Generative Artificial Intelligence (AI) for Music: A Review

Julia Goh
Department of Computer Science
University College London
London, United Kingdom
julia.goh.20@ucl.ac.uk

Philip Treleaven
Department of Computer Science
University College London
London, United Kingdom
p.treleaven@ucl.ac.uk

***Abstract*—The AI music/audio fields are rapidly gaining attention due to Generative AI, following the increasing maturity of Machine Learning (ML) models in the Computer Vision (CV) and Natural Language Processing (NLP) fields. Generally, Generative AI tools are broadly categorised into a) general-purposed such as ChatGPT , b) specialised such as CLIP [1] , and c) application-specific such as MusicLM [2] In this paper, we review music Generative AI tools by first understanding pioneering image Generative AI models, then introducing the common concepts and techniques applied for designing Generative AIs. Finally, an overview of the recent audio-related Generative AI models such as AudioLM [3], and LP-MusicCaps [4] is presented. The paper's contribution is to review the characteristics, advantages and differences of the respective ML models and AI music applications along with possible future work for Generative AI in the music field.**

I. Introduction

Following advances in Natural Language Processing (NLP) models such as BERT [5] and BART [6], the audio/music fields are rapidly gaining focus, such as speech recognition and music generation. The selection of music datasets such as MusicCaps [2] is also rather limited if compared to image datasets.

In terms of ML image generation, Stable Diffusion [7] is one of the most popular and successful models due to its high resolution and creative results. Pioneering works, such as DALL-E [8] and CLIP-Gen [9] have also investigated the feasibility of generating images with Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) [10], which are further reviewed in Section 2. There have been several successful text-to-music models so far. Examples are AudioLM [3] and MusicLM [2] by Google, and MusicGen [11] by Meta. Moreover, a very recent work in music-to-text generation has also been proven feasible in LP-MusicCaps [4].

Another important technology is embedding representations , where numerical representations of real-world objects are used by Generative AI to understand complex knowledge domains. An example is the MULAN joint text-music embedding model introduced in [12] and applied in [2]. Several works [13, 1, 12, 14] have proven that the application of joint embedding helps in defining relationships between cross-modal data and improving the generation quality.

The paper starts with an introduction of the pioneering image Generative AI models to understand the general design of Generative AI on the high level. Then, the available music-related ML models will be presented and compared to give a better idea of their working principles. Where relevant, an overview of existing music generation tools will also be compared, for instance, Melobytes and Mubert. Finally, the commonly used metrics for image and audio/music quality evaluation are introduced, and the paper concludes with discussion on limitations and future work of Generative AI for music.

II. Image Models

In this section, popular Generative AI models for image generation are presented along with some examples of generated images. Finally, the section concludes with comparisons between the different AI image models in terms of their relevance and quality.

A. Stable Diffusion

The Stable Diffusion model [7] is designed for text-to-image generation. It is built on top of the latent diffusion model (LDM) [15]. Its strategy is to gradually and repeatedly add noise to an input, then training takes place for removing the noise and restoring the encoded input into its original state. While excellent quality can easily be achieved through this method, it is computationally expensive and slow due to the number of iterations required. Then, with the help from the popular CLIP [1] text-image joint

embedding, the resulting text-image embeddings are used for model conditioning during training and inference.

From the qualitative perspective, some examples are shown in Figure 1. Compared to other models which often produces low quality images, Stable Diffusion pushed the limits with its ability to produce high quality images [15].

Figure 1. Example Generations from Stable Diffusion [7]

B. DALL-E

DALL-E [8] is another Generative AI for text-to-image generation. Its main component is discrete variational autoencoder (dVAE) [8] instead. As VAEs are known for their limitation in terms of generation resolution and quality, a larger codebook size of 8192 is used to mitigate the issue [8]. This allows the latent to learn and store more features and information. Once the input image is tokenised with dVAE, an autoregressive transformer is applied for modelling the text-image joint space [8]. The components are trained in two stages, first dVAE, followed by the autoregressive transformer. Some generations from DALL-E are shown in Figure 2. While the generations satisfy its respective prompts, the quality of the images is arguably not as attractive as Stable Diffusion (Figure 1).

