pitch, duration, velocity, and dynamics, making it ideal for simulating human performances, though less effective at conveying musical structure. MIDI represents music as a sequence of events, including *note-on, note-off, time shift, and velocity*. Its ability to capture subtle performance nuances makes it valuable for applications like music generation [50], transcription [51], and style transfer [52].

REMI (revamped midi-derived events) [53] is an enhanced representation of MIDI data designed to address the limitations of traditional MIDI in expressing musical rhythm and structure. It introduces the *note duration* event to replace the traditional *note off* event, directly representing the duration of notes and enhancing the capability of rhythm modeling. Additionally, REMI employs a combination of *bar* and *position* tokens to replace the *time shift* event, where *bar* indicates the beginning of a bar and *position* specifies certain positions in a bar. This design establishes a clear metric grid, facilitating more accurate modeling of musical structure. To further enhance the expressiveness of musical rhythm, REMI adds *tempo* events placed before corresponding *position* events. This approach is widely used in music generation tasks [50].

MusicXML is a notation format based on extensible markup language (XML), designed to represent various symbols used in Western music—like rests, slurs, beams, barlines, key and time signatures, articulations, and ornaments—that are often missing in MIDI. It is mainly used to store lead sheets, which include the basic parts of a song, including melodies, lyrics, chords, and repetition marks. The chords can include both literal chord symbols (e.g. Gmaj7) and functional notation relative to the key (e.g. VI7). MusicXML uses a hierarchical structure with nested tags to organize musical data. For instance, a `<measure>` contains `<note>` tags, and each `<note>` includes `<pitch>` and `<duration>`. This format helps computers understand how musical elements are related. However, its detailed structure and XML's verbosity make it challenging to use for training music generation models, especially since it lacks direct audio data and is complex to encode and decode.

The *piano roll* is a common symbolic way to represent music, originating from player pianos that used perforated paper rolls to play songs automatically. In modern digital applications, the piano roll appears as a two-dimensional grid:

time runs along the horizontal axis, and pitch runs along the vertical axis. Notes are shown as rectangles whose lengths indicate duration, making this format an intuitive tool for music analysis and generation. Mathematically, the piano roll is structured as an $h \times w$ real-valued matrix, where h denotes the pitch range, including special entries for rests and note continuations, and w represents the number of time steps, set by the beat resolution. Traditional piano rolls use binary values to show whether a note is played, but modern versions can include extra details like velocity, articulation, and dynamics [54]. This image-like format has proven useful for tasks such as music generation [55, 56] and representation learning [57].

ABC notation is a concise, computer-friendly way of representing music using the ASCII character set. It uses the letters A–G and a–g for notes, while 'z' represents a rest. The notation is structured into two main components: a header and a body. The header contains metadata like title, time signature, key signatures, and default note length. In the body, each note is represented as a token encoding pitch, duration, and rhythmic position, separated by bar lines. ABC notation's compatibility with natural language representations has made it a valuable resource for text-to-symbolic music models, facilitating the generation of music [58–60].

Note graphs provide a graph-based model of musical scores that capture the complex relationships in music using structured graph elements. In the method proposed by Jeong et al. [61], each note is a node, and edges represent musical connections such as sequence, rest, onset, sustain, voice, and slur. These edges are directed both forward and backward, with self-loops for each note, resulting in twelve edge types. This structure offers a detailed way to represent musical scores. Later works have built on this idea. Karystinaios et al. [62] included both notes and rests as nodes, linking them through temporal relationships like synchronization, succession, and overlap. In a follow-up study [63], they proposed a heterogeneous graph framework for voice separation using link prediction. Similarly, Zhang et al. [64] introduced a new graph representation for musical performances and compared it to matrix and sequence-based methods in piece-level classification tasks. These studies show how graph-based models can offer richer insights than traditional music representations.

3 Music-Driven Cross-Modal Interaction

3.1 Music-to-Text Generation

Music-to-text generation lies at the intersection of music information retrieval and natural language processing (NLP), aiming to generate rich textual representations for music. This task can be broadly categorized into three subdomains: *music understanding*, *song understanding*, and *lyrics generation*.

3.1.1 Music Understanding

Music understanding aims to extract semantic information from music and represent it in natural language. A fundamental task in this field is music captioning, which generates coherent textual descriptions of musical segments, capturing genre, emotion, and style to enhance music retrieval and recommendation. Music captioning models typically follow an encoder-decoder framework, where the encoder extracts latent features from the raw audio, and the decoder generates the textual description based on these features (see Figure 4).

Choi et al. [65] first explored this task, proposing a sequence-to-sequence model that encoded musical features into vectors and decoded them into descriptive text. Building on this work, Manco et al. [66] proposed the MusCaps model, which employed an encoder-decoder architecture integrating convolutional neural networks (CNNs) and recurrent neural networks (RNNs), improving audio-text alignment and capturing musical nuances. Xu et al. [67] enhanced the model with reinforcement learning, improving caption quality for complex audio.

Despite these advances, traditional CNN-RNN architectures face limitations in temporal modeling and long-range dependency capture. To address these shortcomings, Chen et al. [68] introduced a Transformer-based model, utilizing a CNN10 encoder to extract spectral features and a Transformer decoder to generate natural text. Mei et al. [69] took this further with the Audio Captioning Transformer (ACT), a fully Transformer-based model using self-attention for audio processing and global context modeling.

Recently, large language models (LLMs) have driven advances in music understanding frameworks. MusiLingo [70] and MU-LLaMA [71] integrated the audio model MERT and LLaMA [72], improving efficiency, accuracy, and generalization in music understanding tasks. Zhao et al. [73] proposed OpenMU, a multimodal large language model for music understanding and OpenMU-Bench, a large-scale benchmark for comprehensive evaluation. These models demonstrated excellent performance in tasks like music captioning, reasoning, and question answering.

Data scarcity remains a critical challenge in this domain. Doh et al. [74] proposed LP-MusicCaps, using an LLM to generate pseudo-music captions from extensive tagged datasets, supporting captioning model training. To improve caption diversity, Augment, Drop & Swap [75] optimizes the learning of music-text representations through targeted

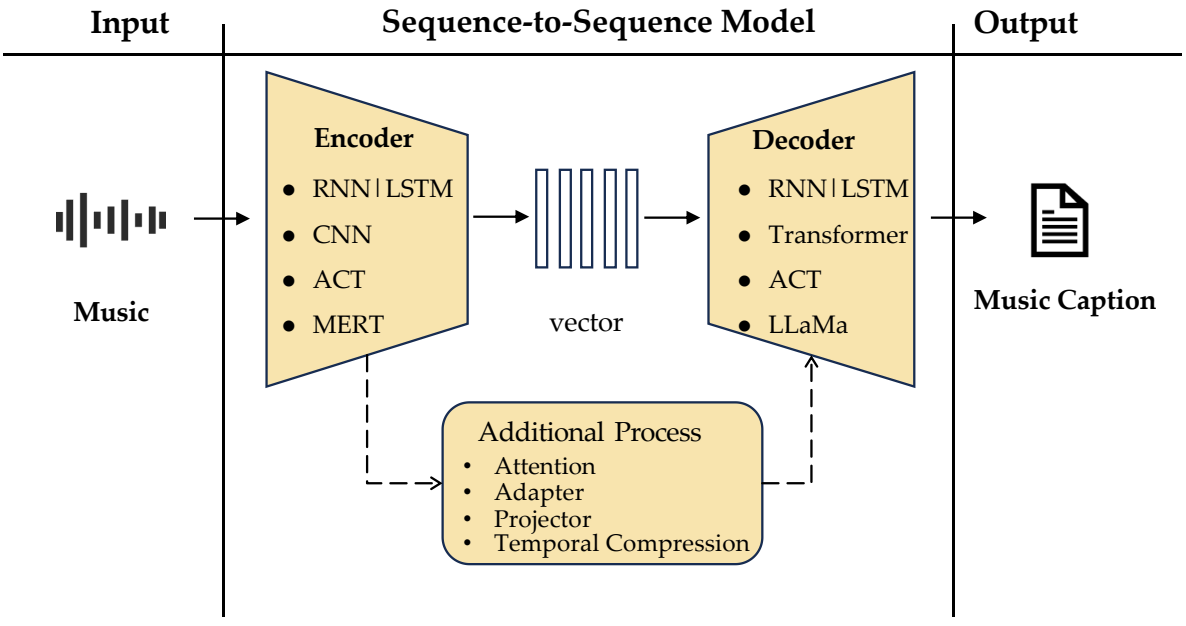


Fig. 4 Overview of music caption methods.

augmentation, deletion, and substitution operations, enhancing lexical variety and semantic expressiveness. Srivatsan et al. [76] incorporated multimodal inputs, such as video frames or musical themes, to construct context-specific prefixes, enabling captions that reflect the music's background and emotion. Similarly, RECAP [77] retrieved similar music captions to construct a prompt, enhancing adaptability and precision for complex music.

Despite significant progress in music understanding, challenges such as data scarcity, coarse and simplistic music descriptions, and limited generalization in complex scenarios hinder its development. In the future, compatibility with diverse music representations, flexible question-answering with reasoning capabilities, and support for cross-cultural music are expected to become key trends.

3.1.2 Song Understanding

Song understanding focuses on the tasks of automatic lyrics transcription and alignment, aiming to extract lyrics from audio signals and align them precisely with the corresponding melody.

Early approaches predominantly relied on traditional audio feature analysis and handcrafted algorithms, such as hidden Markov models (HMMs) [78, 79], Gaussian mixture models (GMMs) [80], and probabilistic models [81]. Although these methods succeed in constrained scenarios, manually designed features and limited model expressiveness make them unsuitable for complex audio.

The advent of deep learning revolutionized lyrics transcription and alignment, with CNNs [82] and RNNs [83] demonstrating particular efficacy in handling complex musical patterns and temporal dependencies. Gupta et al. [84] trained polyphonic acoustic models using genre-specific characteristics, improving lyrics transcription and alignment. An end-to-end model based on Wave-U-Net [85], utilizing connectionist temporal classification (CTC) loss, substantially enhanced alignment accuracy in multi-instrumental music. MSTRE-Net [86] introduced a novel multi-streaming time-delay neural network (MTDNN) architecture, optimizing both speed and accuracy. Gao et al. [87] developed a multi-task learning network for lyrics transcription in polyphonic music, jointly training lyrics and chord transcription, improving transcription accuracy.

With the advent of multimodal large models, lyrics transcription and alignment progressed toward greater complexity and precision. OpenAI's Whisper model, trained with weakly supervised learning, excels in long-form audio transcription. LyricWhiz [88] integrated Whisper with GPT-4 for zero-shot multilingual lyrics transcription, using Whisper for audio transcription and GPT-4 for text refinement. SongComposer [89] employed a two-stage approach—pre-training and supervised fine-tuning—to generate aligned lyrics-melody pairs. Addressing multilingual automatic lyrics transcription, Huang et al. [90] proposed an end-to-end system that expanded vocabularies using existing datasets and integrated linguistic information, demonstrating the superiority of multilingual models over monolingual ones. SongTrans [91] introduced a unified model for simultaneous lyrics and note transcription

and alignment without preprocessing. Gu et al. [92] developed a multimodal framework combining audio and lip movement data to enhance lyrics and note transcription. In addition, they introduced the first multimodal singing datasets, N20EMv1 and N20EMv2.

Additionally, several studies [93–95] have explored automatic lyrics interpretation, generating accessible paraphrases to deepen song comprehension. While these methods do not directly address melody, they offer fresh perspectives on musical semantics. Multimodal approaches integrating lyrics and audio can yield more accurate interpretations. For instance, Bart-fusion [96] used cross-modal attention to fuse audio and lyrics, producing more precise explanations.

Song understanding tasks currently face challenges in robust multilingual transcription and accurate alignment in complex audio. Future trends include unified frameworks based on multimodal pretrained models, cross-lingual alignment, and real-time transcription, to improve lyrics transcription and semantic understanding.

3.1.3 Lyrics Generation

Lyrics generation aims to produce rhythmically aligned and semantically coherent lyrics based on melodies or other inputs. This task has significant potential in music composition, entertainment, and personalized content creation, such as generating lyrics for melodies or assisting songwriters.

Early-stage approaches [97, 98] primarily relied on rule-based and statistical models, utilizing only lyrics, which results in limited generation quality.

To improve alignment between lyrics and melodies, the melody-to-lyrics (M2L) approach has been widely explored. Watanabe et al. [99] constructed a melody-lyrics alignment dataset and proposed an RNN-based melody-conditioned lyrics language model to generate lyrics compatible with melodies. Chen et al. [100] developed an end-to-end system based on SeqGAN, generating lyrics from melody inputs while supporting thematic conditions to improve semantic expression. SongMASS [101] addressed data scarcity and alignment constraints in lyrics-to-melody and melody-to-lyrics generation through masked sequence-to-sequence pre-training and attention-based alignment mechanisms. AI-Lyricist [102] targeted multi-channel music and lexical constraints, producing lyrics suitable for language learning.

Subsequent studies have further advanced this field. LOAF-M2L [103] jointly learned wording and formatting, improving syllable compatibility between lyrics and melodies. Tian et al. [104] proposed an unsupervised hierarchical model that separated training (based on plain text) from inference (melody-guided text generation), achieving high-quality lyrics generation in unsupervised settings. UniLG [105] integrated text and music information through composite templates, effectively capturing the structural information of lyrics and generating content aligned with song structures (e.g. chorus and verse). Zhang et al. [106] introduced a note-level melody-based lyrics generation method, utilizing a Transformer architecture to capture fine-grained syllable-level correlations between melody and lyrics data, facilitating controllable and diverse lyrics generation. Calzolari et al. [107] proposed a lyrics generation approach based on scansion analysis, employing the mBart model to treat lyrics generation as a translation task, producing lyrics that matched both melody and tonal patterns. XAI-Lyricist [108], built on Transformer, provided explanations for prosodic patterns, addressing challenges in evaluating ability to be sung and alignment. Chae et al. [109] introduced a song-form-aware framework for full-song text-to-lyrics generation, enabling multi-granularity syllable count control. Beyond M2L lyrics generation, auxiliary lyric creation systems [110, 111] have also matured, assisting creators in better controlling lyric content, rhyme schemes, and formatting.

Lyrics generation tasks face several challenges, including poor cross-lingual prosody adaptation and limited ability to be sung of generated lyrics. Future works are expected to include multimodal pretrained models for lyrics generation, cross-lingual prosody modeling, and interactive user-driven creation systems, aiming to enhance the alignment between lyrics and melody.

3.2 Multimodal Music Understanding

Multimodal music understanding differs significantly from unimodal music understanding. Unimodal music understanding primarily focuses on music information retrieval (MIR) tasks, such as chord detection, key recognition, melody estimation, pitch detection, automatic music transcription, instrument classification, and music categorization. These tasks only take music as input and analyze specific properties of the music. In contrast, multimodal music understanding aims to enhance the comprehension of music by integrating multiple modalities. It not only enables a deeper analysis of the intrinsic features of music but also achieves an understanding of the relationships between music and other modalities.

3.2.1 Multimodal Representation Learning with Music

Multimodal music representation encodes various modalities into a shared feature space, critical for understanding music and evaluating cross-modal alignment. Among these, music-text joint representation is the most extensively studied. Huang et al. [112] leveraged textual tags and co-listening

statistics from large-scale music video data, aligning text and music embeddings through triplet loss and cross-entropy loss. However, it still failed to understand unstructured text. Won et al. [113] explored classification, regression, and metric learning to align free-form text with music. But they found that relying solely on existing emotion labels was insufficient for modality alignment. MuLaP [114] employed weak language supervision by using masked language modeling, masked audio modeling, and audio-text matching as pretraining objectives for multimodal pretraining of audio and language data. However, its cross-attention-based modality fusion limited transfer learning applications, making it less flexible than shared embedding spaces. MuLan [115], inspired by CLIP, constructed a joint embedding space for music and text using large numbers of music- text pairs. It exhibited strong representation capabilities and performed well in zero-shot downstream tasks, but the model remained closed-source. Unlike the above approaches, which focused on audio representations, CLaMP2 [116] separately designed a multilingual text encoder and a multimodal music encoder. Using contrastive learning, it jointly represented text in 101 languages with ABC notation and MIDI multimodal music data, demonstrating effective applications for traditional MIR tasks.

Many audio-text joint embedding models also possess some degree of music-text joint representation capability, as their datasets contain music. Favory et al. [117, 118] employed contrastive learning to model the latent representations of audio and labels. However, their approach did not generalize to free-form natural language. A major limitation of audio-text joint representation is the lack of paired audio-text data. CLAP [119] addressed this by collecting the LAION-Audio-630K dataset and applying contrastive learning to obtain a stronger audio-text joint embedding space, which efficiently handled audio of varying lengths and was widely used in downstream tasks and evaluation. Zhu et al. [120] constructed an even larger dataset and leveraged large language models (LLMs) to improve the quality of textual descriptions. They employed masked autoencoders (MAE) as audio encoders, achieving superior performance in downstream tasks. Wu et al. [121] revealed limitations in existing audio–text models, especially in capturing the sequential and concurrent relationships of sound events. They proposed a Transformer-based architecture to improve temporal and contextual understanding.

The above works mainly focused on the semantic alignment of audio and text. Some studies have focused on audio-video representation learning. CAV-MAE and MAViL [122] used MAE and contrastive learning to achieve audio-video joint representation. EquiAV [123] applied equivariant learning to audiovisual contrastive learning, effectively improving representation learning performance through a unique design. Diff-Feloy [124] introduced CAVP, a model that enhanced semantic and temporal alignment in audio-video contrastive learning.

While the above studies focused on aligning two modalities, some research has explored joint representation across all modalities. Several approaches extended existing models to make them adaptable to other modalities. For example, AudioCLIP [125] and Wav2CLIP [126] integrated audio encoders into CLIP, enabling joint representation of text, images, and audio. ImageBind [127] learned six-modality joint embeddings by leveraging image-paired data and pretrained models and using image embeddings as a bridge. VIT-LENS [128] extended pretrained ViT models to multiple modalities and aligned their output features with anchor data features. Another approach involved constructing multimodal datasets. Chen et al. [129] created the VAST-27M dataset, a large multimodal video dataset, and trained a highly generalizable multimodal joint representation model, VAST. Unlike others, CLAMP 3 [130] paid more attention to music; it is a framework for MIR across multiple modalities and languages.

3.2.2 Multimodal Music Question-Answering Systems

With the emergence of LLMs, music understanding tasks have moved beyond structured text or predefined labels, enabling free-form textual outputs. Mu-LLaMa created question-answer pairs from existing audio description datasets, constructed the MusicQA dataset, and designed a music understanding adapter module to integrate music representations with LLMs, achieving efficient music question-answering and description generation. Building on this, MusiLingo further explored music-language model alignment by constructing the MusicInstruct dataset based on MusicCaps. It used a single projection layer to align music models with language models, achieving better performance through pretraining and instruction fine-tuning. LLARK [131] integrated existing open-source datasets and employed a generative music encoder with a simple multimodal projection module, enabling music understanding, captioning, and reasoning.

To evaluate these advances, MuChoMusic [132] introduced the first benchmark dataset for assessing audio-language models' music understanding capabilities, containing 1,187 multiple-choice questions. Compared to previous works that relied on synthetic data, this benchmark offered greater reliability. Zhao et al. [73] proposed OpenMU-Bench, which expanded existing datasets to millions of training samples

and introduced two additional subtasks: lyrics understanding and music tool usage. MuMu-LLaMA [133] extended multimodal musicQA to the visual domain, enabling the model to understand relationships between music, images, and video, supporting both music understanding and music generation. Recently, several more general multimodal audio question-answering systems [134–136] have demonstrated reliable music understanding capabilities.

Multimodal music understanding is vital for deep music analysis, supporting the evaluation of music generation models and future systems with reasoning abilities. Future work needs to further explore the intrinsic correspondence between multimodal data and music, enhance the flexibility of model input (e.g. in terms of data formats, question-answering modes), and develop chain-of-thought (CoT) data to enhance music understanding.

3.3 Music-Driven Matched Visual Media

Collaborative creation between music and visual content (images, videos) is a vital area in multimodal research. Next we explore how music drives generation, alignment, and retrieval of visual content. These technologies unlock new possibilities for applications such as automatic music visualization, music video (MV) creation, and dance generation.

3.3.1 Music-Driven Matched Image

Music-guided image generation has typically relied on other modalities as intermediate steps. Specifically, music is first processed by LP-Music-Caps to generate audio captions, which are then converted into image descriptions by LLaMA2 [137], and images are generated using Stable Diffusion XL [138].

Music-driven image generation is challenging due to the stark contrast between music and images. Music is sequential, with dimensions like melody, rhythm, and harmony, while images are spatial, made up of pixels and visual elements. This gap complicates data acquisition and evaluation, and makes achieving real-world applications difficult.

3.3.2 Music-Driven Matched Dance

Dance-related research occupies a significant portion of music-driven visual media studies and is therefore discussed separately. This section is divided into two parts: music-driven dance alignment and music-driven dance generation, as shown in Figure 5.

Music-driven dance alignment refers to the adjustment of the relative timing between music and dance clips to create a rhythm-consistent dance video. Davis et al. [139] introduced the concept of visual beats based on handcrafted features, and aligned them to musical beats by time-warping. AlignNet [140] used a spatiotemporal attention module to extract music and video feature pyramids, and achieved dance-music alignment via frame-level correspondence estimation based on feature similarity metrics. MuDaR [141] utilized an attention-based fully connected layer to extract motion rhythm and facilitate rhythm synchronization using both explicit and implicit methods. M2BNet [142], comprising MFE blocks with SGCN and TCN, effectively extracted intricate motion rhythms from videos by analyzing human skeletons and incorporating temporal information, thereby facilitating beat-level alignment.

