Generative AI: Image-to-Music and Music-to-Image

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 paper explores the use of various machine learning (ML) techniques to design and build Generative Artificial Intelligence (AI) architectures for achieving: a) music-conditioned image generation, and b) image-conditioned music generation. To achieve these architectures, three studies were conducted for developing the Generative AI models. Firstly, the paper starts with a broad analysis of the data collection process for creating the image-music dataset (ImMuTe). Next, the development of a suitable joint embedding model (MuVis) is carried out for Generative AI for Music. The paper then summarises the outcomes of the Generative AI models for: a) music-to-image generation (MusIm), and b) image-to-music generation (Imagic), along with the respective qualitative results. Finally, the paper concludes with statements of contributions, limitations and future work.

I. INTRODUCTION

While image generation technology has progressed rapidly for several years, music generation is a relatively new research interest. This paper investigates the use of various ML techniques to design and build a suitable model architecture for building a music-conditioned image Generative AI, and an image-conditioned music Generative AI.

The research is motivated by the following scenario: "*An artist should also be allowed to express their desired musical requirements in painting and colours, whereas a musician creates arts and painting through melody and tunes*". In practice, a game developer is able to feed shots of their gameplay and environment design for obtaining a suitable soundtrack for the gaming product. On the other hand, branding and advertisement could be made easier by obtaining a matching illustration with the target music as the input.

To this end, three investigations were conducted for developing the Generative AI models. To achieve this, the paper starts with a high-level overview of the data collection process. This is followed by the development of a suitable joint embedding model for the task, then concludes with the resulting Generative AI architectures for music-to-image and image-to-music generation models.

II. RELATED WORK

In this section, the background research is introduced. Firstly, the MusicCaps [1] text-music dataset is presented. Next, pioneering embedding models of Audio Spectrogram Transformer (AST) [2] for embedding audio/music and Vision Transformer (ViT) [3] for embedding image. These will be further discussed when designing the architecture of the image-music embedding model. Finally, the Latent Diffusion [4] and Transformers [5] models are crucial for the music-to-image and image-to-music Generative AI models.

A. MusicCaps

MusicCaps [1] is a text-music dataset, introduced for training the MusicLM [1] text-to-music generative model. It is built with AudioSet [6] as the base. As MusicCaps is related to music only, irrelevant audios (outside of the music category) are removed, leading to the size of MusicCaps being only 5521 music-text pairs. The music annotations are written by music professionals, which made it a very reliable and efficient dataset. It is publicly made available along with the research of MusicLM [1].

B. AST

AST [2] is an audio embedding transformer-based architecture. It transforms the input audio into spectrogram of 128 mels, 25ms Hamming window and 10ms step size [2]. As shown in Figure 1, the spectrogram is then split into 16 x 16 patches before being flatten and projected linearly into 768 dimensions [2]. The transformers encoder used is configured to have 768 embedding dimensions, 12 layers and 12 heads [2]. PE is also added here for including the location information which is absent in the Transformer architecture itself. The spectrogram does not have `[CLS]` tokens by default too. To solve this, learnable `[CLS]` token is prepended to the input sequence (with reference to [7]) before being fed into the transformer encoder. These tokens are the core representations of the audio spectrogram [2].

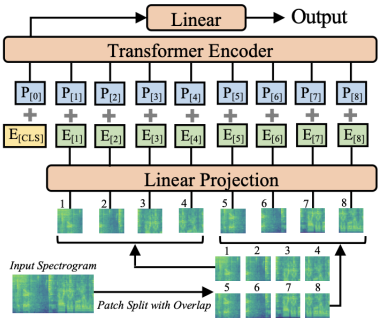

Figure 1. Architecture of AST [2]

C. ViT

ViT [3] is designed to embed images with the transformer-based architecture shown in Figure 2. As images are 2D by default, the inputs are flatten being passed into the transformer due to its input dimension requirement (i.e., 1D). Similar to [7] as well, learnable `[CLS]` token is prepended to the embedded patches for the transformer to learn the compact representations. Commonly seen in transformer-based architectures, PE is applied to the input sequences. A remarkable experiment done as part of the ViT research is self-supervised learning. As with other text-based transformer models [7], masked patch prediction is performed for showing the bright future of self-supervised ViT [3].