Figure 2. Example Generations from DALL-E [8]

C. CLIP-Gen

CLIP-Gen [9] is a transformer-based text-to-image Generative AI model. Similar to its name, it applies the CLIP joint text-image embedding model [1] internally for encoding the text input. For encoding the reference image during training and decoding the generated image during inference, VQ-GAN [16, 10] architecture is used. For better understanding, the CLIP model will be introduced, followed by a deeper brief into the components of CLIP-Gen.

1) CLIP: CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. CLIP consists of two encoding towers for image and text respectively [1]. For the image part, proven architectures such as ResNet-50 [17] are considered, with small modifications in terms of additional layer normalisation and combined patch PE. When it comes to the text encoder, a modified CLIP transformer [1] operating on lower-cased byte pair encoding (BPE) is applied.

To access CLIP, experiments have been done on 1) zero-shot prediction, and 2) visual n-grams comparison in [1]. In experiment 1, CLIP outperforms pure ResNet-50 when asked to recognise actions in videos. CLIP has an improvement of 14.5% with the Kinetics700 dataset and outperforms by 7.7% on UCF101 dataset . In the second experiment, CLIP exhibits a top-5 accuracy of 95%, which is matching Inception-V4 [18] (baseline).

2) CLIP-Gen: Next, CLIP-Gen [9] is a combination of CLIP [1] and VQ-GAN [16, 10]. CLIP is used to extract text-image key features from the joint space whereas VQ-GAN is used as an image tokenizer. Joining the two, the transformer decoder learns to reconstruct the image tokens from VQ-GAN, conditioned on embeddings from CLIP.

Qualitatively, some results can be seen from Figure 3. While CLIP-Gen outperforms DALL-E (Figure 2) slightly, the styles are arguably simple when compared to Stable Diffusion (Figure 1).

Figure 3. Example Generations from CLIP-Gen [9]

D. Discussion

Based on the examples shown above for the respective image generation model, the different aspects of the generated images will be compared in Table I.

III. Audio/Music Models

With increasing attention to Generative AI in the audio/music field, there have been several works published for studying the possibility and feasibility of audio/music generation . These researches of Generative AI for audio/music are presented below along with some other examples of closed-source Generative AI tools for music.

A. LP-MusicCaps

LP-MusicCaps [4] is a recent model for music-to-text generation which wad introduced very

Table I
STYLES AND QUALITY COMPARISON OF THE GENERATED IMAGE

Model	Type	Relevance	Quality	Customisation
Stable Diffusion	Diffusion	High	High	Text prompt and optional image condition
DALL-E	Discrete VAE	Good	Acceptable	Text prompt and optional image condition
CLIP-Gen	Transformer	Good	Acceptable	Text prompt and optional image condition

recently in [4]. The model is motivated by the limited amount of music dataset, and the fact that the procedure of gathering music datasets is complex and costly. The resulting size of the dataset is often limited too, for example, the popular MusicCaps [2] has only 5521 text-music entries. Thus, LP-MusicCaps allows the expansion of music dataset with reliable pseudo-data.

An encoder-decoder model is involved in the architecture of LP-MusicCaps [4]. Instead of designing from scratch, the BART [6] model is applied due to its wide compatibility and outstanding performance on sequential data. For training the text conditioner, the pseudo captions generated from GPT3.5 Turbo are used [4]. In the pseudo captions generation phase, model hallucinations are also checked in case the LLM is making up its output, deviating from facts.

The results [4] has shown that LP-MusicCaps has taken the win in the transfer-learning task in terms of BERT-score. An improvement in novelty score has also proven that repetitions and overlap to the training dataset has been reduced and minimised.