Music-driven dance generation (also known as AI choreography) takes a music sequence as input and outputs a 3D human pose sequence, where human poses are typically represented by the rotation matrices of 24 joints and the root translation, in SMPL format. Early works explored various methods, based on CNNs [143, 144], gated recurrent units (GRUs) [145], long short-term memory (LSTM) [146–149], generative adversarial networks (GANs) [150, 151], and Transformers [152–154]. Li et al. [152] proposed a full-attention cross-modal transformer network that encoded seed motion and music features to generate new dance sequences autoregressively. DanceFormer [154] employed a two-stage approach, generating key poses for each beat and then predicting intermediate motion curves. Integrating professional domain knowledge, ChoreoMaster [155] incorporated choreography rules into a graph-based framework, leveraging cross-modal embeddings to optimize motion paths for highly controllable dance synthesis. Bailando employed VQ-VAE to encode dance sequences into hierarchical codebooks and designed an actor-critic GPT to generate motions by composing codebook units. To ensure rhythmic synchrony, it optimized a composite reward function: $R = \lambda_1 R_{\text{beat}} + \lambda_2 R_{\text{div}}$, where R_{beat} aligned motion kinetic peaks with musical beats and R_{div} promoted choreographic diversity. Based on Bailando, GTN-Bailando [156] employed a genre token network (GTN) to identify music genres, guiding dance generation with genre consistency. To address the limited generalization ability of GTN-Bailando, EnchantDance [157] incorporated a pre-trained audio spectrogram transformer (AST) to robustly predict music genres via transfer learning. ExpressiveBailando [158] incorporated frequency information into VQ-VAE and leveraged MERT to add music genre and beat information. Furthermore, TM2D [159] built up a shared codebook using both music-to-dance and text-to-motion datasets, enabling dance generation conditioned on both text and music.

The aforementioned methods often relied on complex

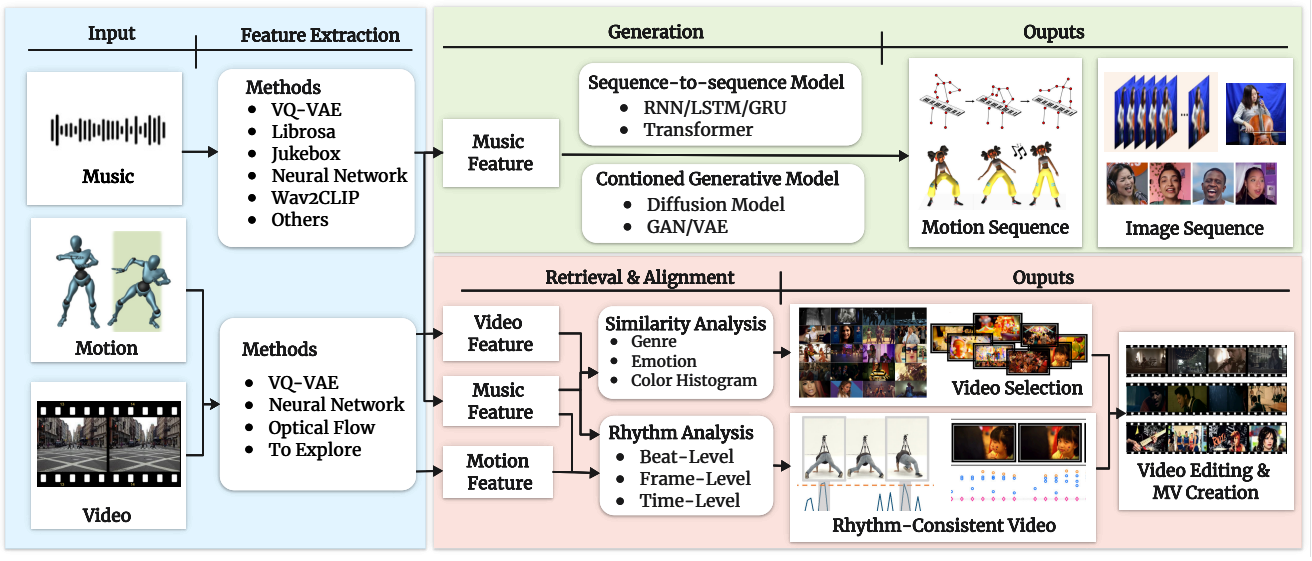


Fig. 5 Overview of the music-driven matched dance/video methods.

models and produced low-diversity dance sequences. Recent advances in diffusion models have driven progress [160, 161] in music-to-dance generation. EDGE [161] leveraged a transformer-based diffusion model and the advanced music feature extractor Jukebox [162] to produce fluid and natural dance movements. EDGE greatly influenced subsequent dance generation methods. To ensure that the generated dance was consistent with the music style, DGSDP [163] incorporated music style prompts into the diffusion model. To address the issue of motion coherence, LongDanceDiff [164] incorporated past movements into the denoising process; BAMD [165], based on a bidirectional autoregressive diffusion framework, considered both preceding and following frames when generating current frame dance movements. Focusing on the guiding role of lyrics in dance generation, Yin et al. [166] proposed LM2D, which employed a multimodal diffusion model conditioned on both music and lyrics with consistency distillation. For long dance generation, Lodge [167] adopted a two-stage coarse-to-fine diffusion framework, using dance primitives as intermediate representations to generate long dances in parallel.

Most of the above methods focused on improving the dance generation network while neglecting the importance of music feature processing. They typically used a frozen Jukebox model or the audio processing toolkit Librosa to extract music features, failing to fully and effectively utilize musical information. As a result, the generated dances often suffer from an insufficient response to musical beats and fail to match the musical style. To address this, BADM incorporated musical beats as input, enhancing the beat alignment of dance generation. M2C [168] demonstrated that the commonly used music features, MFCCs, were highly imbalanced at scale. It proposed a novel normalization process to address this imbalance and used discrete codes as music features for dance generation. In addition to single-person dance generation, recent research has also witnessed a surge in the generation of couple dances [169–171] and group dances [172–175].

Current music-driven dance generation methods are capable of producing relatively smooth movements and integrating common dance movements. However, there are still challenges such as inconsistency in the overall style of dance, insufficient alignment to musical beats, and lack of coherence in long dances. Future work could focus on enhancing motion coherence, beat alignment, and style consistency, as well as group and couple dance generation, hand details, facial expressions, and multimodal guidance from lyrics and videos.

3.3.3 Music-Driven Matched Video

Unlike images, videos include temporal dynamics, aligning more naturally with music. Videos that are semantically and emotionally aligned with music can enhance the expressive potential of musical content by leveraging the intuitive and immediate nature of the visual modality. This task encompasses music-driven video editing and generation, as shown in Figure 5. Music-driven video editing selects and crops video clips based on input music to produce music videos (MVs). Audeosynth [176] extracted optical flow-based video features and music features via note saliency scores, formulating selection and cropping as an optimization problem. It applied Markov chain Monte Carlo sampling to automatically generate video montages using user-input videos. Other works

bypassed user-provided videos by constructing dedicated MV corpora and selecting segments that matched the emotion and style of the music, followed by cropping and blending to produce high-quality MVs. Shin et al. [177] emphasized emotional alignment, using a dimensional emotion model to compute emotional distances between music and video. DJ-MVP [178] leveraged audio similarity and color heuristics to select beat-aligned video segments. Gross et al. [179] employed color histograms, musical key changes, and genre metadata for clip assembly. ApVideor [180] was a music-driven system for fashion video editing by first retrieving music that matched clothing style, then integrating audio and visuals accordingly.

Despite growing interest in music-driven video generation, little research has directly explored the use of music to guide the generation of general video content, which often carries more complex semantic information than music. Music2Video [181], with the assistance of CLIP and Wav2CLIP, leveraged the VQ-GAN method to achieve music video generation guided by both text and music. In contrast, specific tasks such as singing, instrument playing, and dancing have seen a surge in music-driven video generation methods. Song2Face [182], MusicFace [183], and SINGER [184] generated expressive singing videos by aligning facial expressions with vocal tracks. Liu et al. [185] proposed the Music Motion Synchronized Generative Adversarial Network (M2S-GAN) to generate orchestral conductor gestures synchronized with music. Liu et al. [186] and Kao et al. [187] generated 3D body motions for a virtual violinist using RNNs. Similarly, Li et al. [188] and Bogaers et al. [189] leveraged LSTMs to synthesize pianist body movements. Yamamoto et al. [190] generated piano-playing hand motions from scores using inverse kinematics, while Shlizerman et al. [191] employed LSTMs to predict full-body and hand gestures of both violinists and pianists. Unlike previous motion prediction studies, Chen et al. [192] propose a cross-modal GAN for instrument- and pose-oriented image sequence generation. Sound2Sight [193] combines a VAE, GAN, and multimodal transformer to generate drummer performance videos from drum audio and initial frames. Let's Play Music [194] incorporates intra-frame structural features via GCN and inter-frame temporal dynamics via a CNN-GRU, producing high-quality performance videos through a multi-stage STU framework.

Research into music-driven video has mainly focused on music-related scenarios, such as concert video generation and music video editing. This field has been limited by past technological bottlenecks in video processing. However, with recent advances in visual-language models (VLMs) and text-to-video models, we expect more breakthroughs, unlocking greater creativity in music-driven matched visual media.

4 Music-Oriented Cross-Modal Interaction

Music-oriented cross-modal interaction involves music as the output, interacting with visual or textual inputs to retrieve or generate music, enhancing context-aware applications and personalized user experiences. Key tasks include text-driven music retrieval and generation, song generation, controllable music generation, and image and video-driven music retrieval and generation.

4.1 Text-Driven Matched Music

Text-guided music matching encompasses two primary subtasks: text-based music retrieval and text-to-music generation. With the development of generation technologies, research has gradually shifted from simple retrieval tasks to more advanced generative models.

4.1.1 Text-Driven Music Retrieval

Retrieval tasks are relatively straightforward, with a longer research history. Yu et al. [195] explored cross-modal correlation learning, enabling music retrieval using lyrics. Other works [196–198] have employed multimodal metric learning, contrastive learning, and the design of music-specific word embeddings to develop more general and high-performing text-to-music retrieval systems, contributing high-quality datasets and comprehensive evaluation benchmarks. Earlier text-to-music retrieval efforts were constrained by fixed vocabularies, lacking the flexibility of natural language. Doh et al. [199] addressed this by utilizing large language models to generate rich music descriptions, enabling joint training across multiple datasets. These systems relied on audio representations. In contrast, CLAMP [200] stood out as a contrastive learning and pretraining-based cross-modal symbolic music information retrieval model, supporting symbolic semantic search and zero-shot classification.

4.1.2 Text-to-Music Generation

Text-to-image generation has made significant advances and garnered considerable attention in recent years. Using natural language to guide image creation allows individuals without drawing skills to produce artwork, as seen in the AI-generated Space Opera Theatre [201]. Music, another art form with a high entry barrier, has similarly drawn researchers to explore text-to-music generation. This approach lowers the barriers to music creation and acts as a bridge in multimodal data, enabling richer multimodal music tasks. Most works focus

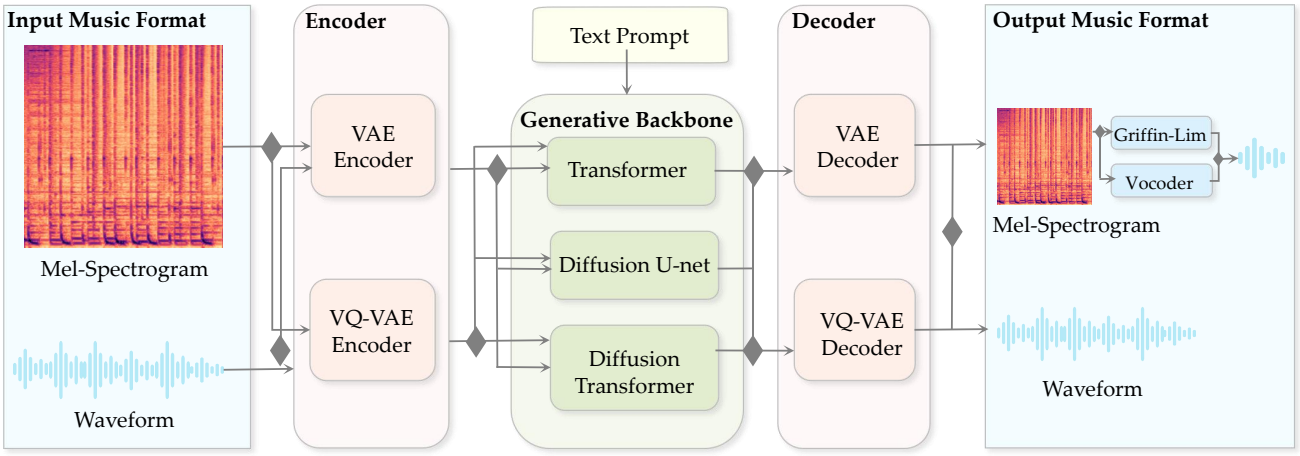


Fig. 6 Overview of the text-to-music generation pipeline based on audio representation. Each ♦ represents alternative choices.

on using text to control high-level musical features for generation, categorized by the data representation used: symbolic representation or audio representation.

Text-to-symbolic music generation refers to guiding a model with text to generate structured symbolic music data, such as MIDI, REMI, and ABC notation. MuseCoco used a two-stage framework, first extracting musical attributes (e.g., rhythm, key) from text, then generating music based on these attributes to enhance musicality and controllability. MeloTrans [202] combined neural networks with human compositional principles, particularly motif development rules, to improve the structure and musicality of symbolic music generation.

Several studies have explored the use of LLMs for symbolic music generation. ChatMusician treated music as a second language, pre-training and fine-tuning LLaMA2 [137] on ABC notation to understand and generate music without external multi-modal structures. MuPT introduced the SMT-ABC symbolic representation and the symbolic music scaling law, revealing the relationship between training data scale and model performance.

In recent years, the rapid development of large-scale generative models has significantly advanced text-to-music generation based on audio representations. These approaches can be broadly classified into Transformer-based autoregressive methods and non-autoregressive methods, including diffusion model-based approaches, as shown in Figure 6. Autoregressive methods typically encode music into discrete tokens and employ Transformer language models as the generative backbone. MusicLM [203], built upon AudioLM [204], incorporated textual control via MuLan, using a semantic-acoustic multi-stage sequence modeling approach to generate minutes-long, coherent music. MusicGen [205], based on AudioGen [206], introduced an efficient token-interleaving pattern, enabling high-quality single-stage mono or stereo music generation. It decomposes audio into K parallel codebook streams via residual vector quantization (RVQ). These streams are interleaved into a single sequence $\mathbf{X}$ to be modeled autoregressively:

$$P(\mathbf{X}) = \prod_i P(x_i | x_{<i}, \mathbf{C}),$$

where $\mathbf{C}$ is the conditioning context. This pattern enables the model to effectively capture dependencies across different codebook levels within a single stage. A significant drawback of autoregressive models is their slow inference speed. To address this, VampNet [207], STEMGEN [208], and MAGNET [209] utilized non-autoregressive Transformer models with masked modeling, achieving efficient text-to-music generation. MusicRL [210] was the first music generation system fine-tuned with human feedback. Building on MusicLM, it employed reinforcement learning (RL) with multiple reward functions and user preference data, enhancing generation quality and alignment with human preferences. To optimize for diverse objectives, MusicRL utilizes a weighted composition of reward functions defined as:

$$R(c, a) = \sum_{k \in \mathcal{K}} w_k R_k(c, a),$$

where c and a represent the text context and generated audio, and w_k is the weight for each individual reward component R_k (such as text-alignment, audio quality, and human preference). This multi-objective approach allows the model to simultaneously improve both technical fidelity and subjective appeal.

Diffusion models are another prevailing non-autoregressive generative framework. Riffusion [211] fine-tuned the text-to-image model Stable Diffusion (SD) using a small dataset of

audio spectrograms paired with text descriptions. Subsequent works, including Make-an-Audio [212], AudioLDM [213], AudioLDM2 [214], Tango [215], Tango2 [216], Diff-a-riff [217] and Auffusion [218] leveraged diffusion models to generate text-guided audio content, spanning speech, sound, and music. While these methods advanced text-to-audio generation, their music-specific capabilities lagged behind dedicated text-to-music models. Dedicated models like MusicLDM [219], Noise2Music [220], and Gen1 [221] employed diffusion models with 1D or 2D UNet architectures to model music waveforms or spectrograms, enabling text-guided music generation of fixed durations. Later efforts focused on refining diffusion-based text-to-music models to improve quality, flexibility, and efficiency. Stable Audio [222] introduced temporal embedding conditions to flexibly control the duration of generated music. MeLoDy [223], adapted from MusicLM, retained semantic token modeling with a language model while using a dual-path diffusion model for acoustic tokens, ensuring quality and enabling efficient music generation.

The DiT (Diffusion Transformer) model [224], excelling in text-to-video generation (e.g., Sora), inspired further innovation in generative frameworks, with some works adapting it for text-to-music generation. Stable Audio 2 [225] replaced the UNet-based diffusion model in Stable Audio by a Transformer-based architecture, enabling longer-duration, higher-quality music generation, alongside creative capabilities like audio style transfer. Although diffusion models excel at generating high-quality content, they suffer from numerous sampling steps and prolonged inference times. Flow matching techniques, which directly learn transformations between distributions, typically offer faster inference and fewer sampling steps. MusicFlow [226] employed cascaded semantic and acoustic flow matching models, achieving competitive results with significantly fewer model parameters while boosting inference speed, thus improving efficiency. FluxMusic [227] integrated rectified flow matching into the DiT model for text-to-music generation, using a joint text-music dual-flow and music single-flow architecture. This combined the high generation quality of DiT with the efficiency of flow matching. Other works, such as ConsistencyTTA [228], Presto [229], and DITTO-2 [230], achieved accelerated music generation through targeted distillation.

Previous text-to-music works were limited to generating instrumental music, but recent advanced studies have begun exploring text-to-song generation. Industry tools such as Suno https://suno.com/blog/introducing-v4-5, Udio https://www.udio.com/blog/introducing-v1-5, Seed-Music [231], and Meruka https://www.mureka.ai have demonstrated remarkable capabilities in generating high-fidelity songs with complex structures. Academic efforts have followed closely, with the development of language model-based song generation approaches. Melodist [232] first defined the text-to-song synthesis task, which integrated singing voice synthesis and vocal-to-accompaniment generation into a unified framework. MelodyLM [233] achieved high-quality, fully text-controlled melody generation by integrating language-model-based vocal synthesis with latent-diffusion-based accompaniment, offering maximum creative flexibility with minimal user requirements. Songcreator [234] employed a dual-sequence language model to capture the intricate dependencies between vocals and accompaniment, enabling a more versatile and controllable end-to-end song generation and editing framework. YuE [235] scaled the LLaMA2 architecture to music generation by introducing a track-decoupled next-token prediction strategy and structural progressive conditioning, enabling the synthesis of full-length songs with high lyrical alignment and sophisticated vocal-accompaniment coordination. SongGen [236] used a single-stage auto-regressive transformer for controllable song generation, which streamlined the synthesis process into a unified framework to mitigate error accumulation; it supported dual-track output modes for flexible downstream applications. LeVo [237] introduced a dual-token modeling framework that concurrently processed mixed and dual-track representations to optimize vocal-instrument harmony, while further enhancing musicality and instruction-following through a multi-preference alignment strategy based on direct preference optimization (DPO). Subsequently, Lam et al. [238] introduced MusiCoT, a CoT prompting technique for music generation, achieving high-fidelity music generation in two stages and opening new avenues for future exploration. Advances in diffusion models have further propelled this field. For instance, DiffRhythm [239] is a simple yet scalable latent diffusion-based model that enables the rapid synthesis of full-length songs, overcoming the inference speed and duration limitations of previous multi-stage or autoregressive approaches. ACE-Step [240] further improved upon DiffRhythm by incorporating song structure understanding.

Text-to-music generation serves as the foundation of multimodal music generation, continuously pushing the boundaries of a model's ability to generate music. This evolution spans from symbolic music (e.g., MIDI representations) to audio-based music, and from instrumental music to songs, with generated music having greater realism, flexibility, and diversity. Although current models generate high-quality songs and some support audio prompts, they still face sev-

eral challenges, including limited fine-grained control (e.g., of BPM and style), imprecise temporal melody control and weak long-term musical structure. Future improvements may be achieved by collecting higher-quality and more diverse datasets, exploring generation framework designs tailored to music, and incorporating domain-specific knowledge from music composition.

4.2 Music Generation with Fine-grained Control

Music generation with fine-grained control aims to produce music that adheres to specific requirements beyond text prompts by utilizing multiple conditions. Going beyond the high-level semantic control in text-to-music generation, this task enforces control over specific musical attributes such as chords, rhythm, and melody.