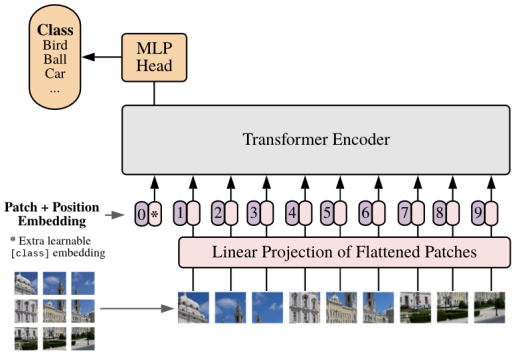

Figure 2. Architecture of ViT [3]

D. Latent Diffusion

Latent diffusion [4] learns to add and reverse noise as shown in Figure 3, known as the diffusion process. In the early days, there has been consistent issues with generation quality and diversity with VAEs and GANs. By denoising and generating through diffusion, latent diffusion [4] has managed to outperform them with its ability to predict and generate high-resolution images. With the help of the UNet [8] architecture as the denoiser, the diffusion process is defined as Equation 1 below [4], with t uniformly sampled from $1, ..., T$ and x_t being the noisy version of input x. The latent diffusion model has succeeded in the tasks of text-to-image generation, image synthesis, semantic landscape synthesis, object removal and more. However, there exists limitation in terms of latency as the noise reversal process is required to be carried out for multiple timesteps.

$$LDM = \mathbb{E}_{x,\epsilon\ N(0,1),t}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2] \tag{1}$$

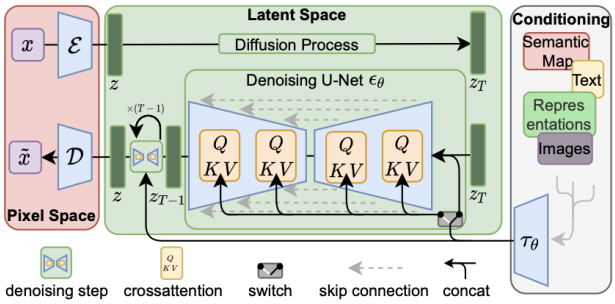

Figure 3. Architecture of the Latent Diffusion Model for Conditional Generation Tasks [4]

E. Transformers

Transformers are popular in NLP as it is efficient in processing sequential data and learning their relationship through multi-head self and cross attention [5], assembled as shown in Figure 4. This also opens up opportunity for semi-supervised and unsupervised learning through unlabelled data [9]. One downside of transformers is that due to its high complexity, it requires large dataset and thus slow to train.

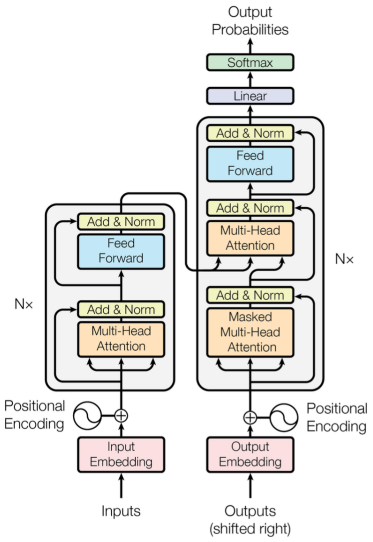

Figure 4. Architecture of Transformers [5]

Based in [5], the encoder maps input symbols $(x_0, x_1, ..., x_n)$ to a continuous representation vector $(z_0, z_1, ..., z_n)$. Given z, the decoder then learns to generate or map to the symbols $(y_0, y_1, ..., y_n)$, one per time. The encoder and decoder blocks are commonly a stack of N (usually $N = 6$) layers respectively, with an embedding space configuration of $d = 512$.

To achieve these, there are several techniques applied. The first is position encoding for representing the position of each token and learning the correct sequence. Next, attention is used to relate the relationship between tokens, which is the way previous tokens affect the prediction of the upcoming tokens.

III. Image-Music Dataset - ImMuTe

There is a limited number of image-music dataset for training and validation. Therefore, we propose an image-music-text triplets synthetic dataset (ImMuTe), created with the help of the Stable Diffusion 2 [10] model and the MusicCaps [1] dataset. While real world dataset is the preferred option, the data collection

process can become tricky. For example, extra manpower and special equipment may be required.

For building a reliable and feasible dataset, a suitable structure is required for direct attribute access. During the image generation phase, several different models are also reviewed and analysed to select the most appropriate one for the task. These methodologies are described next, wrapping up with the qualitative results.