B. Text-to-Audio/Music Generative SOTA Models

In this section, audio/music generation models conditioned on text will be described. They are usually open-source and released as research papers. For better understanding, their respective architectures and distinct characteristics will be described.

1) AudioLM: In [3], AudioLM is introduced for generating high-quality and coherent audio from text inputs. The results are proven through speech and piano continuation experiments. The target acoustic tokens are generated with the encoder and RVQ in Soundstream [19] whereas the target semantic tokens are produced with the intermediate layer of w2v-BERT [14]. These models takes audio as input and outputs the respective types of token. They are then used for training the modelling stages for coherency.

There are three subsequent modelling stages as shown in Figure 4, which are built with transformers decoder for learning the next tokens given current ground-truth tokens [3].

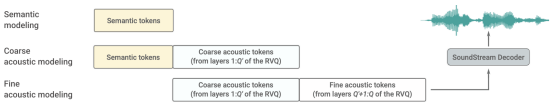

Figure 4. Token Modelling Stages in AudioLM [3]

Each stage is summarised in Table II in terms of the involved tokens type and the modelled distribution.

2) MusicLM: For text-to-music generation, MusicLM is proposed in [2] as an extension to AudioLM [3], with MULAN [12] joint text-music embedding model included for encoding the text input. With examples shown on the Google MusicLM website [2], the model is able to take different types of text input for generating music of different length. Some examples of text input include rich caption, long story-like texts, art captions, and more. On top of conditioning with text, the model can also be conditioned on music or audio.

In this case, the audio embedding tower used in MULAN is ResNet50 [17] whereas the text embedding model remains as BERT [5]. MULAN is used in the model due to its ability to learn the cross-modal characteristics even if the audio-text pair is weakly associated [12, 2]. This lowers the requirements of training data.

In terms of generation quality, a selection of demos can be found here. While it supports both short (several seconds) and long (a few minutes) generation, and accepts short prompts or long story-like captions, the generated music is not always pleasant to listen to. There are cases where the melody is out-of-tune, and the quality sounds coarse. When it is prompted to generate tunes from classical instrumentals, the results do not always sound natural. For example, a request of piano tunes may turn out to resemble more from an organ instead. This may be related to training data quality, but there is a wide room of improvements

Table II
AUDIOLM MODELLING STAGES [3]

Stage	Objective
Semantic	For capturing the temporal structure, the autoregressive distribution $p(z_t \| z_{<t})$ is learnt.
Coarse Acoustic	Conditioned on the semantic tokens. Aims to recover the audio properties and characteristics with a masked language modelling objective. A simple flattening approach is taken for learning the hierarchical structure, resulting in a poor reconstruction.
Fine Acoustic	Conditioned on the coarse acoustic tokens. Aims to upgrade the resulting audio quality by removing the lossy artifacts from compression.

here in terms of melody and quality.

3) MusicGen: In contrast to other works [2], MusicGen, another music Generative AI model relies only on a single transformer decoder for acoustic modelling [11]. In the experiment, the text conditioning stage is added for text-to-music generation. Existing proven encoders such as FLAN-T5 [20] are adapted. The model also allows melody conditioning, where it is trained with data in time-frequency format. There were also experiments done on raw chromogram, but it is observed that overfitting often occurs [11].

A significant contribution of MusicGen is the token interleaving patterns introduced. From Figure 5, there are four different patterns illustrated in pictures.