Several approaches have achieved fine-grained control by explicitly integrating musical prompts into the generation process. For instance, Mustango [241] expanded the scope of text-driven generation by utilizing rich captions that include specific instructions related to chords, beats, tempo, and key. Similarly, in the symbolic domain, MuseBarControl [242] enhanced the control prompts of MuseCoco by introducing bar-level attribute tags and positional embeddings. To achieve precise temporal control, researchers have developed advanced adapter mechanisms. Music ControlNet [243] adopted a strategy similar to ControlNet [244], achieving temporal control over melody, dynamics, and rhythm for the first time. JASCO [245] used a flow-matching paradigm and low-dimensional bottleneck projections to achieve temporally controllable text-to-music generation, offering superior melody adherence, accuracy, and local control. MusiConGen [246] employed an in-attention mechanism and designed an efficient fine-tuning strategy to control rhythm and chords in music. Targeting the complex task of cover song generation, SongEcho [247] employed an instance-adaptive linear modulation method to jointly synthesize new vocals and accompaniment while preserving the original melody. Unlike the aforementioned supervised methods, DITTO [248] explored training-free control. It formulated the task as optimizing the initial noise latent vector during inference, enabling pretrained models to accept arbitrary control signals without additional fine-tuning. However, the performance of such direct inference-time optimization remains limited.

Music generation with fine-grained control has advanced by combining new techniques with inputs like musical attributes and textual information to better meet creators' intentions, thereby enhancing realism, flexibility, and fine-grained control over rhythm, melody, and dynamics. Future progress may involve collaborative human–AI refinement systems.

4.3 Image-Driven Matched Music

4.3.1 Image-Driven Music Retrieval

Image-driven music retrieval aims to automatically match or recommend music based on image content, such as scenes, emotions, or artistic styles, enhancing the user experience in multimedia interactions and personalized recommendations.

Hsia et al. [249] proposed a representation learning framework to bridge the gap between images and music, extracting high-level semantic features and constructing a unified embedding space for image-to-song retrieval. Emotion served as a bridge between images and music, quantifying their similarity in the continuous valence-arousal (VA) space. Sasaki et al. [250] mapped images and music to a VA emotional plane using visual features and acoustic analysis, recommending music by calculating their distance in the plane. Zhao et al. [251] utilized cross-modal deep continuous metric learning to model the shared embedding space while preserving cross-modal similarity with unimodal VA relationships, significantly improving the finesse of emotion matching. Baijal et al. [252] further incorporated style into the retrieval process, recommending music for artworks based on both style and emotion, and for photographs, based solely on emotion. Diff4Steer [253] generated multi-directional music discovery paths from user queries by synthesizing diverse seed embeddings through a lightweight diffusion model. It supported image or text-guided retrieval, combined with nearest neighbor search for flexible and controllable music retrieval.

4.3.2 Image-to-Music Generation

Image-to-music generation transforms visual information into musical sequences that match the semantic or emotional expression of images. The main challenge lies in establishing subjective cross-modal associations while maintaining emotional and artistic consistency of music. Current research focuses on emotion mapping, cross-modal alignment, and generative model optimization, resulting in diverse approaches.

Jukepix [254] proposed a cross-modality model that converted impressionist paintings into multitrack music using deep convolutional generative adversarial networks (DCGANs). Mendis et al. [255] combined a CNN with LSTM to generate melodies matching the emotional tone of images, though only monophonic melodies were supported, and musical instruments and dynamics were not considered. Some works have used the VA emotional space to align music with the emotional tone of images. Wang et al. [256] enhanced

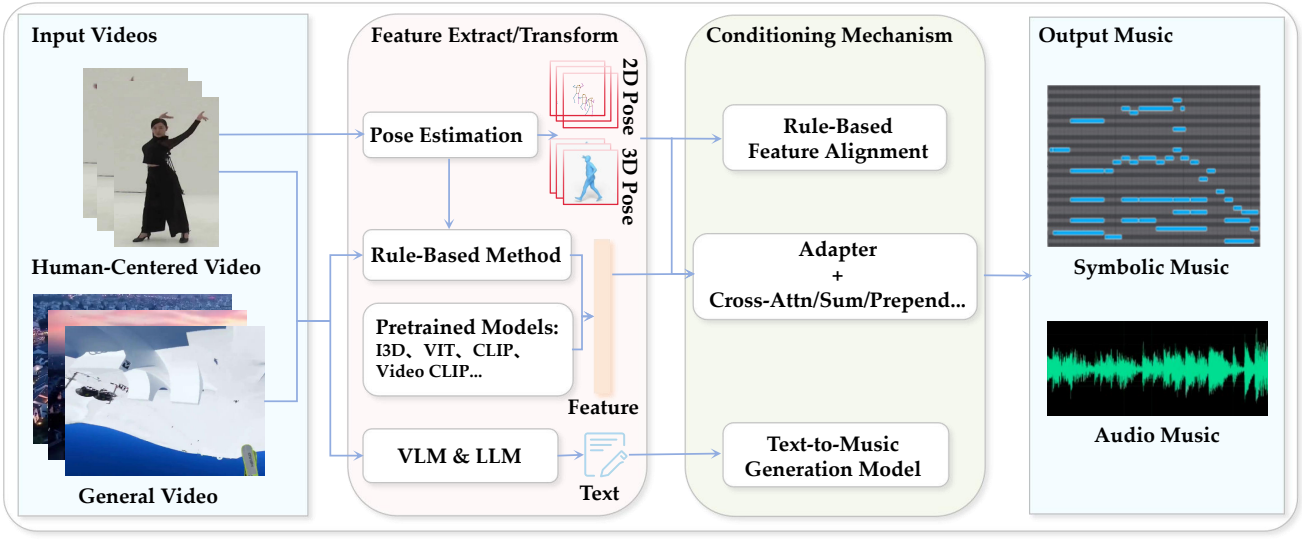


Fig. 7 Overview of the video-to-music generation pipeline based on audio representation.

emotional consistency by establishing a mapping between images and music in a continuous emotional space. Kundu et al. [257] introduced a CNN-Transformer architecture that combined a pre-trained CNN image feature extractor with Transformer to generate musically and emotionally consistent MIDI sequences. They also introduced a VA loss function that directly optimized emotion alignment, avoiding the complexity of contrastive learning. With the advance of LLMs, Hisariya et al. [258] proposed a two-stage framework, where images were first converted into emotional descriptions and then music was generated using a fine-tuned MusicGen model.

With the rise of diffusion models, researchers have explored their use for cross-modal fusion. MELFUSION [259] introduced a music diffusion model with a novel 'visual synapse' that effectively infused image semantics into a text-to-music model. Zhao et al. [260] combined latent diffusion with a Transformer to automatically generate waveform guzheng music from Chinese paintings.

Image-driven music matching has evolved along two main trajectories. Music retrieval bridges visual and auditory modalities through semantic and emotional alignment. In contrast, image-to-music generation has progressed from early GAN- and LSTM-based models generating simple monophonic melodies to more advanced approaches emphasizing emotional consistency and using diffusion models for high-fidelity, culturally specific outputs. However, retrieval systems struggle with subjective emotional interpretation and dataset biases, while generative models face challenges in maintaining long-term musical coherence. Current evaluation methods rely on synthetic datasets or limited cultural contexts, constraining real-world applicability. Future advances may arise from integrating large multimodal models, iterative human feedback, and more diverse datasets to enhance creativity and utility in image-driven matched music.

4.4 Video-Driven Matched Music

Music and videos complement each other: music enhances video atmosphere and emotion, while videos broaden dissemination and perceptual dimensions of music. This synergy drives video-guided matched music tasks, including (i) video-guided music retrieval, which uses video features to find matching music from collections, and (ii) video-guided music generation, which creates tailored music using video features via generative models such as diffusion models and autoregressive models, as shown in Figure 7.

4.4.1 Video-Driven Music Retrieval

Early work, such as Sasaki et al. [261], explored emotion analysis of video and audio features and realized emotion-consistent music recommendation using a non-deep learning approach based on V-A plane distances. Subsequently, CBMVR [262] employed neural networks to embed video and music features into a shared space and achieved bidirectional music-video recommendation using cross-modal ranking constraints and soft intra-modal structural constraints. Suris et al. [263] improved this by extracting audio features with neural networks instead of computing statistical handcrafted features and optimized recommendation with similarity and classification losses. Li et al. [264] introduced emotion label constraints into a dual-stream network and achieved efficient cross-modal music retrieval. Prétet [265] built on CBMVR and explored design choices for music-video cross-modal

recommendation, including optimization of audio features and loss functions. CMVAE [266] used a VAE model to learn associations between music and video through cross-modal reconstruction constraints, enhancing music retrieval with video and text. Earlier works overlooked temporal alignment and artistic correspondence in audio-video cross-modal recommendation. Suris et al. [267] used a Transformer to model relationships between visual features and music tag embeddings, achieving self-supervised audio-video recommendation via contrastive learning. Seg-VM-net [268] also addressed temporal alignment issues. The above methods neglected noise in data under the subjective task of video-music pairing. SSVMR [269] handled label noise, enhanced key video segment capture, and improved model robustness through semi-supervised self-training, saliency-based span mixing, and reverse retrieval. Considering music selection bias in user-created content due to historical preferences, Yi et al. [270] designed a video background music recommendation algorithm to combat the problem and increase recommendation system accuracy. More flexibly, ViML [271] retrieved music matching video content and stylistic descriptions by integrating representations of videos and texts.

4.4.2 Video-to-Music Generation

Video-to-music generation initially focused on symbolic representation-based methods, where generating music for performance videos first attracted researchers. Sight to Sound [272] employed a ResNet model to predict pitch onset events from video frames, aiming to transcribe piano performance videos into MIDI data, mitigating issues like polyphonic instruments and noise interference encountered in audio-to-MIDI transcription. Audeo [273] introduced a two-stage ResNet-GAN network to generate music from silent piano videos, emphasizing audio reconstruction. The same authors later proposed MI Net [274], which leveraged VQ-VAE to directly generate music waveforms for piano performances. Foley Music [275] extended video-to-music generation from piano to multiple instruments using human keypoints. It used a Graph-Transformer framework for better generality, accuracy, and editing flexibility. Recent works [276, 277] have also generated music from video but prioritized semantic and motion alignment over MIDI-level accuracy.

Dance, as an art form highly related to music, has led to ongoing exploration of the generation of music for dance videos. Dance2MIDI [278] and DanceComposer [279] generated rhythm and style-consistent multi-track symbolic music from dance videos, achieving rich symbolic music generation. Dance2MIDI additionally constructed a paired dataset DMIDI of dance and multi-instrumental music. In contrast, D2M-GAN generated audio waveforms directly using VQ representations, I3D, and motion features with GANs, and compiled the TikTok wild dance music dataset. CDCD [280], LORIS [281], and MDM [282] used human motion information (2D keypoints or 3D motion features) as conditions for diffusion models to generate audio music aligned with dance rhythms. CDCD and LORIS also employed I3D features as additional condition controls. CDCD supported broader cross-modal generation tasks such as text-to-image generation. The direct audio generation works mentioned above all required training models from scratch, resulting in low-quality music due to the limited availability of public dance datasets. Li et al. [283] enhanced text-to-music models for dance video scoring using encoder-based text inversion methods. MotionComposer [284] improved synchronization and style consistency between rhythm and motion videos through retrieval-augmented generation (RAG) technology and progressive training strategies.

In addition to the aforementioned human-centric videos, the internet contains plenty of video-music data pairs with diverse scenes, making general video background music generation an increasingly prominent research area. CMT [285] employed a rule-based approach to align video motion features with corresponding attributes of symbolic music, achieving video background music generation in the absence of paired video-music datasets. Video2Music [286] extracted semantic, motion, and emotional features from videos, encoded them using a Transformer encoder, and subsequently input them into a Transformer decoder to autoregressively generate MIDI notes. V-MusProd [287] presented a multi-stage Transformer framework for video background music generation. Diff-BGM leveraged visual features and textual descriptions from videos to guide a latent diffusion model to generate piano rolls with a segment-aware cross-attention module. Nevertheless, symbolic music-based approaches faced data-scale issues and often generated less expressive and diverse outputs.

Recent studies have explored the direct generation of music waveforms from video. V2Meow [288] used a multi-stage autoregressive model to generate visually aligned music from multidimensional video features, allowing high-level feature control via text. MuMu-LLaMA, a multimodal music understanding and generation model, leveraged a large language model as a bridge, building on text-to-music models to enable video-to-music generation, utilizing the MUVideo dataset focused on instrumental performances. VidMuse [289] constructed a large-scale dataset of 360,000 video-music pairs and employed long-short-term visual modules to generate music synchronized with video. S2L2-V2M [290], a novel

framework, leveraged a memory-augmented state space model to enable efficient and effective long-term video-to-music generation using large language models. M2M-Gen [291], VEH [292] and SONIQUE [293] explored the use of text as a bridge to generate background music for videos, avoiding the reliance on paired video and music data. These approaches prioritized aligning video semantics with music but lacked fine-grained rhythm control. VMAS [294] refined the standard cross-entropy loss to focus on video-music alignment by detecting video and music beats. MuVi [295], based on a DiT model with a flow-matching objective, achieved semantic alignment and rhythmic synchronization in video-to-music generation through a visual adapter. VeM [296] introduced a hierarchical 'music conductor' strategy within a diffusion framework to strictly enforce semantic and rhythmic synchronization via storyboard-guided attention. VidMusician [297] and GVMGen [298] compiled diverse paired video-music datasets and developed video background music generation methods with rhythm and semantic alignment based on MusicGen. MuVi, VidMusician, and GVMGen harnessed the generative capabilities of pretrained music generation models, with MuVi and VidMusician incorporating control conditions via token-level addition, which aided rhythm alignment. More specifically, FilmComposer [299] and HPM [300] respectively employed large model-driven agent collaboration and designed the Film Score ControlNet to achieve controllable generation of film background music. Video2Song [301] harnessed retrieval-augmented VLMs and LLMs to generate structured, video-aligned prompts for music generation, functioning as a proto-agent that exhibited rudimentary agentic behaviors. Furthermore, several recent video generation studies [302, 303] have integrated background music generation as a subtask.

Video-to-music generation has great potential but is currently limited to instrumental music. It faces challenges like insufficient high-quality datasets, weak music-video correspondence, and a lack of fine-grained rhythm alignment. To enhance video-to-music generation, future work should focus on collecting diverse, high-quality music-video datasets, refining audio-video feature alignment, and developing reasoning-capable generative models, thereby improving quality and flexibility, meeting user needs, and advancing commercialization.

4.5 Singing Voice Generation

4.5.1 Singing Voice Synthesis (SVS)

SVS can be viewed as an extension of the speech synthesis task, aiming to generate natural, expressive, and high-fidelity singing voices based on musical scores and lyrics. Current deep learning-driven SVS systems primarily adopt an acoustic model with vocoder architecture similar to speech synthesis. In fact, numerous SVS systems utilize modified versions of speech synthesis models. Despite significant advances in two-stage human voice generation, training the neural acoustic model and vocoder separately can cause a mismatch between training and inference, leading to suboptimal performance. Recent research has shifted towards exploring end-to-end unified frameworks to improve synthesis quality.

ByteSing [304] enhanced acoustic modeling by combining a duration-allocated encoder–decoder structure with powerful WaveRNN vocoders, achieving significant control over pitch and spectral fidelity in the generated singing voice. XiaoiceSing [305] followed the main architecture of FastSpeech [306] while incorporating musical features, improving fundamental frequency prediction through residual connections, and designing syllable duration loss to enhance rhythm modeling.

GANs have been widely explored for SVS. WGANSing [307] proposed a block-wise acoustic modeling approach using vocoder features to decouple pitch and timbre, enabling temporal dependency modeling and smooth synthesis. HiFiSinger [308] addressed high sampling rates by proposing a sub-frequency GAN and a multi-length GAN to optimize frequency bandwidth and long sequence modeling, introducing fundamental frequency (F_0) and V/UV as acoustic features for high-fidelity SVS. SingGAN [309] designed source excitation modules and global-local discriminators to optimize continuity, high-frequency reconstruction and low-frequency details. XiaoiceSing2 [310] employed a GAN architecture with a FastSpeech-based generator and multi-band discriminator to address mel-spectrogram smoothing issues observed in XiaoiceSing.

DiffSinger pioneered the use of diffusion models in SVS, generating mel-spectrograms through iterative Markov chain denoising. To enhance inference efficiency and detail restoration, it proposed a shallow diffusion mechanism. Instead of starting from Gaussian noise x_T, the reverse process begins at a boundary time-step K ($K \ll T$), where the initial latent x_K is constructed from an auxiliary predictor's output x_{aux} and a noise term ϵ:

$$x_K = \sqrt{\bar{\alpha}_K} x_{\text{aux}} + \sqrt{1 - \bar{\alpha}_K}\epsilon.$$

By refining the coarse spectrogram x_{aux} via K iterations of the denoising function $\epsilon_\theta(x_t, t, \mathcal{C})$, where $\mathcal{C}$ represents musical conditions, DiffSinger effectively mitigated the over-smoothing issues of traditional GANs while significantly

accelerating the sampling process. CoMoSpeech [311] applied a consistency constraint to distill a consistency model from a well-designed diffusion-based teacher model, requiring only a single sampling step to generate high-quality audio.

End-to-end models simplify traditional cascaded workflows, improving synthesis efficiency and consistency. Jukebox used a multiscale VQ-VAE to compress raw audio to discrete codes and modeled them using autoregressive Transformers. VISinger [312] followed the VITS [313] architecture, and was enhanced with a length regulator, frame prior network, and modified duration predictor. VISinger 2 [314] integrated a digital signal processing (DSP) synthesizer into the decoder to address text-to-phase mapping, glitch artifacts, and low sampling rate issues. Muse-SVS [315] introduced a joint embedding space that uniformly represents pitch, energy, and phoneme duration, allowing precise control over attribute variations and accurate emotional expression with different intensities. Prompt-Singer [316] enabled natural language control over singer gender and vocal range by decoupling pitch from melody representation, advancing human-AI interactive music creation.

Despite significant advances, current systems still face challenges in generating expressive results, real-time controllability, and producing coherent long-form outputs with natural characteristics. Future improvements may arise from multimodal learning approaches, self-supervised techniques to address data scarcity, and integration of music theory.

4.5.2 Singing Voice Conversion (SVC)

SVC aims to transform the timbre, singing techniques, or style of a source singer into those of a target singer while preserving the song content and melody. Luo et al. [317] proposed a dual-encoder framework that separately modeled latent representations of singer identity and singing techniques, enabling many-to-many conversion through vector operations. Lu et al. [318] introduced a VAW-GAN-based framework that disentangled singer identity and prosody from phonetic content to generate spectral features for unseen target singers with improved F_0 rendering. Li et al. [319] proposed an end-to-end architecture with dual encoders for content and acoustic features, enhanced by adversarial and regressive modules to improve timbre and melody modeling. FastSVC [320] utilized a conformer-based phoneme recognizer and a lightweight waveform generator to achieve high conversion performance with fast inference speed. Jayashankar et al. [321] circumvented disentanglement training and utilized ASR-finetuned self-supervised features as inputs to a HiFi-GAN vocoder. Ning et al. [322] proposed a VITS-based singing voice conversion model using Whisper-extracted bottleneck features and multi-scale pitch information, trained with a three-stage strategy to adapt it to limited target speaker data.

The emergence of diffusion models has brought breakthrough advances in SVC. DiffSVC [323] pioneered the application of denoising diffusion probabilistic models to SVC tasks, achieving high-quality spectrogram generation by predicting Gaussian noise and surpassing traditional methods in naturalness and voice similarity. However, its iterative sampling process limited inference speed. To address this, CoMoSVC [324] achieved single-step sampling through consistency model distillation, improving inference speed while maintaining high-quality generation. Similarly, LCM-SVC [325] reduced the inference steps of latent diffusion models (LDMs) using latent consistency distillation techniques, significantly accelerating the generation process. To address the issue of timbre leakage, LDM-SVC [326] pretrained a variational autoencoder based on the open-source So-VITS-SVC project and introduced a singer-guided training method to further suppress the timbre of the original singer. Additionally, the visual analysis system SingVisio [327] enhanced model generation mechanism interpretability through interactive diffusion process visualization.

SVC has evolved from early approaches leveraging disentangled representations via VAEs and adversarial training to advanced diffusion-based frameworks, achieving high-fidelity timbre transformation while preserving melody and content. Despite progress, challenges persist in fine-grained disentanglement of timbre, pitch, and expressive nuances, particularly in cross-linguistic or low-resource scenarios. Future directions may focus on improved disentanglement mechanisms, lightweight architectures, and standardized benchmarks.

4.6 Multimodal-Driven Matched Music

The foregoing discussion outlined research on acquiring matched music through methods driven by various single modalities. In contrast, other studies aim to develop a unified model that leverages multiple modalities for collaborative music acquisition driven by diverse inputs. Shen et al. [328] enhanced music recommendation by constructing a multimodal dataset, defining handcrafted and deep features, and proposing an attention-based multimodal autoencoder (AMAE) that combined text and visual features. However, generating music rather than retrieving it represents a prevailing trend in current developments. MuMu-LLaMA employed a large language model as a bridge between multiple pretrained modality encoders and a music generator, achieving the generation of music conditioned on text, images, and video. Mozart's Touch [329] utilized BLIP [330] to convert visual

content into text, directly expanding the multimodal music generation capabilities of the text-to-music model MusicGen. VMB [331], built upon Stable Audio, demonstrated superior performance in multimodal music generation tasks through dual-track retrieval augmentation. More comprehensively, AudioX [332], a unified diffusion transformer, generated high-quality audio and music from diverse multimodal inputs using a novel masked training strategy. Instead of directly generating audio, XMusic [333] introduced a novel framework that leveraged flexible multimodal prompts to produce generalized and emotionally controllable symbolic music.

Works on arbitrary-to-arbitrary generation have treated multimodal music generation as a subtask. CoDi [334] used multiple modality encoders and diffusion models with alignment to generate arbitrary output combinations from varied inputs, but struggled with complex tasks. NExT-GPT [335], an end-to-end multimodal LLM, integrated components with alignment learning and instruction tuning for robust input-to-output generation and strong complex task handling. AnyGPT [336] employed a unified multimodal tokenizer, converting diverse modalities (images, music, speech) into discrete semantic tokens for uniform processing and generation in an LLM.

While exhibiting basic capabilities in generating matched music, multimodal music generation offers a promising, flexible cross-modal interaction system between music and multimodal data, with significant potential for enhancing user music creation experiences. However, integrating multiple modalities poses modeling challenges, with room for improvement in deep multimodal alignment and understanding, adaptability, and personalized generation capabilities.

5 Bidirectional Music Cross-Modal Interaction

Bidirectional cross-modal music interaction involves the dynamic interplay between music and other modalities, enabling both the generation of music from diverse inputs and the extraction and conversion of music information into other content modalities. Specifically, it takes the following forms:

(1) *Multimodal music generation and understanding:* a model generates music from visual, textual, or other inputs and extracts information from music for understanding.
(2) *Multimodal music editing:* editing music using other modalities as conditions.
(3) *Music agents:* enabling deep, multimodal, multi-turn dialogues for personalized, context-aware music creation and interaction.

5.1 Multimodal Music Understanding and Generation

Some studies [337] have focused on symbolic music generation and understanding. The ABC notation shares similarities with text in terms of data format. ChatMusician, based on LLaMA2, underwent continual pretraining and fine-tuning on music data represented in ABC notation. It treated music as a second language, achieving strong performance in music understanding, generation, and linguistic capabilities. However, it faced challenges such as hallucination and memory degradation. Other studies [338] have focused on audio-based music understanding and generation. MuMu-LLaMA integrated multiple pretrained modality feature encoders via designed adapters, unifying different modality information into a LLaMA model. Multimodal music understanding and generation were achieved by organizing multimodal datasets and designing multi-stage training strategies.

5.2 Multimodal Music Editing

Some autoregressive text-to-music generation models [203, 205] have supported melody-guided generation. Hou et al. [339] enhanced the DiT model using ControlNet, introducing top-k CQT as a melody prompt to achieve the above task.

For comprehensive music editing, AUDIT [340] and InstructME [341], both based on diffusion models, were trained from scratch to implement audio editing guided by human instructions. They supported various audio editing tasks, including addition, deletion, replacement, restoration, and super-resolution. InstructME further supported mixing operations.

Several works have explored fine-tuning strategies for music editing. Instruct-MusicGen [342] incorporated an audio fusion module and a text fusion module into MusicGen, fine-tuning it with instructions to achieve music editing. Lin et al. [343, 344] enhanced content-controllable music generation by designing parameter-efficient fine-tuning (PEFT) methods for MusicGen. AP-Adapter [345] incorporated audio as an input for AudioLDM2, which extracted audio features using AudioMAE and integrated them into the model through an adapter, to achieve music editing. Some studies have explored training-free strategies for music editing. Manor et al. [346] proposed two zero-shot audio editing methods: Zero-shot Text-to-Audio (ZETA) editing and Zero-shot Unsupervised (ZEUS) editing. Zhang et al. [347] determined editing direction by computing the difference between two embedding sets of different captions. MELODYFLOW [348] introduced a flow-matching architecture with regularized latent inversion, significantly improving stereo generation

fidelity and consistency in text-guided editing. MEDIC [349] proposed a disentangled inversion control framework using a triple-branch architecture and coordinated attention to enable zero-shot music editing. Additionally, some studies have focused on music style transfer. Li et al. [350] and Kim et al. [351] achieved music style transfer through textual inversion and training-free approaches, respectively.

5.3 Music Agents

Music agents mark a new era in AI music systems, leveraging large language models (LLMs) as a central brain to bridge multimodal perception and generation through reasoning and planning. AudioGPT combined audio foundation models with a large language model, enabling music understanding and generation via multi-turn dialogue, though limited by prompt engineering, token constraints, and model performance. Loop Copilot [352] and MusicAgent [353] used LLMs to integrate music tools, automating task analysis and model selection for various music tasks. ByteComposer [354] simulated human composition workflows—concept analysis, draft creation, self-evaluation and revision—aligning AI composition with human creative logic to improve interactivity and interpretability. ComposerX [355], built on GPT-4, was the first multi-agent symbolic music composition system, improving quality through role-based collaboration.

However, these systems primarily function as tool integrators and are often constrained by the performance of underlying models, leaving significant room for improvement. Music agents hold the potential to transcend these limitations by infusing the robust generative capabilities of platforms like Suno with the advanced planning and reasoning faculties of LLMs. To achieve professional-level autonomous composition, future agents must address the critical hurdle of long-term planning. Unlike linear generation, music demands rigid structural coherence; thus, future agents could adopt hierarchical planning mechanisms, such as recursive LLMs or multi-agent collaboration, to draft global blueprints before micro-generation. Furthermore, incorporating reinforcement learning from human feedback (RLHF) and agentic closed-loop feedback is essential, enabling agents to 'listen' to and critique their own generated results for iterative refinement. Crucially, by orchestrating diverse downstream tools, agents could facilitate the synergistic co-creation of symbolic and audio music, combining the structural editability of symbolic representations with the high fidelity of audio synthesis. Finally, expanding beyond composition, agents could integrate music-driven visual generation (Sec. 3.3) and visual-to-music generation (Sec. 4.4) models to serve as multimodal producers. By integrating visual media, agents could orchestrate complex applications like MV sequence generation or AR concerts, providing intelligent support for the entire creative lifecycle.

6 Datasets and Evaluation

The development of cross-modal interaction between music and multimodal data relies on the emergence of diverse datasets involving music and other modalities. Researchers have introduced new evaluation metrics to provide more comprehensive performance validation and support for this field.

6.1 Datasets

We have compiled a list of existing multimodal music datasets in Table 2, together with essential descriptions and applicable tasks, providing researchers with a concise overview of available datasets in their areas of interest. Music-text datasets are relatively abundant and form the basis for advances in text-to-music generation and music understanding. In contrast, music-image paired datasets are limited, likely due to the absence of temporal structure in images, which complicates modality alignment. Music-video datasets are an emerging area, yet despite the abundance of music-containing videos on the internet, collecting large-scale, high-quality paired datasets remains challenging. On one hand, many video audio tracks contain noise, such as ambient sounds or speech, in addition to music. On the other hand, due to the ambiguity and diversity of music-video pairing patterns, user-produced videos often feature music that is not well-aligned; more commonly, while semantically harmonious, the music and video lack precise temporal synchronization. Automated filtering methods are underdeveloped, while manual filtering is both expensive and subjective, making large-scale data acquisition difficult.

Early music-video datasets primarily focused on MV collections [262, 271, 287]. Recently, driven by the growing demand for general video music generation, newer datasets [289, 297, 298] have drawn from a broader range of video scenarios. Dance videos, a subclass of music-video datasets, are often studied independently. AIST++ is currently the most widely used dataset for music generation from dance. However, its constrained scene diversity and repetitive choreography and songs limit the performance of dance-to-music generation. The development of more complex and large-scale in-the-wild dance datasets is now a trend in this area.

6.2 Evaluation

Table 3 summarizes the objective evaluation metrics currently used in music-related cross-modal tasks, along with brief

Table 2 Overview of multimodal music datasets.

Type	Dataset	Modality	Size	Open	Note	Applications
Music-driven cross- modal interaction	Jamendo Lyrics [85]	audio, text	80	Yes	80 songs with different genres and languages along with lyrics time-aligned on a word-by-word level (with start and end times) to the music	Automatic lyrics transcription (ALT) & alignment
	MTG-Jamendo [30]	audio, text	55k	Yes	Over 55,000 full audio tracks with 195 tags covering genre, instrument, and mood/theme.	Music auto-tagging
	lyrics-melody dataset [356]	audio, text	12,197	Yes	Paired lyrics and melody alignment, from different music sources where the alignment relationship between syllables and music attributes is extracted.	Lyrics generation
	WASABI Song Corpus [357]	audio, text	1.73M	Yes	A large corpus of songs enriched with metadata extracted from music databases on the Web	Lyrics interpretation
	SongCompose-PT [89]	audio, text	283k	Yes	Large-scale song pretraining dataset that includes lyrics, melodies, and paired lyrics-melodies in either Chinese or English	Lyrics generation
	Chinese-Lyric-Corpus [358]	text	50k	Yes	Chinese lyrics corpus crawled from NetEase Cloud Music	Lyrics generation
	LP-MusicCaps [74]	audio, text	22M	Yes	LLM based pseudo music captioning dataset	Music caption
	MusicInstruct [70]	audio, text	5.5k	Yes	Music-instruction pairs containing different styles and types	Music understanding
	OpenMU-Bench [73]	audio, text	1M	Yes	Extended from existing datasets with rich annotations by GPT-3.5	Music understanding
	MusicQA [71]	audio, text	12.5k	Yes	Question-answer pairs built on MusicCaps and MagnaTagATune [359]	Music understanding
	GrooveNet [146]	audio, motion	0.4h	Yes	Music-dance paired data collected through Mocap with single dance genre	Music-to-dance generation
	Dance w/Melody [147]	audio, motion	1.6h	Yes	Music-dance paired data collected through Mocap in 4 dance genres	Music-to-dance generation
	Music2Dance [149]	audio, motion	1h	Yes	Music-dance paired data collected through Mocap with hand motion, 2 dance genres	Music-to-dance generation
	EA-MUD [360]	audio, motion	0.4h	Yes	Music-dance paired data, 4 dance genres	Music-to-dance generation
	FineDance [361]	audio, motion	346/14.6h	Yes	Music-dance paired data in SMPL-X format, with detailed hand motions, 22 dance genres	Music-to-dance generation
	PhantomDance [154]	audio, motion	260/9.6h	Yes	Music-dance paired data in SMPL format, 13 dance genres	Music-to-dance generation
	POPDG [362]	audio, motion	307/3.5h	Yes	Music-dance paired data for popular dance in SMPL format, 19 dance genres	Music-to-dance generation
	AIST-M [169]	audio, motion	1h	Yes	Music-dance paired data of interactive couple-dances in SMPL format from AIST, 10 dance genres	Duet dance generation

Continued on next page

Table 2 (continued) Overview of multimodal music datasets.

Type	Dataset	Modality	Size	Open	Note	Applications
	DD100 [170]	audio, motion	2h	Yes	Music-dance paired data of interactive couple-dances in SMPL-X format, 10 dance genres	Duet dance generation
	InterDance [171]	audio, motion	4h	To be released	Music-dance paired data of interactive couple-dances in SMPL-X format, 15 dance genres	Duet dance generation
	MDD [363]	audio, motion, text	10.3h	To be released	Music-synchronized 3D couples motion with over 10K text annotations	Text/Music-driven duet dance generation
	AIOZ-GDANCE [172]	audio, motion	16.7h	Yes	Music-dance paired data of group dances in SMPL format, 19 dance genres	Group dance generation
	MDC [174]	audio, motion	73/1h	No	Music-dance paired data of group dances in fbx format	Group dance generation
	SingingHead [364]	audio, video, motion	448, 27h	Yes	2D portrait videos of 76 singers with singing audio, background music and 3D facial motions	Singing head animation
	Song2Face [182]	audio, motion	2h	No	Vocal audio with 3D facial motions (blendshapes) of 7 singers	Singing head animation
	SingingFace [183]	audio, motion	600/40h	No	Singing audio, background music with 3D facial motions and eye state of 7 singers	Singing head animation
	SHV [184]	audio, video	200/20h	To be released	In-the-wild singing video with background music and background image	Singing head animation
	Conductor-Motion100 [185]	audio, motion	100h	Yes	Orchestral conductor motions and aligned music mel spectrogram	Conducting motion generation
	URMP [365]	audio, video, MIDI	44	Yes	Classical chamber music with multi-instrument annotations, frame/note-level transcriptions	Audio-visual synchronization, instrument performance motion generation
	Violin&Piano Performance Dataset [366]	audio, motion	215	Yes	Violin and piano audio with performance motion	Instrument performance motion generation
	AudioSet-Drums [193, 367]	audio, video	8k	Yes	Drum performance video with visible drummer and clear drum beat from AudioSet	Instrument performance motion generation
	MMIP [368]	audio, multi-angle video, motion, MIDI	3.5h	Yes	Synchronized multi-modal recordings of professional musicians performing on guitar, digital piano, and drums	Instrument performance motion generation
	Music-AVQA [369], Music-AVQA-R [370]	audio, text, video	9,288	Yes	Question-answer pairs about music performance	Audio-visual question answering, music understanding
Music-oriented cross-modal interaction	MusicCaps [203]	audio, text	5.5k	Yes	10s music examples labeled with aspect lists and captions written by musicians	Text-to-music generation
	MusicBench [241]	audio, text	52.8k	Yes	Extended from MusicCaps	Controllable text-to-music generation
	MUCaps [133]	audio, text	18.5k	Yes	10s music with synthesized language descriptions	Text-to-music generation
	MUSICGEN [205]	audio, text	400k	No	Instrument-only music tracks	Text-to-music generation
	Stable Audio Open Dataset [225]	audio, text	486k	Yes	CC recordings from Freesound and the Free Music Archive (FMA)	Text-to-music generation

Continued on next page

Table 2 (continued) Overview of multimodal music datasets.

Type	Dataset	Modality	Size	Open	Note	Applications
	XMIDI [333]	text, MIDI	108,023	Yes	Largest known symbolic music dataset with precise emotion and genre labels	Text-to-symbolic music generation
	MUImage [133]	audio, text, image	14,520	Yes	Music samples and corresponding images from the Balanced-AudioSet	Image-to-music generation
	IMEMNet [251]	Image, audio	144k	Yes	Image-music pairs with continuous valence-arousal emotion annotations	Image-driven music retrieval
	EIMG Dataset [256]	Image, MIDI	3k	Yes	Images with VA labels and music clips annotated with continuous valence-arousal values, paired by emotion similarity score in VA space	Image-to-music generation
	EMID [371]	Image, audio	10,738	Yes	Music clips paired with 3 images in the same emotional category, and rich annotations	Image-to-music generation
	MeLBench [259]	image, text, audio	11,250	Yes	Manually annotated triplets of image, text and music, including 15 predefined genres, various image types and detailed text descriptions	Image and text-to-music generation
	Extended MusicCaps [259]	image, text, audio	7,684	Yes	Extended from MusicCaps dataset by adding corresponding images to each text-music pair	Image and text-to-music generation
	Emotion Painting Music Dataset [258]	image, audio	1,200	No	Paired paintings and music clips categorized into five emotions (Happy, Angry, Sad, Fun, Neutral)	Image-to-music generation
	Chinese Paintings and Guzheng Audio Dataset (CPGAD) [260]	image, audio	22,103	No	Paired dataset of Chinese paintings and guzheng music with semantic annotations	Image-to-music generation
	YouTube8M-MusicTextClips [271]	audio, text, video	4k	Yes	10s music videos with synthesized language descriptions	Video-to-music recommendation
	HIMV-200k [262]	audio, video	200k	Yes	Music videos derived from the YouTube-8M dataset	Video-to-music recommendation
	TT-150k [266]	audio, text, video	150k	No	Micro-videos based on 3k music	Video background music recommendation
	SymMV [287]	audio, video, MIDI	1,140	Yes	Music videos with rich musical annotations	Video-to-music generation
	BGM909 [55]	audio, text, video, MIDI	909	Yes	Music videos with rich annotations	Video-to-music generation
	MUVideo [133]	audio, text, video	14,504	Yes	Music samples and corresponding videos from the Balanced-AudioSet	Video-to-music generation
	MuVi-Sync [286]	audio, video	748	Yes	Music videos with rich features (chords, tonality, scene offsets)	Video-to-music generation
	V2M [289]	audio, video	360k	Yes	Diverse video-music pairs	Video-to-music generation
	GVMGen [298]	audio, video	147 h	Yes	Diverse video-music pairs including Chinese traditional music	Video-to-music generation
	DVMSet [297]	audio, video	3,839	To be released	Diverse video-music pairs	Video-to-music generation

Continued on next page

Table 2 (continued) Overview of multimodal music datasets.

Type	Dataset	Modality	Size	Open	Note	Applications
	MusicPro-7k [299]	audio, text, video	7,418	Yes	Film music dataset with film clips, music, visual and music descriptions, melodies, and rhythm spots per sample	Film-to-music generation
	FilmScoreDB [300]	video, music	32,520	To be released	90.35 hours of film clips and corresponding original scores by renowned film composers	Film-to-music generation
	HarmonySet [372]	audio, video	48,328	Yes	Diverse video-music pairs	Video-to-music generation
	LORIS [281]	audio, video	12,446	Yes	25 s paired videos, including 8,585 figure skating and 1,950 floor exercise clips	Sports-to-music generation
	TikTok [373]	audio, video	445	Yes	In-the-wild dance videos with complex backgrounds	Dance-to-music generation
	Suno Music Generation Dataset [374]	audio, text, video, image	660k	Yes	Metadata for 659,788 songs generated by artificial intelligence on the suno.com platform	Any-to-music generation
Bidirectional cross-modal interaction	AIST++ [152]	audio, video	1,408	Yes	Multi-view dance dataset, 10 dance genres	Music-dance mutual generation
	AIST-M2B [142]	audio, motion	1408	Yes	Algorithm for extracting motion rhythm labels from music-dance paired data in the AIST++ dataset	Motion rhythm analysis, dance-music alignment
	Dance50 [141]	audio, video	14k, 47h	No	K-pop dance clips collected from website with 2D skeleton keypoints extracted by OpenPose	Motion rhythm analysis, Dance-music alignment
	AnyInstruct [336]	audio, text, image	108k	Yes	Multimodal interleaved instruction-following data, constructed by a generative model	Any-to-any generation
	MMtrail [375]	audio, text, video	20M	Yes	27k+ hours of trailer clips, multi-modal captions	Multimodal understanding and generation
	EmoMV [376]	audio, video	5,986	Yes	Humanly and automatically annotated labels for emotional correspondence	Music-video correspondence learning
	MusicPile-sft [59]	text, ABC notation	1,139,473	Yes	Samples with a ratio of music verbal to score of 2:1, created using music from various regions	Music understanding and generation
	MUEdit [133]	audio, text	5k	Yes	Instruction-response pairs for music and text	Instruction-based music editing
Music-driven & Music-oriented interaction	Youtube-Music-1M [336]	audio, text	1M	Yes	Music-text pairs from Youtube	Music captioning, text-to-music generation
	WikiMusicText (WikiMT) [200]	ABC, text	1010	Yes	Sourced from Wikifonia.org, lead sheets with title, artist, genre, and description.	Text-to-music generation, music captioning
	Song Describer Dataset (SDD) [377]	audio, text	706	Yes	High-quality benchmark for evaluating music-and-language models	Music captioning, text-to-music generation, text-to-music retrieval

Table 3 Overview of objective evaluation metrics for music-related cross-modal tasks.

Type	Metric	Description	Applications
	Accuracy	Ratio of correctly classified instances to total number of instances	Multimodal representation learning
	Recall	Ratio of all relevant items successfully retrieved in the top-K results	Multimodal representation learning

Continued on next page

Table 3 (continued) Overview of objective evaluation metrics for music-related cross-modal tasks.

Type	Metric	Description	Applications
	Mean average precision (mAP)	Mean of average precision over all queries, measuring both relevance and ranking quality	Multimodal representation learning
	BLEU [378]	n-gram overlap between generated text and reference answers	Music understanding
	Meteor [379]	Text similarity considering precision, recall, and alignment based on exact, stem, synonym, and paraphrase matches	Music understanding
	Rouge-L [380]	Text similarity by longest common subsequence (LCS), capturing semantic continuity and word order	Music understanding
	BERT-score [381]	Semantic similarity using BERT embeddings of generated and reference text	Music understanding
	Word error rate (WER)	Edit distance (Levenshtein distance) between the hypothesis (predicted transcription) and the reference (ground-truth transcript)	Lyrics transcription
	Beat align score (BAS) [152]	Motion-music correlation using the average distance between each kinematic beat and its nearest music beat	Music-to-motion generation
	Diversity (Dist) [152]	Generation diversity using the average Euclidean distance in feature space across generated motions	Music-to-motion generation
	Landmark distance (LMD) [382]	Evaluation of lip synchronization using Euclidean distance between mouth landmarks between ground-truth and generated motions	Music-to-motion generation
	Frechét inception distance (FID) [152]	Motion quality using FID based on geometric and kinetic features between ground-truth(GT) and generated motions	Music-to-motion generation
	Standard structural similarity (SSIM) [383]	Quality of generated video frames compared to ground-truth using differences between the properties of the pixels	Music-to-video generation
	Signal to noise ratio (PSNR)	Quality of generated video frames compared to ground-truth using absolute error between pixels.	Music-to-video generation
Music-oriented cross-modal interaction	Frechét audio distance (FAD)	Similarity of distributions of generated audio and real music based on VGGish [384]	Any-to-music generation
	Frechét distance (FD)	Similarity between distributions of generated audio and real music based on PANNS [385] or OpenL3 [386]	Any-to-music generation
	Kullback-Leibler divergence (KL)	Divergence between class predictions of ground truth and generated music based on PANNS [385] or PaSST [387]	Any-to-music generation
	Inception score (IS)	Diversity and quality of generated music, measured by entropy of class predictions using a pre-trained music classifier (e.g., VGGish [384] or PANNS [385])	Any-to-music generation
	CLAP [119] score	Similarity of CLAP embeddings of music and text	Text-to-music generation
	ClaMP3 [130] score	A more recent metric similar to CLAP score	Text-to-music generation
	Phoneme error rate (PER)	Phoneme recognition accuracy using FireRedASR [388]	Any-to-song generation
	Vocal range	Quantifies the breadth of the vocal range at the song level (pitch estimated with RMVPE [389])	Any-to-song generation
	Imagebind score [127]	Similarity of ImageBind embeddings of music and video	Video-to-music generation
	Beats coverage score (BCS)	Ratio of generated beats to ground truth beats	Dance-to-music generation
	Beats hit score (BHS)	Ratio of aligned generated beats to number of ground truth beats	Dance-to-music generation
	Average sample-wise accuracy (ASA)	Proportion of correctly predicted attributes per sample, averaged over the test set	Text-to-music generation
	Tone span (TS)	Number of half-tone steps between the lowest and highest tones in a melody	Any-to-symbolic-music generation
	Scale consistency (SC)	The fraction of tones in a melody that belong to standard musical scales	Any-to-symbolic-music generation
	Polyphony rate (PR)	The fraction of time steps with multiple simultaneous pitches, measuring polyphonic complexity	Any-to-symbolic-music generation
	Pitch entropy (PE)	Quantification of pitch diversity using information entropy	Any-to-symbolic-music generation

Continued on next page

Table 3 (continued) Overview of objective evaluation metrics for music-related cross-modal tasks.