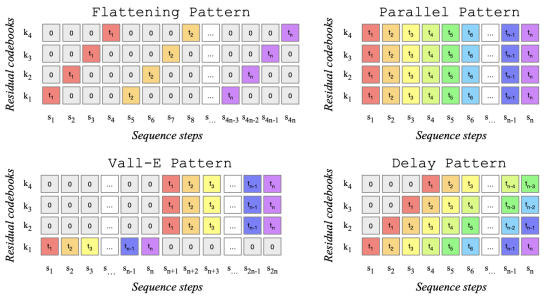

Figure 5. Codebook interleaving patterns in MusicGen [11]

The first being the straightforward **Flattening Pattern** (Equation 1). Assuming we have K codebooks in the encodec, the codes are flatten such that the representation at time step t is extracted sequentially one codebook per timestep, then concatenated. However, complexity is higher due to certain level of lost in gain [11].

$$p_{t,k}(V_{t-1}, ..., V_0) \triangleq \mathbb{P}[Vt, k | V_{t-1}, ..., V_0] \\ \forall_t > 0, \forall_k, \mathbb{P}[\tilde{V}_{t,k}] = p_{t,k}(\tilde{V}_{t-1}, ..., \tilde{V}_0) \quad (1)$$

With Flattening Pattern as the base, **Parallel Pattern** is introduced and defined as Equation 2 [11]:

$$P_s = (s, k) : k \in 1, ..., K \quad (2)$$

We can also have the **Vall-E Pattern** defined as Equation 3 [11]:

$$\begin{cases} P_s = (s, 1) & if\ s \leq T \\ P_s = (s, k) : k \in 2, ..., K & otherwise \end{cases} \quad (3)$$

Lastly, the **Delay Pattern** decomposition is described as Equation 4 [11]:

$$P_s = (s - k + 1, k) : k \in 1, ..., K, s - k \geq 0 \quad (4)$$

Apart from contributions, MusicGen has limitations such as low-quality of the generated audio with the pretrained `musicgen-small` model on HuggingFace. Test generation of the model is also accessible here. If compared to [2], MusicGen has some limitations on the length of the input sequence. Future work also includes finding ways of supporting fine-grained control over the generated acoustic coherence and quality [11].

C. Generative Music Applications

There are ready applications on the market where users can generate high quality royalty-free music or subscribed to access more premium features. These are mostly closed-source, thus the applications will only be introduced on a high-level basis.

1) Melobytes: On top of text-to-music generation, Melobytes offers more features such as image-to-music, image-to-song, and video-to-music generations. For text-to-music generation, Melobytes allows a wider range of customisation than Mubert. These include input text language, tempo (beats per minute), tonality (major or minor key), acappella generation, and more. They also allow addition of lyrics with AI completion support.

2) Mubert: Mubert is a text-to-music generation platform, where a user can generate a track for free, customised through text-prompt. It allows generation as short as 15 seconds, up to 25 minutes. There are some ready samples here for immediate listening. Overall, the generation quality is very good such that a) the music fits the prompt and category well, b) the composition is very coherent, and c)

Table III
STYLES AND QUALITY COMPARISON OF THE GENERATED MUSIC

Model	Relevance	Quality	Music Coherence	Customisation
MusicLM	Acceptable	Bad	Inconsistent	Text prompt and optional melody condition
MusicGen	Acceptable	Bad	Inconsistent	Text prompt and optional melody condition
Mubert	High	Excellent	Smooth and logical	Text prompt and duration
Melobytes	High	Excellent	Smooth and logical	Tempo, lyrics, tonality, vocal and many more

the quality is aligned with professional composed tunes. There is a paid contributing opportunity in Mubert, where artists can upload their samples, tracks and loops to Mubert Studio. This also helps in expanding its reference database to contain a large amount of professional contents, thus largely improving its music generation quality.

D. Discussion

In Table III, we will compare the different aspects of the generated music from different models and applications stated in the previous subsections.

IV. CONCLUSION AND FUTURE WORK

While there has been several eye-catching works showing the feasibility of audio and music generative models, there remains a wide room of improvements. Some of the challenges and future work are listed and described below.

1) Ethics: A big concern in generative models is ethics. The generated works of these models usually requires no license and can be used by anyone in the public (with citations). However, as references and training data may overlap with existing copyrighted work, this may turn out unfair for the original creators. A possible approach may be constantly evaluating legal matters and policies, then redefining them whenever required.