Type	Metric	Description	Applications
	Groove consistency (GC)	Rhythmic stability = 1− average Hamming distance of onset vectors between consecutive measures	Any-to-symbolic-music generation
	Repetition rate	how often the repeat sign : | appears in a generated set	Any-to-symbolic-music generation
	Intra similarity	Average value of texture latent similarity matrix, excluding the diagonal	Any-to-symbolic-music generation
	Time signature error rate (TSER)	Assesses whether the beat count in each measure corresponds to the time signature in the ABC notation header	Any-to-symbolic-music generation
	Instrument range error rate (IRER)	Assesses how often generated notes exceed the expert-defined instrument range	Any-to-symbolic-music generation
	Score information completeness rate (SICR)	Assesses completeness of the ABC notation	Any-to-symbolic-music generation
	Average attribute accuracy (AAA)	How often attributes in the generated ABC notation match those defined, averaged over all musical attributes	Any-to-symbolic-music generation
	Hit rate@K	Proportion of top-K recommended songs that contain keywords or conceptually similar terms associated with the input	Any-to-music retrieval
	Recall@K	Proportion of relevant items retrieved within the top-K results	Any-to-music retrieval
	Entropy (H@K)	Distribution of ground-truth genres using the formula on the top-K results	Any-to-music retrieval
	Emotion matching (EM)	Average inverse Euclidean distance between predicted music VA values and image VA values in the valence-arousal space	Image-to-music generation
	Image music similarity metric (IMSM)	The similarity between image-music pairs via CLIP image-text and CLAP audio-text embeddings	Image-to-music generation
	Mel-cepstral distortion (MCD)	The mean Euclidean distance of mel-cepstral coefficients, measuring spectral distortion between synthesized and original voices	Singing voice synthesis/conversion
	Root-mean-square error (RMSE) of F_0 values	The square root of the mean of squared errors between the natural and synthesized F_0 (fundamental frequency) values, measuring pitch prediction accuracy	Singing voice synthesis
	Correlation coefficients (Corr.) of F_0 values	Linear correlation coefficient relating natural and synthesized F_0 values	Singing voice synthesis
	Duration root mean square error (Dur RMSE)	Root mean square error of phoneme durations	Singing voice synthesis
	Duration correlation (Dur CORR)	Correlation coefficient relating predicted phoneme duration and the actual duration	Singing voice synthesis
	Real-time factor (RTF)	Ratio of times of transforming embeddings into latent representations/mel-spectrograms to the duration of audio, measuring inference speed	Singing voice conversion
	Character error rate (CER)	Accuracy of lyrics in converted audio, using Whisper	Singing voice conversion
	Speaker similarity (SIM)	Cosine distance of speaker embeddings, measuring similarity of singer's timbre	Singing voice conversion
	F_0 Pearson correlation coefficient (FPC)	Correlation between F_0 values of original and converted audio, reflecting the accuracy of pitch conversion	Singing voice conversion
Bidirectional cross-modal interaction in music	Chroma similarity	Cosine similarity of chromagrams of original and edited music	Music editing

descriptions and their associated tasks. Given the diversity of tasks, we analyze the significance of these metrics in four critical ways. First, for tasks like music retrieval and captioning, standard information retrieval metrics (e.g. accuracy, mAP) and NLP metrics (e.g. BLEU, WER) serve as the fundamental baseline for quantifying a model's ability to identify musical attributes or transcribe lyrics with linguistic precision. Second, in the domain of audio and video generation, distribution-based metrics like FAD and IS quantify the statistical distance between generated and reference feature distributions. They are critical for jointly assessing fidelity and diversity, ensuring that the generated signals align with the perceptual characteristics of real-world data. Third, interaction tasks require dual alignment checks: semantically, metrics like CLAP score quantify the semantic correspondence (e.g. style, and emotional alignment) between music and text, while temporally, metrics such as BCS are indispensable for ensuring visual movements synchronize with musical rhythms. Finally, specific to symbolic generation, domain-specific metrics like pitch entropy and groove consistency are essential to validate adherence to harmonic rules and structural patterns.

However, unlike the well-established metrics for assessing musical attributes like quality, measures for evaluating cross-modal alignment between music and other modalities are still in the exploratory stage. A prevalent approach involves extracting features using contrastive learning-based multimodal pretraining models [119, 127, 130], and computing similarity scores across modalities. However, most of these models are trained on general audio data rather than music-specific datasets, resulting in limited representational capacity for music and often overlooking temporal characteristics. Moreover, due to the artistic nature of music, its alignment with other modalities is inherently more flexible. For example, a single video may be effectively paired with multiple music tracks, each creating a distinct but coherent audiovisual experience. This raises an open challenge: how to design evaluation metrics that balance tolerance for flexible pairings with the need for alignment accuracy.

For subjective evaluation, human assessment remains the gold standard. Given that music is an art form driven by human perception, objective metrics often fail to capture aesthetic nuances, emotional resonance, and culturally appropriate stylistic attributes. Researchers typically define task-specific criteria and conduct user studies. These primarily take two forms: rating-based evaluations, such as the human mean opinion score (MOS), where users assign scores to different aspects (e.g., semantics, rhythm, overall quality) of a sample; and preference-based evaluations, where users choose between options, yielding rates that reflect comparative preferences.

7 Discussion

Despite progress in research involving music and other modalities, notable challenges remain. One critical issue is the lack of high-quality multimodal music datasets, especially ones with paired music–visual data. For example, although numerous music videos exist online, many of them suffer from poor quality. On the other hand, most existing works on cross-modal interactions of music focus on instrumental music, while a large portion of online music takes the form of songs, leading to issues such as the limited availability of instrumental music and data imbalance between music genres. While source separation models can extract instrumental tracks from songs, their limited performance introduces noise into the audio.

Additionally, the subjective nature of cross-modal alignment and varying levels of users' creative abilities contribute to inconsistencies in internet-sourced data. Manual annotation is also time-consuming and requires expert knowledge.

Another limitation is that current research typically treats pure instrumental music and vocal music separately; most cross-modal music interaction studies are based on instrumental music. However, directly generating full songs is both more efficient and better aligned with the way music is commonly represented on the Internet, making it more user-friendly in practical applications.

Moreover, existing models often exhibit deficiencies in the depth and breadth of modality fusion when handling multimodal data. Many approaches simply concatenate features from different modalities or force them into a shared latent space without truly guiding the model to comprehend the semantic associations and logical relationships between modalities. As a result, there is a lack of high-performance music understanding models and multimodal models tailored for music, comparable to the performance of vision-language models (VLMs) in the visual domain.

Furthermore, music agents also face certain limitations. While they have achieved progress in tasks such as music composition assistance and basic music analysis, their performance in complex music understanding and generation remains weak. When processing multimodal inputs, music agents may suffer from information loss, leading to decision-making and outputs that lack coherence. Meanwhile, current music agents lack effective feedback mechanisms to enable self-improvement or alignment with personalized preferences.

Moreover, current music agents are based on fixed, manually designed workflows. They lack deep integration with large language models and efficient integration of diverse music tools, limiting their ability to autonomously plan workflows and handle various music processing tasks.

Despite current limitations, several promising trends are emerging in this field:

(1) *High-quality data construction:* advances in technology will improve multimodal data collection and automated annotation, yielding richer, high-quality music data. Automated annotation benefits from the synergy between music generation and understanding, further augmented by LLMs.

(2) *Adaptation of speech techniques:* music and speech share fundamental signal representations (e.g., spectrograms) and generative backbones (e.g., Transformers). This implies significant potential for transferring acoustic priors from large-scale speech models to music tasks. However, bridging the domain gap requires more than direct application; future research should incorporate hierarchical tokenization to decouple high-level semantics from acoustic nuances, long-context memory mechanisms to accommodate music's extended temporal nature, and relative positional encodings to ensure strict alignment with musical bars and beats.

(3) *Hybrid symbolic-audio frameworks:* while end-to-end audio generation dominates in acoustic fidelity, symbolic generation remains indispensable for its explicit structure and editability. Unlike raw audio, symbolic representations offer transparent formats allowing precise refinement. Consequently, the evolving paradigm points towards hybrid models that leverage symbolic priors for structural control and neural synthesizers for high-fidelity rendering. This synergy fosters artist-friendly systems, enabling interactive editing and guiding AI music generation towards professional-grade quality.

(4) *Multimodal CoT:* the construction of comprehensive multimodal music CoT datasets holds promise for helping models learn logical correspondences across different modalities.

item *LLM-Driven multimodal integration:* researchers are exploring novel model architectures and leveraging LLMs to improve multimodal integration, thereby capitalizing on the consistency between music and other modalities for controllable cross-modal generation and understanding.

(5) *Alignment with human preferences:* preference data can also be constructed and utilized through methods such as DPO to align generative models more closely with human preferences.

8 Conclusions

This paper has provided a comprehensive review of research into cross-modal interaction between music and multimodal data. Based on an overview of music data formats, it systematically examines the development status of music-driven cross-modal interaction, music-oriented cross-modal interaction, and bidirectional cross-modal music interaction. In addition, we comprehensively summarize existing multimodal music data and the evaluation metrics used in the cross-modal interactions between music and other modalities. Furthermore, we discuss existing research limitations and future research directions. This survey aims to serve as a reference for researchers in the field, fostering further advances in cross-modal interaction between music and multimodal data, and driving innovation in the application of related technologies in the music industry and beyond.

Declaration

Declaration of competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author contributions

Sifei Li and Mining Tan were primarily responsible for drafting and revising the manuscript. Minyan Luo and Feier Shen wrote Sections 3.1 and 3.3, respectively. Zijiao Yin contributed to proofreading of the manuscript. Weiming Dong led the project. Fan Tang, Weiming Dong, and Changsheng Xu provided overall guidance and valuable suggestions throughout the work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62572458).

References

[1] Yang LC, Chou SY, Yang YH. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017, 324–331.

[2] Dong HW, Hsiao WY, Yang LC, Yang YH. Musegan: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018, volume 32.

[3] Dash A, Agres K. AI-Based Affective Music Generation Systems: A Review of Methods and Challenges. *ACM Computing Surveys*, 2024, 56(11): 1–34.

[4] Ji S, Yang X, Luo J. A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. *ACM Computing Surveys*, 2023, 56(1): 1–39.

[5] Wang L, Zhao Z, Liu H, Pang J, Qin Y, Wu Q. A Review of Intelligent Music Generation Systems. *Neural Computing and Applications*, 2024, 36(12): 6381–6401.

[6] Ji S, Luo J, Yang X. A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions. *arXiv preprint arXiv:2011.06801*, 2020.

[7] Li S, Ji S, Wang Z, Wu S, Yu J, Zhang K. A Survey on Music Generation from Single-Modal, Cross-Modal, and Multi-Modal Perspectives: Data, Methods, and Challenges. *arXiv preprint arXiv:2504.00837*, 2025.

[8] Ma Y, Øland A, Ragni A, Del Sette BM, Saitis C, Donahue C, Lin C, Plachouras C, Benetos E, Shatri E, et al.. Foundation Models for Music: A Survey. *arXiv preprint arXiv:2408.14340*, 2024.

[9] Li W, Cai Y, Wu Z, Zhang W, Chen Y, Qi R, Dong M, Chen P, Dong X, Shi F, et al.. A Survey of Foundation Models for Music Understanding. *arXiv preprint arXiv:2409.09601*, 2024.

[10] Ji S, Wu S, Wang Z, Li S, Zhang K. A Comprehensive Survey on Generative AI for Video-to-Music Generation. *arXiv preprint arXiv:2502.12489*, 2025.

[11] Wang Z, Bao C, Zhuo L, Han J, Yue Y, Tang Y, Huang VSJ, Liao Y. Vision-to-Music Generation: A Survey. *arXiv preprint arXiv:2503.21254*, 2025.

[12] Siedenburg K, Saitis C, McAdams S, Popper AN, Fay RR. Timbre: Acoustics, Perception, and Cognition, 2019, springer, Cham, Switzerland.

[13] Moore BC. Human and Machine Hearing: Extracting Meaning from Sound. *The Journal of the Acoustical Society of America*, 2018, 144(2): 567–567.

[14] Thickstun J, Harchaoui Z, Kakade S. Learning Features of Music From Scratch. In *International Conference on Learning Representations (ICLR)*, 2017.

[15] Bertin-Mahieux T, Ellis DPW, Whitman B, Lamere P. The Million Song Dataset. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2011, volume 2, 10.

[16] Wang Z, Chen K, Jiang J, Zhang Y, Xu M, Dai S, Bin G, Xia G. POP909: A Pop-Song Dataset for Music Arrangement Generation. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[17] Defferrard M, Benzi K, Vandergheynst P, Bresson X. FMA: A Dataset for Music Analysis. *International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[18] Tzanetakis G. Automatic Musical Genre Classification of Audio Signals. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2001.

[19] Raffel C. Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching. Ph.D. thesis, Columbia University, New York, NY, USA, 2016.

[20] Hawthorne C, Stasyuk A, Roberts A, Simon I, Huang CZA, Dieleman S, Elsen E, Engel J, Eck D. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In *International Conference on Learning Representations (ICLR)*, 2019.

[21] Foscarin F, McLeod A, Rigaux P, Jacquemard F, Sakai M. ASAP: A Dataset of Aligned Scores and Performances for Piano Transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2020, 534–541.

[22] Wang Y, Wang X, Zhu P, Wu J, Li H, Xue H, Zhang Y, Xie L, Bi M. Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis. In *INTERSPEECH*, 2022.

[23] Duan Z, Fang H, Li B, Sim KC, Wang Y. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2013, 1–9.

[24] Wu S, Li X, Yu F, Sun M. TunesFormer: Forming Irish Tunes with Control Codes by Bar Patching. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023, volume 3528 of *CEUR Workshop Proceedings*.

[25] Ferreira LN, Lelis LHS, Whitehead J. Computer-Generated Music for Tabletop Role-Playing Games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2020.

[26] Liu J, Dong Y, Cheng Z, Zhang X, Li X, Yu F, Sun M. Symphony Generation with Permutation Invariant Language Model. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[27] Donahue C, Mao HH, McAuley J. The NES Music Database: A Multi-Instrumental Dataset with Expressive Performance Attributes. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[28] Kong Q, Li B, Chen J, Wang Y. GiantMIDI-Piano: A Large-Scale MIDI Dataset for Classical Piano Music. *International Society for Music Information Retrieval Conference (ISMIR)*, 2022, 5(1).

[29] Ens J, Pasquier P. Building the MetaMIDI Dataset: Linking Symbolic and Audio Musical Data. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021, 182–188.

[30] Bogdanov D, Won M, Tovstogan P, Porter A, Serra X. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, 2019.

[31] Pegoraro Santana IA, Pinhelli F, Donini J, Catharin L, Man-

golin RB, da Costa YMeG, Delisandra Feltrim V, Domingues MA. Music4All: A New Music Database and Its Applications. In *International Conference on Systems, Signals and Image Processing*, 2020, 399–404.

[32] Lanzendörfer L, Grötschla F, Funke E, Wattenhofer R. DISCO-10M: A Large-Scale Music Dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, 36: 54451–54471.

[33] Wu S, Li X, Sun M. Chord-conditioned melody harmonization with controllable harmonicity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5.

[34] Hung HT, Ching J, Doh S, Kim N, Nam J, Yang YH. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[35] Liu J, Li C, Ren Y, Chen F, Zhao Z. Diffsinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. *AAAI Conference on Artificial Intelligence (AAAI)*, 2022, 36(10): 11020–11028.

[36] Huang R, Chen F, Ren Y, Liu J, Cui C, Zhao Z. Multi-Singer: Fast Multi-Singer Singing Voice Vocoder With A Large-Scale Corpus. In *ACM International Conference on Multimedia (ACM MM)*, 2021, 3945–3954.

[37] Zhang L, Li R, Wang S, Deng L, Liu J, Ren Y, He J, Huang R, Zhu J, Chen X, et al.. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, 35: 6914–6926.

[38] Mehri S, Kumar K, Gulrajani I, Kumar R, Jain S, Sotelo J, Courville A, Bengio Y. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. In *International Conference on Learning Representations (ICLR)*, 2017.

[39] Engel J, Resnick C, Roberts A, Dieleman S, Norouzi M, Eck D, Simonyan K. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *International Conference on Machine Learning (ICML)*, 2017, 1068–1077.

[40] Goel K, Gu A, Donahue C, Ré C. It's Raw! Audio Generation with State-Space Models. In *International Conference on Machine Learning (ICML)*, 2022, 7616–7633.

[41] Hawthorne C, Simon I, Roberts A, Zeghidour N, Gardner J, Manilow E, Engel J. Multi-instrument Music Synthesis with Spectrogram Diffusion. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[42] Gong Y, Lai CI, Chung YA, Glass J. SSAST: Self-Supervised Audio Spectrogram Transformer. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022, volume 36, 10699–10709.

[43] Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, 29: 3451–3460.

[44] Wang R, Hamdulla A. Fusion of MFCC and IMFCC for Whispered Speech Recognition. In *International Conference on Pattern Recognition and Machine Learning (PRML)*, 2022, 285–289.

[45] Yoo G, Hong S, Kim H. Emotion Recognition and Multi-Class Classification in Music with MFCC and Machine Learning. *International Journal on Advanced Science, Engineering and Information Technology*, 2024.

[46] Ma Y, Li Y, Zhang G, Lin C, Chen X, Ragni A, Yin H, Benetos E, Gyenge N, Liu R, et al.. On the Effectiveness of Speech Self-Supervised Learning for Music. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[47] Li Y, Yuan R, Zhang G, Ma Y, Chen X, Yin H, Xiao C, Lin C, Ragni A, Benetos E, et al.. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. In *International Conference on Learning Representations (ICLR)*, 2024.

[48] Huang S, Li Q, Anil C, Bao X, Oore S, Grosse RB. TimbreTron: A WaveNet (CycleGAN (CQT (Audio))) Pipeline for Musical Timbre Transfer. In *International Conference on Learning Representations (ICLR)*, 2018.

[49] Demerlé N, Esling P, Doras G, Genova D. Combining audio control and style transfer using latent diffusion. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[50] Lu P, Xu X, Kang C, Yu B, Xing C, Tan X, Bian J. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*, 2023.

[51] Benetos E, Dixon S, Duan Z, Ewert S. Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine*, 2018, 36(1): 20–30.

[52] Brunner G, Konrad A, Wang Y, Wattenhofer R. MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2018, 747–754.

[53] Huang YS, Yang YH. Pop Music Transformer: Beat-Based Modeling and Generation of Expressive Pop Piano Compositions. In *ACM International Conference on Multimedia (ACM MM)*, 2020, 1180–1188.

[54] Zhao K, Li S, Cai J, Wang H, Wang J. An Emotional Symbolic Music Generation System Based on LSTM Networks. In *IEEE Information Technology, Networking, Electronic and Automation Control Conference*, 2019, 2039–2043.

[55] Li S, Qin Y, Zheng M, Jin X, Liu Y. Diff-BGM: A Diffusion Model for Video Background Music Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, 27348–27357.

[56] Min L, Jiang J, Xia G, Zhao J. Polyffusion: A Diffusion Model for Polyphonic Score Generation With Internal and External

Controls. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023, 231–238.

[57] Wang Z, Zhang Y, Zhang Y, Jiang J, Yang R, Zhao J, Xia G. PianoTree VAE: Structured Representation Learning for Polyphonic Music. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[58] Wang Y, Wu S, Hu J, Du X, Peng Y, Huang Y, Fan S, Li X, Yu F, Sun M. NotaGen: Advancing Musicality in Symbolic Music Generation with Large Language Model Training Paradigms. *arXiv preprint arXiv:2502.18008*, 2025.

[59] Yuan R, Lin H, Wang Y, Tian Z, Wu S, Shen T, Zhang G, Wu Y, Liu C, Zhou Z, Xue L, Ma Z, Liu Q, Zheng T, Li Y, Ma Y, Liang Y, Chi X, Liu R, Wang Z, Lin C, Liu Q, Jiang T, Huang W, Chen W, Fu J, Benetos E, Xia G, Dannenberg RB, Xue W, Kang S, Guo Y. ChatMusician: Understanding and Generating Music Intrinsically with LLM. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, 6252–6271.

[60] Qu X, Bai Y, Ma Y, Zhou Z, Lo KM, Liu J, Yuan R, Min L, Liu X, Zhang T, et al.. Mupt: A generative symbolic music pretrained transformer. *arXiv preprint arXiv:2404.06393*, 2024.

[61] Jeong D, Kwon T, Kim Y, Nam J. Graph Neural Network for Music Score Data and Modeling Expressive Piano Performance. In *International Conference on Machine Learning (ICML)*, 2019, 3060–3070.

[62] Karystinaios E, Widmer G. Cadence Detection in Symbolic Classical Music using Graph Neural Networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[63] Karystinaios E, Foscarin F, Widmer G. Musical Voice Separation as Link Prediction: Modeling a Musical Perception Task as a Multi-Trajectory Tracking Problem. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, 3866–3874.

[64] Zhang H, Karystinaios E, Dixon S, Widmer G, Cancino-Chacón CE. Symbolic Music Representations for Classification Tasks: A Systematic Evaluation. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[65] Choi K, Fazekas G, McFee B, Cho K, Sandler M. Towards Music Captioning: Generating Music Playlist Descriptions. *arXiv preprint arXiv:1608.04868*, 2016.

[66] Manco I, Benetos E, Quinton E, Fazekas G. Muscaps: Generating Captions for Music Audio. *International Joint Conference on Neural Networks (IJCNN)*, 2021: 1–8.

[67] Xu X, Dinkel H, Wu M, Yu K. A CRNN-GRU Based Reinforcement Learning Approach to Audio Captioning. *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020: 225–229.

[68] Chen K, Wu Y, Wang Z, Zhang X, Nian F, Li S, Shao X. Audio Captioning Based on Transformer and Pre-Trained CNN. *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020: 21–25.

[69] Mei X, Liu X, Huang Q, Plumbley MD, Wang W. Audio Captioning Transformer. *arXiv preprint arXiv:2107.09817*, 2021.

[70] Deng Z, Ma Y, Liu Y, Guo R, Zhang G, Chen W, Huang W, Benetos E. MusiLingo: Bridging Music and Text with Pretrained Language Models for Music Captioning and Query Response. In *North American Chapter of the Association for Computational Linguistics*, 2024, 3643–3655.

[71] Liu S, Hussain AS, Sun C, Shan Y. 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, 286–290.

[72] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al.. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.

[73] Zhao M, Zhong Z, Mao Z, Yang S, Liao WH, Takahashi S, Wakaki H, Mitsufuji Y. OpenMU: Your Swiss Army Knife for Music Understanding. *arXiv preprint arXiv:2410.15573*, 2024.

[74] Doh S, Choi K, Lee J, Nam J. LP-MusicCaps: LLM-Based Pseudo Music Captioning. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[75] Manco I, Salamon J, Nieto O. Augment, Drop & Swap: Improving Diversity in LLM Captions for Efficient Music-Text Representation Learning. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024, 938–945.

[76] Srivatsan N, Chen K, Dubnov S, Berg-Kirkpatrick T. Retrieval Guided Music Captioning via Multimodal Prefixes. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023: 7762–7770.

[77] He Z, Hao W, Song X. RECAP: Retrieval Augmented Music Captioner. *arXiv preprint arXiv:2212.10901*, 2022.

[78] Fujihara H, Goto M, Ogata J, Okuno HG. LyricSynchronizer: Automatic Synchronization System Between Musical Audio Signals and Lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 2011, 5(6): 1252–1261.

[79] Mesaros A, Virtanen T. Automatic Alignment of Music Audio and Lyrics. In *International Conference on Digital Audio Effects*, 2008.

[80] Fujihara H, Goto M. A Music Information Retrieval System Based on Singing Voice Timbre. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2007, 467–470.

[81] Dzhambazov G. Knowledge-Based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals. Ph.D. thesis, Universitat Pompeu Fabra, 2017.

[82] Bittner RM, McFee B, Salamon J, Li P, Bello JP. Deep Salience Representations for F0 Estimation in Polyphonic Music. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017, 63–70.

[83] Vaglio A, Hennequin R, Moussallam M, Richard G, d'Alché Buc F. Multilingual Lyrics-to-Audio Alignment. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[84] Gupta C, Yılmaz E, Li H. Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 496–500.

[85] Stoller D, Durand S, Ewert S. End-to-End Lyrics Alignment for Polyphonic Music using an Audio-to-Character Recognition Model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, 181–185.

[86] Demirel E, Ahlbäck S, Dixon S. MSTRE-Net: Multistreaming Acoustic Modeling for Automatic Lyrics Transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[87] Gao X, Gupta C, Li H. Automatic Lyrics Transcription of Polyphonic Music with Lyrics-Chord Multi-Task Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, 30: 2280–2294.

[88] Zhuo L, Yuan R, Pan J, Ma Y, Li Y, Zhang G, Liu S, Dannenberg RB, Fu J, Lin C, et al.. LyricWhiz: Robust Multilingual Zero-Shot Lyrics Transcription by Whispering to ChatGPT. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[89] Ding S, Liu Z, Dong X, Zhang P, Qian R, He C, Lin D, Wang J. SongComposer: A Large Language Model for Lyric and Melody Composition in Song Generation. *arXiv preprint arXiv:2402.17645*, 2024.

[90] Huang J, Benetos E. Towards Building an End-to-End Multilingual Automatic Lyrics Transcription Model. In *European Signal Processing Conference*, 2024, 146–150.

[91] Wu S, He J, Yuan R, Wei H, Wei X, Lin C, Xu J, Lin J. SongTrans: An Unified Song Transcription and Alignment Method for Lyrics and Notes. *arXiv preprint arXiv:2409.14619*, 2024.

[92] Gu X, Ou L, Zeng W, Zhang J, Wong N, Wang Y. Automatic Lyric Transcription and Automatic Music Transcription from Multimodal Singing. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2024, 20(7): 1–29.

[93] Sterckx L, Naradowsky J, Byrne B, Demeester T, Develder C. Break it down for me: a study in automated lyric annotation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017, 1–7.

[94] Ogihara M, Manolovitz B, Shinohara VY, Ren G, Tavares TF. Graph-Based Representation, Analysis, and Interpretation of Popular Music Lyrics Using Semantic Embedding Features. *Handbook of Artificial Intelligence for Music*, 2021.

[95] Tadokoro T, Yamanishi R. Visualization of Collocations in Lyrics Through Query Interpretation by Large Language Model. *Procedia Computer Science*, 2024, 246: 3810–3819.

[96] Zhang Y, Jiang J, Xia G, Dixon S. Interpreting Song Lyrics with an Audio-Informed Pre-trained Language Model. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[97] Barbieri G, Pachet F, Roy P, Degli Esposti M. Markov Constraints for Generating Lyrics with Style. In *ECAI 2012*, 2012, 115–120.

[98] Malmi E, Takala P, Toivonen H, Raiko T, Gionis A. Dopelearning: A Computational Approach to Rap Lyrics Generation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, 195–204.

[99] Watanabe K, Matsubayashi Y, Fukayama S, Goto M, Inui K, Nakano T. A Melody-conditioned Lyrics Language Model. In *North American Chapter of the Association for Computational Linguistics*, 2018, 163–172.

[100] Chen Y, Lerch A. Melody-conditioned Lyrics Generation with SeqGANs. In *IEEE International Symposium on Multimedia (ISM)*, 2020, 189–196.

[101] Sheng Z, Song K, Tan X, Ren Y, Ye W, Zhang S, Qin T. Songmass: Automatic Song Writing with Pre-training and Alignment Constraint. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021, volume 35, 13798–13805.

[102] Ma X, Wang Y, Kan MY, Lee WS. AI-Lyricist: Generating Music and Vocabulary Constrained Lyrics. In *ACM International Conference on Multimedia (ACM MM)*, 2021, 1002–1011.

[103] Ou L, Ma X, Wang Y. Loaf-m2l: Joint Learning of Wording and Formatting for Singable Melody-to-Lyric Generation. *arXiv preprint arXiv:2307.02146*, 2023.

[104] Tian Y, Narayan-Chen A, Oraby S, Cervone A, Sigurdsson GA, Tao C, Zhao W, Chung T, Huang J, Peng N. Unsupervised Melody-to-Lyrics Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[105] Qian T, Lou F, Shi J, Wu Y, Guo S, Yin X, Jin Q. UniLG: A Unified Structure-aware Framework for Lyrics Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023, 983–1001.

[106] Zhang Z, Yu Y, Takasu A. Controllable Syllable-Level Lyrics Generation From Melody With Prior Attention. *IEEE Transactions on Multimedia*, 2024.

[107] Chen Y, Teufel S. Scansion-based Lyrics Generation. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024, 14370–14381.

[108] Liang Q, Ma X, Doshi-Velez F, Lim B, Wang Y. XAI-Lyricist: Improving the Singability of AI-Generated Lyrics with Prosody Explanations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, 7877–7885.

[109] Chae Y, Shin E, Suntae H, Paik S, Lee K. Song Form-aware Full-Song Text-to-Lyrics Generation with Multi-Level Granularity Syllable Count Control. *arXiv preprint arXiv:2411.13100*, 2024.

[110] Zhang R, Mao X, Le Li LJ, Chen L, Hu Z, Xi Y, Fan C, Huang M. Youling: an AI-Assisted Lyrics Creation System. *Em-*

pirical Methods in Natural Language Processing (EMNLP), 2020: 85.

[111] Zhang L, Zhang R, Mao X, Chang Y. QiuNiu: A Chinese Lyrics Generation System with Passage-level Input. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, 76–82.

[112] Huang Q, Jansen A, Zhang L, Ellis DPW, Saurous RA, Anderson J. Large-Scale Weakly-Supervised Content Embeddings for Music Recommendation and Tagging. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 8364–8368.

[113] Won M, Salamon J, Bryan NJ, Mysore GJ, Serra X. Emotion Embedding Spaces for Matching Music to Stories. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021, 777–785.

[114] Manco I, Benetos E, Quinton E, Fazekas G. Learning Music Audio Representations Via Weak Language Supervision. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: 456–460.

[115] Huang Q, Jansen A, Lee J, Ganti R, Li JY, Ellis DPW. MuLan: A Joint Embedding of Music Audio and Natural Language. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022, 559–566.

[116] Wu S, Wang Y, Yuan R, Guo Z, Tan X, Zhang G, Zhou M, Chen J, Mu X, Gao Y, et al.. CLaMP 2: Multimodal Music Information Retrieval Across 101 Languages Using Large Language Models. *arXiv preprint arXiv:2410.13267*, 2024.

[117] Favory X, Drossos K, Virtanen T, Serra X. COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations. In *ICML Workshop on Self-supervision in Audio and Speech*, 2020.

[118] Favory X, Drossos K, Virtanen T, Serra X. Learning Contextual Tag Embeddings for Cross-Modal Alignment of Audio and Tags. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 596–600.

[119] Wu Y, Chen K, Zhang T, Hui Y, Berg-Kirkpatrick T, Dubnov S. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: 1–5.

[120] Zhu G, Darefsky J, Duan Z. Cacophony: An Improved Contrastive Audio-Text Model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, 32: 4867–4879.

[121] Wu HH, Nieto O, Bello J, Salamon J. Audio-Text Models Do Not Yet Leverage Natural Language. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: 1–5.

[122] Huang PY, Sharma V, Xu H, Ryali C, Li Y, Li SW, Ghosh G, Malik J, Feichtenhofer C, et al.. Mavil: Masked Audio-Video Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 36.

[123] Kim J, Lee H, Rho K, Kim J, Chung JS. EquiAV: Leveraging Equivariance for Audio-Visual Contrastive Learning. In *International Conference on Machine Learning (ICML)*, 2024.

[124] Luo S, Yan C, Hu C, Zhao H. Diff-foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 36.

[125] Guzhov A, Raue F, Hees J, Dengel A. AudioClip: Extending Clip to Image, Text and Audio. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: 976–980.

[126] Wu HH, Seetharaman P, Kumar K, Bello J. Wav2CLIP: Learning Robust Audio Representations from Clip. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: 4563–4567.

[127] Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I. ImageBind: One Embedding Space to Bind Them All. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 15180–15190.

[128] Lei W, Ge Y, Zhang J, Sun D, Yi K, Shan Y, Shou MZ. VIT-LENS: Towards Omni-Modal Representations. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024: 26637–26647.

[129] Chen S, Li H, Wang Q, Zhao Z, Sun M, Zhu X, Liu J. Vast: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, 36: 72842–72866.

[130] Wu S, Zhancheng G, Yuan R, Jiang J, Doh S, Xia G, Nam J, Li X, Yu F, Sun M. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, 2605–2625.

[131] Gardner JP, Durand S, Stoller D, Bittner RM. LLark: A Multimodal Instruction-Following Language Model for Music. In *International Conference on Machine Learning (ICML)*, 2024.

[132] Weck B, Manco I, Benetos E, Quinton E, Fazekas G, Bogdanov D. MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[133] Liu S, Hussain AS, Wu Q, Sun C, Shan Y. MuMu-LLaMA: Multi-Modal Music Understanding and Generation via Large Language Models. *arXiv preprint arXiv:2412.06660*, 2024.

[134] Chu Y, Xu J, Zhou X, Yang Q, Zhang S, Yan Z, Zhou C, Zhou J. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint arXiv:2311.07919*, 2023.

[135] Gong Y, Luo H, Liu AH, Karlinsky L, Glass JR. Listen, Think, and Understand. In *International Conference on Learning Representations (ICLR)*, 2024.

[136] Chu Y, Xu J, Yang Q, Wei H, Wei X, Guo Z, Leng Y, Lv Y, He J, Lin J, et al.. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*, 2024.

[137] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux MA, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

[138] Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J, Rombach R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

[139] Davis A, Agrawala M. Visual Rhythm and Beat. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[140] Wang J, Fang Z, Zhao H. AlignNet: A Unifying Approach to Audio-Visual Alignment. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020: 3298–3306.

[141] Yu J, Pu J, Cheng Y, Feng R, Shan Y. Learning Music-Dance Representations Through Explicit-Implicit Rhythm Synchronization. *IEEE Transactions on Multimedia*, 2022, 26: 8454–8463.

[142] Jiang D, Zhang Y, He S, Ming A. M2Beats: When Motion Meets Beats in Short-Form Videos. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[143] Ye Z, Wu H, Jia J, Bu Y, Chen W, Meng F, Wang Y. ChoreoNet: Towards Music to Dance Synthesis with Choreographic Action Unit. *ACM International Conference on Multimedia (ACM MM)*, 2020.

[144] Ferreira JP, Coutinho TM, Gomes TL, Neto JF, Azevedo R, Martins R, Nascimento ER. Learning to Dance: A Graph Convolutional Adversarial Network to Generate Realistic Dance Motions from Audio. *Computers & Graphics*, 2021, 94: 11–21.

[145] Ren X, Li H, Huang Z, Chen Q. Self-supervised Dance Video Synthesis Conditioned on Music. *ACM International Conference on Multimedia (ACM MM)*, 2020.

[146] Alemi O, Françoise J, Pasquier P. GrooveNet: Real-Time Music-Driven Dance Movement Generation using Artificial Neural Networks. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

[147] Tang T, Jia J, Mao H. Dance with Melody: An LSTM-autoencoder Approach to Music-oriented Dance Synthesis. *ACM International Conference on Multimedia (ACM MM)*, 2018: 1598–1606.

[148] Yalta N, Watanabe S, Nakadai K, Ogata T. Weakly-Supervised Deep Recurrent Neural Networks for Basic Dance Step Generation. *International Joint Conference on Neural Networks (IJCNN)*, 2019: 1–8.

[149] Zhuang W, Wang C, Chai J, Wang Y, Shao M, Xia S. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022, 18(2): 1–21.

[150] Lee HY, Yang X, Liu MY, Wang TC, Lu YD, Yang MH, Kautz J. Dancing to music. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, 32.

[151] Kim J, Oh H, Kim S, Tong H, Lee S. A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 3490–3500.

[152] Li R, Yang S, Ross DA, Kanazawa A. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[153] Li S, Yu W, Gu T, Lin C, Wang Q, Qian C, Loy CC, Liu Z. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[154] Li B, Zhao Y, Shi Z, Sheng L. DanceFormer: Music Conditioned 3D Dance Generation with Parametric Motion Transformer. *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[155] Chen K, Tan Z, Lei J, Zhang SH, Guo YC, Zhang W, Hu SM. ChoreoMaster: Choreography-Oriented Music-Driven Dance Synthesis. *ACM Transactions on Graphics*, 2021, 40(4): 1–13, doi:10.1145/3450626.3459932.

[156] Zhuang H, Lei S, Xiao L, Li W, Chen L, Yang S, Wu Z, Kang S, Meng H. Gtn-bailando: Genre consistent long-term 3d dance generation based on pre-trained genre token network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5.

[157] Han B, Zhang T, Ling Z, Ren Y, Yin X, Han F. EnchantDance: Unveiling the Potential of Music-Driven Dance Movement. *arXiv preprint arXiv:2312.15946*, 2023.

[158] Huang Q, He X, Tang B, Zhuang HW, Chen L, Gao S, Wu Z, Huang H, Meng HM. Enhancing Expressiveness in Dance Generation Via Integrating Frequency and Music Style Information. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: 8185–8189.

[159] Gong K, Lian D, Chang H, Guo C, Jiang Z, Zuo X, Mi MB, Wang X. TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 9942–9952.

[160] Qi Q, Zhuo L, Zhang A, Liao Y, Fang F, Liu S, Yan S. DiffDance: Cascaded Human Motion Diffusion Model for Dance Generation. In *ACM International Conference on Multimedia (ACM MM)*, 2023, 1374–1382.

[161] Tseng JH, Castellon R, Liu CK. EDGE: Editable Dance

Generation From Music. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: 448–458.

[162] Dhariwal P, Jun H, Payne C, Kim JW, Radford A, Sutskever I. Jukebox: A Generative Model for Music. *arXiv preprint arXiv:2005.00341*, 2020.

[163] Wang H, Zhu Y, Geng X. Flexible Music-Conditioned Dance Generation with Style Description Prompts. *arXiv preprint arXiv:2406.07871*, 2024.

[164] Yang S, Yang Z, Wang Z. LongDanceDiff: Long-term Dance Generation with Conditional Diffusion Model. *arXiv preprint arXiv:2308.11945*, 2023.

[165] Zhang C, Tang Y, Zhang N, Lin RS, Han M, Xiao J, Wang S. Bidirectional Autoregessive Diffusion Model for Dance Generation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024: 687–696.

[166] Yin W, Zhao X, Yu Y, Yin H, Kragic D, Björkman M. LM2D: Lyrics- and Music-Driven Dance Synthesis. *arXiv preprint arXiv:2403.09407*, 2024.

[167] Li R, Zhang Y, Zhang Y, Zhang H, Guo J, Zhang Y, Liu Y, Li X. LODGE: A Coarse to Fine Diffusion Network for Long Dance Generation Guided by the Characteristic Dance Primitives. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024: 1524–1534.

[168] Marchellus M, Park IK. M2C: Concise Music Representation for 3D Dance Generation. *IEEE/CVF International Conference on Computer Vision Workshops*, 2023: 3118–3127.

[169] Yao S, Sun M, Li B, Yang F, Wang J, Zhang R. Dance with You: The Diversity Controllable Dancer Generation via Diffusion Models. In *ACM International Conference on Multimedia (ACM MM)*, 2023, 8504–8514.

[170] Siyao L, Gu T, Yang Z, Lin Z, Liu Z, Ding H, Yang L, Loy CC. Duolando: Follower GPT with Off-Policy Reinforcement Learning for Dance Accompaniment. In *International Conference on Learning Representations (ICLR)*, 2024.

[171] Li R, Zhang Y, Zhang Y, Zhang Y, Su M, Guo J, Liu Z, Liu Y, Li X. InterDance: Reactive 3D Dance Generation with Realistic Duet Interactions. *arXiv preprint arXiv:2412.16982*, 2024.

[172] Le N, Pham T, Do T, Tjiputra E, Tran QD, Nguyen A. Music-Driven Group Choreography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[173] Le N, Do T, Do K, Nguyen H, Tjiputra E, Tran QD, Nguyen A. Controllable Group Choreography Using Contrastive Diffusion. *ACM Transactions on Graphics*, 2023, 42(6): 1–14.

[174] Wang Z, Jia J, Wu H, Xing J, Cai J, Meng F, Chen G, Wang Y. GroupDancer: Music to Multi-People Dance Synthesis with Style Collaboration. In *ACM International Conference on Multimedia (ACM MM)*, 2022.

[175] Le N, Do K, Bui X, Do T, Tjiputra E, Tran QD, Nguyen A. Scalable Group Choreography via Variational Phase Manifold Learning. In *European Conference on Computer Vision (ECCV)*, 2024, 293–311.

[176] Liao Z, Yu Y, Gong B, Cheng L. Audeosynth. *ACM Transactions on Graphics*, 2015, 34: 1–10.

[177] Shin KH, Kim HR, Lee IK. Automated Music Video Generation Using Emotion Synchronization. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2016, 2594–2597.

[178] Fan J, Li W, Bizzocchi J, Bizzocchi J, Pasquier P. DJ-MVP: An Automatic Music Video Producer. *International Conference on Advances in Computer Entertainment Technology*, 2016.

[179] Gross S, Wei X, Zhu J. Automatic Realistic Music Video Generation from Segments of Youtube Videos. *arXiv preprint arXiv:1905.12245*, 2019.