Another sensitive issue comes from religion or culture differences. Public feedback should always be taken into consideration so that the related problems can be eliminated efficiently. Where possible, negative keys may be used during training so that the model is able to learn and avoid these in the generations. One example is the use of negative keys in InfoNCE Contrastive Loss . Post-processing filters may also be introduced to mask and replace sensitive information.

2) Reinforcement Learning (RL): There are recent interest around incorporating RL during the training phase of a model. This means that the Markov decision process (MDP) in RL is adapted and a reward/penalty function is defined for the model to learn. However, completely replacing the training style with RL may not be relevant. Some issues are reward function is difficult to define due to the black-box nature of certain models. Where applicable, we may be able to apply RL in the fine-tuning phase such that the model is trained to specialise in certain categories of task.

3) Advancement in Image-Music Generation: There has been an increasing number of music generation models conditioned on text prompts, but there has not been an attempt for exploring the possibilities on image prompts. Similarly, there has been various research on text-conditioned image generation models but this is not the case when it comes to audio-conditioned image generation models. There is a strong potential in this field and successful publications in image-music generations could bring ML and AI a step forward.

4) VAE-GAN: VAEs have an advantage in efficient sampling as well as great level of diversity in its latent space whereas GANs are known for its excellent generation quality under short generation time. While diffusion has good diversity as well as generation quality, it requires a long time for data generation. However, VAEs and GANs are able to achieve good result (diversity and quality) when being combined. VAE-GANs have potentials and can be further worked on in the audio/music field.

REFERENCES

[1] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[2] Andrea Agostinelli et al. “Musiclm: Generating music from text”. In: *arXiv preprint arXiv:2301.11325* (2023).

[3] Zalán Borsos et al. “Audiolm: a language modeling approach to audio generation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[4] SeungHeon Doh et al. “Lp-musiccaps: Llm-based pseudo music captioning”. In: *arXiv preprint arXiv:2307.16372* (2023).

[5] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Mike Lewis et al. “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. In: *arXiv preprint arXiv:1910.13461* (2019).

[7] *Stable Diffusion*. 2022. URL: https://github.com/CompVis/stable-diffusion.

[8] Aditya Ramesh et al. “Zero-shot text-to-image generation”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[9] Zihao Wang et al. “Clip-gen: Language-free training of a text-to-image generator with clip”. In: *arXiv preprint arXiv:2203.00386* (2022).

[10] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[11] Jade Copet et al. “Simple and Controllable Music Generation”. In: *arXiv preprint arXiv:2306.05284* (2023).

[12] Qingqing Huang et al. “Mulan: A joint embedding of music audio and natural language”. In: *arXiv preprint arXiv:2208.12415* (2022).

[13] Yusong Wu et al. “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[14] Yu-An Chung et al. “W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training”. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2021, pp. 244–250.

[15] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

[16] Jiahui Yu et al. “Vector-quantized image modeling with improved vqgan”. In: *arXiv preprint arXiv:2110.04627* (2021).

[17] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[18] Christian Szegedy et al. “Inception-v4, inception-resnet and the impact of residual connections on learning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[19] Neil Zeghidour et al. “Soundstream: An end-to-end neural audio codec”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), pp. 495–507.

[20] Hyung Won Chung et al. “Scaling instruction-finetuned language models”. In: *arXiv preprint arXiv:2210.11416* (2022).

BIOGRAPHIES

Julia Goh, is a student at University College London, London, WC1E 6BT, U.K. Her research interests include machine learning, artificial intelligence, and computer graphics. Contact her at julia.goh.20@ucl.ac.uk.

Philip Treleaven, is a professor of computing at University College London, London, WC1E 6BT, U.K, and director at the UK Centre for Financial Computing & Analytics. His research interests include data science, algorithms, and blockchain technologies. Treleaven received a Ph.D. from The University of Manchester. He is a Member of the IEEE and the IEEE Computer Society. Contact him at p.treleaven@ucl.ac.uk.