[180] Zhang H, Zhang Kj, Cao YA, Zheng J, Huang X, Yang Cy, Sun L. A Music-Driven System for Generating Apparel Display Video. *Multimedia Tools and Applications*, 2019, 79: 5649–5670.

[181] Jang J, Shin S, Kim Y. Music2Video: Automatic Generation of Music Video with Fusion of Audio and Text.(2022). *arXiv preprint arXiv:2201.03809*, 2022.

[182] Iwase S, Kato T, Yamaguchi S, Tsuchiya Y, Morishima S. Song2Face: Synthesizing Singing Facial Animation from Audio. *SIGGRAPH Asia Technical Communications*, 2020.

[183] Liu P, Deng W, Li H, Wang J, Zheng Y, Ding Y, Guo X, Zeng M. MusicFace: Music-driven Expressive Singing Face Synthesis. *Computational Visual Media*, 2023, 10: 119–136.

[184] Li Y, Zhou Z, Wang Z, Xue W, Luo W, Guo YT. SINGER: Vivid Audio-driven Singing Video Generation with Multi-scale Spectral Diffusion Model. *arXiv preprint arXiv:2412.03430*, 2024.

[185] Liu F, Chen D, Zhou RZ, Yang S, Xu F. Self-Supervised Music Motion Synchronization Learning for Music-Driven Conducting Motion Generation. *Journal of Computer Science and Technology*, 2022, 37: 539–558.

[186] Liu JW, Lin HY, Huang YF, Kao HK, Su L. Body Movement Generation for Expressive Violin Performance Applying Neural Networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: 3787–3791.

[187] Kao HK, Su L. Temporally Guided Music-to-Body-Movement Generation. *ACM International Conference on Multimedia (ACM MM)*, 2020: 147–155.

[188] Li B, Maezawa A, Duan Z. Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDI Performance. *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[189] Bogaers A, Yumak Z, Volk A. Music-Driven Animation Generation of Expressive Musical Gestures. *International Conference on Multimodal Interaction (ICMI)*, 2020.

[190] Yamamoto K, Ueda E, Suenaga T, Takemura K, Takamatsu J, Ogasawara T. Generating Natural Hand Motion in Playing a Piano. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010: 3513–3518.

[191] Shlizerman E, Dery L, Schoen H, Kemelmacher-Shlizerman I. Audio to Body Dynamics. *IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition (CVPR), 2018: 7574–7583.

[192] Chen L, Srivastava S, Duan Z, Xu C. Deep Cross-Modal Audio-Visual Generation. *ACM International Conference on Multimedia (ACM MM)*, 2017.

[193] Chatterjee M, Cherian A. Sound2sight: Generating visual dynamics from sound and context. In *European Conference on Computer Vision (ECCV)*, 2020, 701–719.

[194] Zhu H, Li Y, Zhu F, Zheng A, He R. Let's Play Music: Audio-Driven Performance Video Generation. *International Conference on Pattern Recognition*, 2020: 3574–3581.

[195] Yu Y, Tang S, Raposo F, Chen L. Deep Cross-Modal Correlation Learning for Audio and Lyrics in Music Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2017, 15: 1–16.

[196] Won M, Oramas S, Nieto O, Gouyon F, Serra X. Multimodal Metric Learning for Tag-Based Music Retrieval. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: 591–595.

[197] Doh S, Won M, Choi K, Nam J. Toward Universal Text-To-Music Retrieval. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: 1–5.

[198] Doh S, Lee J, Jeong D, Nam J. Musical Word Embedding for Music Tagging and Retrieval. *arXiv preprint arXiv:2404.13569*, 2024.

[199] Doh S, Lee M, Jeong D, Nam J. Enriching Music Descriptions with A Finetuned-LLM and Metadata for Text-to-Music Retrieval. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: 826–830.

[200] Wu S, Yu D, Tan X, Sun M. CLaMP: Contrastive Language-Music Pre-Training for Cross-Modal Symbolic Music Information Retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[201] Roose K. An AI-Generated Picture Won an Art Prize. Artists Aren't Happy. *The New York Times*, 2022.

[202] Wang Y, Yang W, Dai Z, Zhang Y, Zhao K, Wang H. MeloTrans: A Text to Symbolic Music Generation Model Following Human Composition Habit. *arXiv preprint arXiv:2410.13419*, 2024.

[203] Agostinelli A, Denk TI, Borsos Z, Engel J, Verzetti M, Caillon A, Huang Q, Jansen A, Roberts A, Tagliasacchi M, Sharifi M, Zeghidour N, Frank C. MusicLM: Generating Music From Text. *arXiv preprint arXiv:2301.11325*, 2023.

[204] Borsos Z, Marinier R, Vincent D, Kharitonov E, Pietquin O, Sharifi M, Roblek D, Teboul O, Grangier D, Tagliasacchi M, Zeghidour N. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, 31: 2523–2533.

[205] Copet J, Kreuk F, Gat I, Remez T, Kant D, Synnaeve G, Adi Y, Défossez A. Simple and Controllable Music Generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 36.

[206] Kreuk F, Synnaeve G, Polyak A, Singer U, Défossez A, Copet J, Parikh D, Taigman Y, Adi Y. AudioGen: Textually Guided Audio Generation. In *International Conference on Learning Representations (ICLR)*, 2023.

[207] Flores García H, Seetharaman P, Kumar R, Pardo B. VampNet: Music Generation via Masked Acoustic Token Modeling. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023, 359–366.

[208] Parker J, Spijkervet J, Kosta K, Yesiler F, Kuznetsov B, Wang JC, Avent M, Chen J, Le D. STEMGEN: A Music Generation Model That Listens. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: 1116–1120.

[209] Ziv A, Gat I, Le Lan G, Remez T, Kreuk F, Copet J, Défossez A, Synnaeve G, Adi Y. Masked Audio Generation using a Single Non-Autoregressive Transformer. In *International Conference on Learning Representations (ICLR)*, 2024.

[210] Cideron G, Girgin S, Verzetti M, Vincent D, Kastelic M, Borsos Z, McWilliams B, Ungureanu V, Bachem O, Pietquin O, Geist M, Hussenot L, Zeghidour N, Agostinelli A. MusicRL: Aligning Music Generation to Human Preferences. In *International Conference on Machine Learning (ICML)*, 2024.

[211] Forsgren S, Martiros H. Riffusion - Stable Diffusion for Real-Time Music Generation. *Online Publication*, 2022.

[212] Huang R, Huang J, Yang D, Ren Y, Liu L, Li M, Ye Z, Liu J, Yin X, Zhao Z. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2023, volume 202 of *Proceedings of Machine Learning Research*, 13916–13932.

[213] Liu H, Chen Z, Yuan Y, Mei X, Liu X, Mandic D, Wang W, Plumbley MD. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2023, 21450–21474.

[214] Liu H, Tian Q, Yuan Y, Liu X, Mei X, Kong Q, Wang Y, Wang W, Wang Y, Plumbley M. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, 32: 2871–2883.

[215] Ghosal D, Majumder N, Mehrish A, Poria S. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[216] Majumder N, Hung CY, Ghosal D, Hsu WN, Mihalcea R, Poria S. Tango 2: Aligning Diffusion-Based Text-to-Audio Generations Through Direct Preference Optimization. In *ACM International Conference on Multimedia (ACM MM)*, 2024, 564–572.

[217] Nistal J, Pasini M, Aouameur C, Grachten M, Lattner S. Diff-A-Riff: Musical Accompaniment Co-creation via Latent Diffusion Models. *arXiv preprint arXiv:2406.08384*, 2024.

[218] Xue J, Deng Y, Gao Y, Li Y. Auffusion: Leveraging the Power of Diffusion and Large Language Models for Text-to-Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[219] Chen K, Wu Y, Liu H, Nezhurina M, Berg-Kirkpatrick T, Dubnov S. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: 1206–1210.

[220] Huang Q, Park DS, Wang T, Denk TI, Ly A, Chen N, Zhang Z, Zhang Z, Yu J, Frank C, Engel J, Le QV, Chan W, Han W. Noise2Music: Text-conditioned Music Generation with Diffusion Models. *arXiv preprint arXiv:2302.03917*, 2023.

[221] Li P, Chen BY, Yao Y, Wang Y, Wang A, Wang A. JEN-1: Text-Guided Universal Music Generation with Omnidirectional Diffusion Models. *2024 IEEE Conference on Artificial Intelligence (CAI)*, 2023: 762–769.

[222] Evans Z, Carr C, Taylor J, Hawley SH, Pons J. Fast timing-conditioned latent audio diffusion. In *International Conference on Machine Learning (ICML)*, 2024.

[223] Lam MW, Tian Q, Li T, Yin Z, Feng S, Tu M, Ji Y, Xia R, Ma M, Song X, et al.. Efficient Neural Music Generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 36.

[224] Peebles W, Xie S. Scalable Diffusion Models with Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 4195–4205.

[225] Evans Z, Parker JD, Carr C, Zukowski Z, Taylor J, Pons J. Stable Audio Open. *arXiv preprint arXiv:2407.14358*, 2024.

[226] Prajwal KR, Shi B, Le M, Vyas A, Tjandra A, Luthra M, Guo B, Wang H, Afouras T, Kant D, Hsu WN. MusicFlow: Cascaded Flow Matching for Text Guided Music Generation. In *International Conference on Machine Learning (ICML)*, 2024.

[227] Fei Z, Fan M, Yu C, Huang J. FLUX that Plays Music. *arXiv preprint arXiv:2409.00587*, 2024.

[228] Bai Y, Dang T, Tran DN, Koishida K, Sojoudi S. ConsistencyTTA: Accelerating Diffusion-Based Text-to-Audio Generation with Consistency Distillation. *INTERSPEECH*, 2024.

[229] Novack Z, Zhu G, Casebeer J, McAuley J, Berg-Kirkpatrick T, Bryan NJ. Presto! Distilling Steps and Layers for Accelerating Music Generation. *arXiv preprint arXiv:2410.05167*, 2024.

[230] Novack Z, McAuley J, Berg-Kirkpatrick T, Bryan NJ. DITTO-2: Distilled Diffusion Inference-Time T-Optimization for Music Generation. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[231] Bai Y, Chen H, Chen J, Chen Z, Deng Y, Dong X, Hantrakul L, Hao W, Huang Q, Huang Z, Jia D, La F, Le D, Li B, Li C, Li H, Li X, Liu S, Lu W, Lu Y, Shaw A, Spijkervet J, Sun Y, Wang B, Wang JC, Wang Y, Wang Y, Xu L, Yang Y, Yao C, Zhang S, Zhang Y, Zhang Y, Zhao H, Zhao Z, Zhong D, Zhou S, Zou P. Seed-Music: A Unified Framework for High Quality and Controlled Music Generation. *arXiv preprint arXiv:2409.09214*, 2024.

[232] Hong Z, Huang R, Cheng X, Wang Y, Li R, You F, Zhao Z, Zhang Z. Text-to-song: Towards controllable music generation incorporating vocals and accompaniment. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, 6248–6261.

[233] Li R, Hong Z, Wang Y, Zhang L, Huang R, Zheng S, Zhao Z. Accompanied Singing Voice Synthesis with Fully Text-controlled Melody. *arXiv preprint arXiv:2407.02049*, 2024.

[234] Lei S, Zhou Y, Tang B, Lam MW, Liu H, Wu J, Kang S, Wu Z, Meng H, et al.. Songcreator: Lyrics-based universal song generation. *Advances in Neural Information Processing Systems*, 2024, 37: 80107–80140.

[235] Yuan R, Lin H, Guo S, Zhang G, Pan J, Zang Y, Liu H, Liang Y, Ma W, Du X, et al.. YuE: Scaling Open Foundation Models for Long-Form Music Generation. *arXiv preprint arXiv:2503.08638*, 2025.

[236] Liu Z, Ding S, Zhang Z, Dong X, Zhang P, Zang Y, Cao Y, Lin D, Wang J. SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation. In *International Conference on Machine Learning (ICML)*, 2025.

[237] Lei S, Xu Y, Lin Z, Zhang H, Tan W, Chen H, Yu J, Zhang Y, Yang C, Zhu H, et al.. LeVo: High-Quality Song Generation with Multi-Preference Alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[238] Lam MW, Xing Y, You W, Wu J, Yin Z, Jiang F, Liu H, Liu F, Li X, Lu WT, et al.. Analyzable Chain-of-Musical-Thought Prompting for High-Fidelity Music Generation. *arXiv preprint arXiv:2503.19611*, 2025.

[239] Ning Z, Chen H, Jiang Y, Hao C, Ma G, Wang S, Yao J, Xie L. DiffRhythm: Blazingly Fast and Embarrassingly Simple End-to-End Full-Length Song Generation with Latent Diffusion. *arXiv preprint arXiv:2503.01183*, 2025.

[240] Gong J, Zhao S, Wang S, Xu S, Guo J. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 2025.

[241] Melechovsky J, Guo Z, Ghosal D, Majumder N, Herremans D, Poria S. Mustango: Toward Controllable Text-to-Music Generation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, 8286–8309.

[242] Shu Y, Xu H, Zhou Z, Hengel Avd, Liu L. MuseBarControl: Enhancing Fine-Grained Control in Symbolic Music Generation through Pre-Training and Counterfactual Loss. *arXiv preprint arXiv:2407.04331*, 2024.

[243] Wu SL, Donahue C, Watanabe S, Bryan NJ. Music ControlNet: Multiple Time-Varying Controls for Music Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, 32: 2692–2703.

[244] Zhang L, Rao A, Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 3836–3847.

[245] Tal O, Ziv A, Gat I, Kreuk F, Adi Y. Joint Audio and Symbolic Conditioning for Temporally Controlled Text-to-Music Generation. *arXiv preprint arXiv:2406.10970*, 2024.

[246] Lan YH, Hsiao WY, Cheng HC, Yang YH. Musicongen:

Rhythm and Chord Control for Transformer-Based Text-to-Music Generation. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[247] Li S, Li Y, Wang Z, Zhang Y, Wu F, Deussen O, Lee TY, Dong W. SongEcho: Cover Song Generation via Instance-Adaptive Element-wise Linear Modulation. In *International Conference on Learning Representations (ICLR)*, 2026.

[248] Novack Z, McAuley JJ, Berg-Kirkpatrick T, Bryan NJ. DITTO: Diffusion Inference-Time T-Optimization for Music Generation. In *International Conference on Machine Learning (ICML)*, 2024.

[249] Hsia CC, Lai KH, Chen Y, Wang CJ, Tsai MF. Representation Learning for Image-Based Music Recommendation. *arXiv preprint arXiv:1808.09198*, 2018.

[250] Sasaki S, Hirai T, Ohya H, Morishima S. Affective music recommendation system using input images. In *ACM SIGGRAPH 2013 Posters*, 2013, 1–1.

[251] Zhao S, Li Y, Yao X, Nie W, Xu P, Yang J, Keutzer K. Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space. In *ACM International Conference on Multimedia (ACM MM)*, 2020, 2945–2954.

[252] Baijal A, Agarwal V, Hyun D. Analyzing Images for Music Recommendation. In *IEEE International Conference on Consumer Electronics*, 2021, 1–6.

[253] Bao X, Li JY, Wan ZY, Su K, Denk T, Lee J, Kuzmin D, Sha F. Diff4Steer: Steerable Diffusion Prior for Generative Music Retrieval with Semantic Guidance. *arXiv preprint arXiv:2412.04746*, 2024.

[254] Wang X, Gao Z, Qian H, Xu Y. Jukepix: A Cross-Modality Approach to Transform Paintings into Music Segments. In *IEEE International Conference on Robotics and Biomimetics*, 2018, 456–461.

[255] Mendis N, Bandara R. Show and Listen: Generating Music Based on Images using Machine Learning. In *International Conference on Advances in ICT for Emerging Regions (ICTER)*, 2022, 25–30.

[256] Wang Y, Chen M, Li X. Continuous Emotion-Based Image-to-Music Generation. *IEEE Transactions on Multimedia*, 2023.

[257] Kundu S, Singh S, Iwahori Y. Emotion-Guided Image to Music Generation. *arXiv preprint arXiv:2410.22299*, 2024.

[258] Hisariya T, Zhang H, Liang J. Bridging Paintings and Music–Exploring Emotion Based Music Generation through Paintings. *arXiv preprint arXiv:2409.07827*, 2024.

[259] Chowdhury S, Nag S, Joseph KJ, Srinivasan BV, Manocha D. MELFuSION: Synthesizing Music from Image and Language Cues Using Diffusion Models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024: 26816–26825.

[260] Zhao E, Zheng J, Cao M. CPTGZ: Generating Chinese Guzheng Music from Chinese Paintings Based on Diffusion Model. *IEEE Access*, 2024.

[261] Sasaki S, Hirai T, Ohya H, Morishima S. Affective Music Recommendation System Based on the Mood of Input Video. In *International Conference on MultiMedia Modeling*, 2015, 299–302.

[262] Hong S, Im W, Yang H. CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint. *ACM International Conference on Multimedia Retrieval (ICMR)*, 2018.

[263] Surís D, Duarte A, Salvador A, Torres J, Giró-i Nieto X. Cross-Modal Embeddings for Video and Audio Retrieval. In *ECCV Workshops*, 2018, 0–0.

[264] Li B, Kumar A. Query by Video: Cross-Modal Music Retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2019, 604–611.

[265] Prétet L, Richard G, Peeters G. Cross-Modal Music-Video Recommendation: A Study of Design Choices. *International Joint Conference on Neural Networks (IJCNN)*, 2021: 1–9.

[266] Yi J, Zhu Y, Xie J, Chen Z. Cross-Modal Variational Auto-Encoder for Content-Based Micro-Video Background Music Recommendation. *IEEE Transactions on Multimedia*, 2021, 25: 515–528.

[267] Surís D, Vondrick C, Russell BC, Salamon J. It's Time for Artistic Correspondence in Music and Video. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: 10554–10564.

[268] Prétet L, Richard G, Souchier C, Peeters G. Video-to-Music Recommendation Using Temporal Alignment of Segments. *IEEE Transactions on Multimedia*, 2023, 25: 2898–2911.

[269] Cheng X, Zhu Z, Li H, Li Y, Zou Y. SSVMR: Saliency-Based Self-Training for Video-Music Retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5.

[270] Yi J, Chen Z. Deconfounded Cross-Modal Matching for Content-Based Micro-Video Background Music Recommendation. *ACM Transactions on Intelligent Systems and Technology*, 2024, 15: 1–25.

[271] McKee D, Salamon J, Sivic J, Russell B. Language-Guided Music Recommendation for Video via Prompt Analogies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 14784–14793.

[272] Koepke AS, Wiles O, Moses Y, Zisserman A. Sight to Sound: An End-to-End Approach for Visual Piano Transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 1838–1842.

[273] Su K, Liu X, Shlizerman E. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, 33: 3325–3337.

[274] Su K, Liu X, Shlizerman E. Multi-Instrumentalist Net: Unsupervised Generation of Music from Body Movements. *arXiv preprint arXiv:2012.03478*, 2020.

[275] Gan C, Huang D, Chen P, Tenenbaum JB, Torralba A. Foley Music: Learning to Generate Music from Videos. In *European Conference on Computer Vision (ECCV)*, 2020, 758–775.

[276] Wang Y, Guo W, Huang R, Huang J, Wang Z, You F, Li R,

Zhao Z. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 37: 128118–128138.

[277] Wang H, Ma J, Pascual S, Cartwright R, Cai W. V2a-mapper: A Lightweight Solution for Vision-to-Audio Generation by Connecting Foundation Models. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024, volume 38, 15492–15501.

[278] Han B, Li Y, Shen Y, Ren Y, Han F. Dance2MIDI: Dance-driven multi-instrument music generation. *Computational Visual Media*, 2024, 10(4): 791–802.

[279] Liang X, Li W, Huang L, Gao C. DanceComposer: Dance-to-Music Generation Using a Progressive Conditional Music Generator. *IEEE Transactions on Multimedia*, 2024, 26: 10237–10250.

[280] Zhu Y, Wu Y, Olszewski K, Ren J, Tulyakov S, Yan Y. Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation. In *International Conference on Learning Representations (ICLR)*, 2023.

[281] Yu J, Wang Y, Chen X, Sun X, Qiao Y. Long-term Rhythmic Video Soundtracker. In *International Conference on Machine Learning (ICML)*, 2023, 40339–40353.

[282] Tan V, Nam J, Nam J, Noh Jy. Motion to Dance Music Generation using Latent Diffusion Model. *SIGGRAPH Asia Technical Communications*, 2023.

[283] Li S, Dong W, Zhang Y, Tang F, Ma C, Deussen O, Lee TY, Xu C. Dance-to-Music Generation with Encoder-based Textual Inversion. In *ACM SIGGRAPH Asia Conference*, 2024, SA '24.

[284] Wang J, Liu L, Wang J. MotionComposer: Enhancing Rhythmic Music Generation with Adaptive Retrieval Reference. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, 1–5.

[285] Di S, Jiang S, Liu Z, Wang L, Zhu Z, He H, Liu S, Yan. Video Background Music Generation with Controllable Music Transformer. *ACM International Conference on Multimedia (ACM MM)*, 2021.

[286] Kang J, Poria S, Herremans D. Video2music: Suitable Music Generation from Videos Using an Affective Multimodal Transformer Model. *Expert Systems with Applications*, 2024, 249: 123640.

[287] Zhuo L, Wang Z, Wang B, Liao Y, Peng S, Bao C, Lu M, Li X, Liu S. Video Background Music Generation: Dataset, Method and Evaluation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023: 15591–15601.

[288] Su K, Li JY, Huang Q, Kuzmin D, Lee J, Donahue C, Sha F, Jansen A, Wang Y, Verzetti M, Denk TI. V2Meow: Meowing to the Visual Beat via Video-to-Music Generation. *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[289] Tian Z, Liu Z, Yuan R, Pan J, Huang X, Liu Qf, Tan X, Chen Q, Xue W, Guo YT. VidMuse: A Simple Video-to-Music Generation Framework with Long-Short-Term Modeling. *arXiv preprint arXiv:2406.04321*, 2024.

[290] Kai WH, Xing KX. Video-driven musical composition using large language model with memory-augmented state space. *The Visual Computer*, 2024: 1–13.

[291] Sharma M, Haseeb MT, Xia G, Tsuruoka Y. M2m-gen: A multimodal framework for automated background music generation in japanese manga using large language models. *arXiv preprint arXiv:2410.09928*, 2024.

[292] Tong X, Chen S, Yu P, Liu N, Qv H, Ma T, Zheng B, Yu F, Zhu SC. Video Echoed in Harmony: Learning and Sampling Video-Integrated Chord Progression Sequences for Controllable Video Background Music Generation. *IEEE Transactions on Computational Social Systems*, 2024.

[293] Zhang L, Fuentes M. SONIQUE: Video Background Music Generation Using Unpaired Audio-Visual Data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, 1–5.

[294] Lin YB, Tian Y, Yang L, Bertasius G, Wang H. VMAS: Video-to-Music Generation via Semantic Alignment in Web Music Videos. *arXiv preprint arXiv:2409.07450*, 2024.

[295] Li R, Zheng S, Cheng X, Zhang Z, Ji S, Zhao Z. MuVi: Video-to-Music Generation with Semantic Alignment and Rhythmic Synchronization. *arXiv preprint arXiv:2410.12957*, 2024.

[296] Tong X, Zhu Y, Chen J, Zhan C, Wang T, Zhang S, Liu N, Ge T, Xu D, Jin X, et al.. Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation. *arXiv preprint arXiv:2511.09585*, 2025.

[297] Li S, Yang B, Yin C, Sun C, Zhang Y, Dong W, Li C. VidMusician: Video-to-Music Generation with Semantic-Rhythmic Alignment via Hierarchical Visual Features. *arXiv preprint arXiv:2412.06296*, 2024.

[298] Zuo H, You W, Wu J, Ren S, Chen P, Zhou M, Lu Y, Sun L. Gvmgen: A general video-to-music generation model with hierarchical attentions. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025, 23099–23107.

[299] Xie Z, He Q, Zhu Y, He Q, Li M. FilmComposer: LLM-Driven Music Production for Silent Film Clips. *arXiv preprint arXiv:2503.08147*, 2025.

[300] Qi F, Ni L, Xu C. Harmonizing Pixels and Melodies: Maestro-Guided Film Score Generation and Composition Style Transfer. *arXiv preprint arXiv:2411.07539*, 2024.

[301] Yin Z, Li S, Dong W. Video2Song: Video-to-Song Generation via Retrieval-Augmented Prompt Generation. In *Proceedings of the SIGGRAPH Asia 2025 Posters*, 2025, 15:1–15:2.

[302] Kong W, Tian Q, Zhang Z, Min R, Dai Z, Zhou J, Xiong J, Li X, Wu B, Zhang J, et al.. Hunyuanvideo: A Systematic Framework for Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024.

[303] Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, Vyas A, Shi B, Ma CY, Chuang CY, Yan D, Choudhary D, Wang D, Sethi G, Pang G, Ma H, Misra I, Hou J, Wang J, Jagadeesh K, Li K, Zhang L, Singh M, Williamson M, Le M, Yu M, Singh MK, Zhang P, Vajda P, Duval Q, Girdhar R,

Sumbaly R, Rambhatla SS, Tsai S, Azadi S, Datta S, Chen S, Bell S, Ramaswamy S, Sheynin S, Bhattacharya S, Motwani S, Xu T, Li T, Hou T, Hsu WN, Yin X, Dai X, Taigman Y, Luo Y, Liu YC, Wu YC, Zhao Y, Kirstain Y, He Z, He Z, Pumarola A, Thabet A, Sanakoyeu A, Mallya A, Guo B, Araya B, Kerr B, Wood C, Liu C, Peng C, Vengertsev D, Schonfeld E, Blanchard E, Juefei-Xu F, Nord F, Liang J, Hoffman J, Kohler J, Fire K, Sivakumar K, Chen L, Yu L, Gao L, Georgopoulos M, Moritz R, Sampson SK, Li S, Parmeggiani S, Fine S, Fowler T, Petrovic V, Du Y. Movie Gen: A Cast of Media Foundation Models. *arXiv preprint arXiv:2410.13720*, 2025.

[304] Gu Y, Yin X, Rao Y, Wan Y, Tang B, Zhang Y, Chen J, Wang Y, Ma Z. Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders. In *IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, 1–5.

[305] Lu P, Wu J, Luan J, Tan X, Zhou L. XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System. *INTERSPEECH*, 2020.

[306] Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, Liu TY. Fastspeech: Fast, Robust and Controllable Text to Speech. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, 32.

[307] Chandna P, Blaauw M, Bonada J, Gómez E. Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan. In *European Signal Processing Conference*, 2019, 1–5.

[308] Chen J, Tan X, Luan J, Qin T, Liu TY. Hifisinger: Towards High-Fidelity Neural Singing Voice Synthesis. *arXiv preprint arXiv:2009.01776*, 2020.

[309] Huang R, Cui C, Chen F, Ren Y, Liu J, Zhao Z, Huai B, Wang Z. Singgan: Generative Adversarial Network for High-Fidelity Singing Voice Generation. In *ACM International Conference on Multimedia (ACM MM)*, 2022, 2525–2535.

[310] Wang C, Zeng C, He X. Xiaoicesing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network. *INTERSPEECH*, 2023.

[311] Ye Z, Xue W, Tan X, Chen J, Liu Q, Guo Y. CoMoSpeech: One-Step Speech and Singing Voice Synthesis via Consistency Model. In *ACM International Conference on Multimedia (ACM MM)*, 2023, 1831–1839.

[312] Zhang Y, Cong J, Xue H, Xie L, Zhu P, Bi M. Visinger: Variational Inference with Adversarial Learning for End-to-End Singing Voice Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, 7237–7241.

[313] Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning (ICML)*, 2021, 5530–5540.

[314] Zhang Y, Xue H, Li H, Xie L, Guo T, Zhang R, Gong C. VISinger 2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer. *INTERSPEECH*, 2023: 4444–4448.

[315] Kim S, Kim Y, Jun J, Kim I. Muse-svs: Multi-singer Emotional Singing Voice Synthesizer that Controls Emotional Intensity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[316] Wang Y, Hu R, Huang R, Hong Z, Li R, Liu W, You F, Jin T, Zhao Z. Prompt-Singer: Controllable Singing-Voice-Synthesis with Natural Language Prompt. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, 4780–4794.

[317] Luo YJ, Hsu CC, Agres K, Herremans D. Singing Voice Conversion with Disentangled Representations of Singer and Vocal Technique Using Variational Autoencoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 3277–3281.

[318] Lu J, Zhou K, Sisman B, Li H. Vaw-gan for Singing Voice Conversion with Non-parallel Training Data. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2020, 514–519.

[319] Li Z, Tang B, Yin X, Wan Y, Xu L, Shen C, Ma Z. Ppg-based Singing Voice Conversion with Adversarial Representation Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 7073–7077.

[320] Liu S, Cao Y, Hu N, Su D, Meng H. Fastsvc: Fast Cross-domain Singing Voice Conversion with Feature-wise Linear Modulation. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2021, 1–6.

[321] Jayashankar T, Wu J, Sari L, Kant D, Manohar V, He Q. Self-supervised Representations for Singing Voice Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5.

[322] Ning Z, Jiang Y, Wang Z, Zhang B, Xie L. Vits-Based Singing Voice Conversion Leveraging Whisper and Multi-Scale F0 Modeling. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, 1–8.

[323] Liu S, Cao Y, Su D, Meng H. Diffsvc: A Diffusion Probabilistic Model for Singing Voice Conversion. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, 741–748.

[324] Lu Y, Ye Z, Xue W, Tan X, Liu Q, Guo Y. Comosvc: Consistency model-based singing voice conversion. In *IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, 184–188.

[325] Chen S, Gu Y, Cui J, Zhang J, Chen R, Dai L. LCM-SVC: Latent Diffusion Model Based Singing Voice Conversion with Inference Acceleration via Latent Consistency Distillation. In *IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, 309–313.

[326] Chen S, Gu Y, Zhang J, Li N, Chen R, Chen L, Dai L. LDM-SVC: Latent Diffusion Model Based Zero-Shot Any-to-Any Singing Voice Conversion with Singer Guidance. *arXiv preprint arXiv:2406.05325*, 2024.

[327] Xue L, Wang C, Wang M, Zhang X, Han J, Wu Z. SingVisio: Visual analytics of diffusion model for singing voice conversion. *Computers & Graphics*, 2024, 124: 104058.

[328] Shen T, Jia J, Li Y, Wang H, Chen B. Enhancing Music Recommendation with Social Media Content: An Attentive Multimodal Autoencoder Approach. In *International Joint Conference on Neural Networks (IJCNN)*, 2020, 1–8.

[329] Xu T, Li J, Chen X, Yao X, Liu S. Mozart's Touch: A Lightweight Multi-Modal Music Generation Framework Based on Pre-Trained Large Models. *arXiv preprint arXiv:2405.02801*, 2024.

[330] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*, 2022, 12888–12900.

[331] Wang B, Zhuo L, Wang Z, Bao C, Wu C, Nie X, Dai J, Han J, Liao Y, Liu S. Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation. *arXiv preprint arXiv:2412.09428*, 2024.

[332] Tian Z, Jin Y, Liu Z, Yuan R, Tan X, Chen Q, Xue W, Guo Y. AudioX: Diffusion Transformer for Anything-to-Audio Generation. *arXiv preprint arXiv:2503.10522*, 2025.

[333] Tian S, Zhang C, Yuan W, Tan W, Zhu W. XMusic: Towards a Generalized and Controllable Symbolic Music Generation Framework. *arXiv preprint arXiv:2501.08809*, 2025.

[334] Tang Z, Yang Z, Zhu C, Zeng M, Bansal M. Any-to-Any Generation via Composable Diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, 36.

[335] Wu S, Fei H, Qu L, Ji W, Chua TS. NExT-GPT: Any-to-Any Multimodal LLM. In *International Conference on Machine Learning (ICML)*, 2024.

[336] Zhan J, Dai J, Ye J, Zhou Y, Zhang D, Liu Z, Zhang X, Yuan R, Zhang G, Li L, Yan H, Fu J, Gui T, Sun T, Jiang YG, Qiu X. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, 9637–9662.

[337] Liang X, Zhao Z, Zeng W, He Y, He F, Wang Y, Gao C. Pianobart: Symbolic Piano Music Generation and Understanding with Large-Scale Pre-Training. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2024, 1–6.

[338] Huang R, Li M, Yang D, Shi J, Chang X, Ye Z, Wu Y, Hong Z, Huang J, Liu J, et al.. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024, volume 38, 23802–23804.

[339] Hou S, Liu S, Yuan R, Xue W, Shan Y, Zhao M, Zhang C. Editing Music with Melody and Text: Using ControlNet for Diffusion Transformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, 1–5.

[340] Wang Y, Ju Z, Tan X, He L, Wu Z, Bian J, et al.. Audit: Audio Editing by Following Instructions with Latent Diffusion Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, 36: 71340–71357.

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

[342] Zhang Y, Ikemiya Y, Choi W, Murata N, Martínez Ramírez MA, Lin L, Xia GG, Liao WH, Mitsufuji Y, Dixon S. Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning. *arXiv preprint arXiv:2405.18386*, 2024.

[343] Lin L, Xia GG, Jiang J, Zhang Y. Content-Based Controls For Music Large Language Modeling. *arXiv preprint arXiv:2310.17162*, 2023.

[344] Lin L, Xia G, Zhang Y, Jiang J. Arrange, Inpaint, and Refine: Steerable Long-Term Music Audio Generation and Editing via Content-Based Controls. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, 7690–7698.

[345] Tsai FD, Wu SL, Kim H, Chen BY, Cheng HC, Yang YH. Audio Prompt Adapter: Unleashing Music Editing Abilities for Text-to-Music with Lightweight Finetuning. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[346] Manor H, Michaeli T. Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion. In *International Conference on Machine Learning (ICML)*, 2024.

[347] Zhang Y, Ikemiya Y, Xia G, Murata N, Martínez Ramírez MA, Liao WH, Mitsufuji Y, Dixon S. MusicMagus: Zero-Shot Text-to-Music Editing via Diffusion Models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, 7805–7813.

[348] Lan GL, Shi B, Ni Z, Srinivasan S, Kumar A, Ellis B, Kant D, Nagaraja V, Chang E, Hsu WN, et al.. High Fidelity Text-Guided Music Editing via Single-Stage Flow Matching. *arXiv preprint arXiv:2407.03648*, 2024.

[349] Liu H, Wang J, Li X, Huang R, Liu Y, Xu J, Zhao Z. MEDIC: Zero-shot Music Editing with Disentangled Inversion Control. *arXiv preprint arXiv:2407.13220*, 2024.

[350] Li S, Zhang Y, Tang F, Ma C, Dong W, Xu C. Music Style Transfer with Time-Varying Inversion of Diffusion Models. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024, volume 38, 547–555.

[351] Kim S, Kwon JB, Wang H, Yoo S, Lin Y, Cha J. A Training-Free Approach for Music Style Transfer with Latent Diffusion Models. *arXiv preprint arXiv:2411.15913*, 2024.

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

[353] Yu D, Song K, Lu P, He T, Tan X, Ye W, Zhang S, Bian J. MusicAgent: An AI Agent for Music Understanding and Generation with Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, 246–255.

[354] Liang X, Du X, Lin J, Zou P, Wan Y, Zhu B. ByteComposer: A Human-Like Melody Composition Method Based on Language Model Agent. *arXiv preprint arXiv:2402.17785*, 2024.

[355] Deng Q, Yang Q, Yuan R, Huang Y, Wang Y, Liu X, Tian Z, Pan J, Zhang G, Lin H, Li Y, Ma Y, Fu J, Lin C, Benetos E, Wang W, Xia G, Xue W, Guo YT. ComposerX: Multi-Agent Symbolic Music Composition with LLMs. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[356] Yu Y, Srivastava A, Canales S. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2021, 17(1): 1–20.

[357] Buffa M, Cabrio E, Fell M, Gandon F, Giboin A, Hennequin R, Michel F, Pauwels J, Pellerin G, Tikat M, et al.. The WASABI dataset: cultural, lyrics and audio analysis metadata about 2 million popular commercially released songs. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, 2021, 515–531.

[358] Information on https://github.com/gaussic/Chinese-Lyric-Corpus.

[359] Law E, West K, Mandel MI, Bay M, Downie JS. Evaluation of Algorithms Using Games: The Case of Music Tagging. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2009, 387–392.

[360] Sun G, Wong Y, Cheng Z, Kankanhalli MS, Geng W, Li X. DeepDance: Music-to-Dance Motion Choreography With Adversarial Learning. *IEEE Transactions on Multimedia*, 2021, 23: 497–509.

[361] Li R, Zhao J, Zhang Y, Su M, Ren Z, Zhang H, Li X. FineDance: A Fine-Grained Choreography Dataset for 3D Full Body Dance Generation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022: 10200–10209.

[362] Luo Z, Ren M, Hu X, Huang Y, Yao L. POPDG: Popular 3D Dance Generation with PopDanceSet. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024: 26984–26993.

[363] Gupta P, Fotso-Puepi JA, Li Z, Mehta J, Bera A. MDD: A Dataset for Text-and-Music Conditioned Duet Dance Generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, 13932–13941.

[364] Wu S, Li Y, Zhang W, Jia J, Zhu Y, Yan Y, Zhai G, Yang X. Singinghead: A large-scale 4d dataset for singing head animation. *arXiv preprint arXiv:2312.04369*, 2023.

[365] Li B, Liu X, Dinesh K, Duan Z, Sharma G. Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Transactions on Multimedia*, 2018, 21(2): 522–535.

[366] Sarasúa A, Caramiaux B, Tanaka A, Ortiz M. Datasets for the analysis of expressive musical gestures. In *International Conference on Movement Computing*, 2017, 1–4.

[367] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: 776–780.

[368] Kyriakou T, Aristidou A, Charalambous P. Multi-Modal Instrument Performances (MMIP): A Musical Database. In *Computer Graphics Forum*, 2025, e70025.

[369] Li G, et al.. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[370] Ma J, Hu M, Wang P, Sun W, Song L, Pei H, Liu J, Du Y. Look, Listen, and Answer: Overcoming Biases for Audio-Visual Question Answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[371] Zou J, Mei J, Ye G, Huai T, Shen Q, Dong D. EMID: An Emotional Aligned Dataset in Audio-Visual Modality. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, 2023, 41–48.

[372] Zhou Z, Mei K, Lu Y, Wang T, Rao F. HarmonySet: A Comprehensive Dataset for Understanding Video-Music Semantic Alignment and Temporal Synchronization. *arXiv preprint arXiv:2503.01725*, 2025.

[373] Zhu Y, Olszewski K, Wu Y, Achlioptas P, Chai M, Yan Y, Tulyakov S. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision (ECCV)*, 2022, 182–199.

[374] Information on https://huggingface.co/datasets/nyuuzyou/suno.

[375] Chi X, Wang Y, Cheng A, Fang P, Tian Z, He Y, Liu Z, Qi X, Pan J, Zhang R, et al.. MMTrail: A Multimodal Trailer Video Dataset with Language and Music Descriptions. *arXiv preprint arXiv:2407.20962*, 2024.

[376] Thao HTP, Roig G, Herremans D. EmoMV: Affective Music-Video Correspondence Learning Datasets for Classification and Retrieval. *Information Fusion*, 2023, 91: 64–79.

[377] Manco I, Weck B, Doh S, Won M, Zhang Y, Bogdanov D, Wu Y, Chen K, Tovstogan P, Benetos E, et al.. The Song Describer Dataset: A Corpus of Audio Captions for Music-and-Language Evaluation. *arXiv preprint arXiv:2311.10057*, 2023.

[378] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, 311–318.

[379] Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, 65–72.

[380] Lin CY. Rouge: A Package for Automatic Evaluation of

Summaries. In *Text Summarization Branches Out*, 2004, 74–81.

[381] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

[382] Chen L, Li Z, Maddox RK, Duan Z, Xu C. Lip Movements Generation at a Glance. *European Conference on Computer Vision (ECCV)*, 2018.

[383] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612.

[384] Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, Plakal M, Platt D, Saurous RA, Seybold B, et al.. CNN Architectures for Large-Scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 131–135.

[385] Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. Panns: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, 28: 2880–2894.

[386] Cramer AL, Wu HH, Salamon J, Bello JP. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, 3852–3856.

[387] Koutini K, Schlüter J, Eghbal-zadeh H, Widmer G. Efficient Training of Audio Transformers with Patchout. In *INTERSPEECH*, 2022, 2753–2757.

[388] Xu KT, Xie FL, Tang X, Hu Y. FireRedASR: Open-Source Industrial-Grade Mandarin Speech Recognition Models from Encoder-Decoder to LLM Integration. *arXiv preprint arXiv:2501.14350*, 2025.

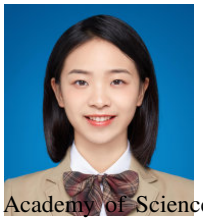
[389] Information on https://github.com/yxlllc/RMVPE.

Author biographies

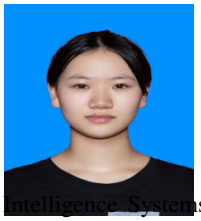
Sifei Li received a B.S. in automation in 2022 from Shandong University, China. She is pursuing a Ph.D. degree in the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. Her research interests include multimodal music generation and computational art.

Mining Tan received a B.S. in automation from the University of Science and Technology, Beijing, China, in 2024. She is working towards an M.S. degree in control theory and control engineering in the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. Her current research interests include deep learning and multimedia computing.

Feier Shen received the B.S. degree in Artificial Intelligence in 2025 from the University of Chinese Academy of Sciences, China. She is a Ph.D. student at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. Her research interests include motion generation and computational art.



Minyan Luo received the B.S. degree in Artificial Intelligence in 2025 from the University of Chinese Academy of Sciences, China. She is a Ph.D. student at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. Her research interests include computer vision, computer graphics, and computational art.



Zijiao Yin is an undergraduate student at the University of Chinese Academy of Sciences, majoring in Artificial Intelligence, and will receive her B.Sc. degree in 2025. She will continue her studies as a Ph.D. student at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. Her research interests include computer vision, computer graphics, and computational art.

Fan Tang received a B.Sc. in computer science from North China Electric Power University, Beijing, China, in 2013, and a Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is an Assistant Professor in the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.

Weiming Dong is a Professor in MAIS. He received his B.Sc. and M.Sc. degrees in 2001 and 2004, both from Tsinghua University, China. He received his Ph.D. in Computer Science from the University of Lorraine, France, in 2007. His research interests include artificial intelligence generated content and computational art.

Changsheng Xu is a Professor in MAIS. He received the Best Associate Editor Award of ACM Transactions on Multimedia Computing, Communications, and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He has served as in various editorial roles for over 20 prestigious IEEE and ACM multimedia journals, conferences, and workshops. He is an IEEE Fellow, IAPR Fellow and an ACM Distinguished Scientist.