A. Methodology

ImMuTe consists of image-music-text data pairs, with accessible attributes shown in Table I.

Table I
ATTRIBUTES IN IMMUTE

Attribute	**Description**
ytid	This represents the YouTube ID of the clip, where the data should then be downloaded with this ID. This is dependent on MusicCaps [1].
start_s	This indicates the starting time in seconds of the clip. This is dependent on MusicCaps [1].
end_s	This indicates the ending time in seconds of the clip. This is dependent on MusicCaps [1].
caption	This represents the text caption of the music clip. This is dependent on MusicCaps [1].
aspect_list	This is a list of keywords describing the music. This is dependent on MusicCaps [1].
image_link	This is the link to the image associated with the text caption and music clip. The link could be used to download the image from GitHub. This is new in ImMuTe.

For selecting a suitable state-of-the-art (SOTA) model for the task, we further reviewed and compared several popular models in terms of styles and metrics. These are presented in Table II. Finally, the popular Stable Diffusion [11] is chosen as the image generation model for the ImMuTe dataset due to its ability to generate images with high quality and attractive styles. The general generation pipeline is described next.

Table II
QUALITY AND METRICS COMPARISON BETWEEN THE GENERATION MODELS

Model	**Relevance**	**Quality**	**Metrics**
Stable Diffusion [11]	High creativity	High resolution	Different configurations and variants of LDM is being experimented. When pushing its limits, the large LDM-4 with fine-tuning managed to achieve an FID score of 1.50, which is more than 0.3 improvement from the second-place model - CoModGAN [12].
DALL-E [13]	Good	Acceptable	It achieved an FID score of 27.5 [13, 14], outperforming other architectures such as AttnGAN [15]. In terms of IS score, DALL-E records 17.9, which is reasonable although not the best.
CLIP-Gen [14]	Good	Acceptable	In terms of FID, CLIP-Gen outperforms similar models such as DALL-E [13] and CogView [16]. For CapS, while CLIP-Gen turns out second, achieving a score of 0.13751 [14] which is very close to CogView's 0.17403.

ImMuTe relies on MusicCaps [1] for the music clip and text data. For generating the corresponding image, the text description from MusicCaps [1] is passed into the text-to-image generation model. Different pretrained Stable Diffusion 2 models are considered, namely `stable-diffusion-2-1` and `stable-diffusion-2-1-base`. To ensure that the generation is non-bias [17] and up to quality, we compare generations from the different checkpoints. This is to make sure that the generated images are varying in styles while remaining to be relevant to the text prompt.

B. Results

As we are unable to show audio in the paper, only text-image pairs are displayed here. This shows the style and quality of the generated images in general. The contributions and limitations are then further discussed later in the paper.

Example 1:
Text: "The low quality recording features a ballad song that contains sustained strings, mellow piano melody and soft female vocal singing over it. It sounds sad and soulful, like something you would hear at Sunday services."

Images Generated: see Figure 5 below.

Figure 5. Generated Image for Example 1

Example 2:
Text: "A male vocalist sings this trippy Electronic song. The tempo is slow with synpaperer arrangement, digital drums and electronically articulated music. The song is catchy, youthful, enthusiastic, trippy, psychedelic , new age, and groovy. This song is an Electro -Hop/ Hip-Hop."
Images Generated: see Figure 6 below.

Figure 6. Generated Image for Example 2

Example 3:
Text: "This music is a pensive instrumental. The tempo is medium with a melancholic cello, rhythmic ukulele and a percussion instrument. The song is pensive, contemplative, meditative, melancholic, solemn and dreamy."
Images Generated: see Figure 7 below.

Figure 7. Generated Image for Example 3

IV. Image-Music Joint Embedding Model - MuVis

As AI audio and music are the current hot topic in ML, popular pioneering embedding models include CLIP [18] for learning the text-image joint space, and MULAN [19] which is the text-audio version of the joint embedding model. However, embedding model related to image-music pair remains new and there is limited significant open source work and research in the field. Thus, we propose an effective image-music joint embedding model (MuVis) in this study. The model aims to help in: a) music tagging with image labels, b) image tagging with music labels, c) music retrieval from image collection, and d) image retrieval from music collection. The design methodology is presented next along with the quantitative results.

A. Methodology

1) Music Embedding Network: The proven encoder architecture - AST [2] is considered for music embedding. Inheriting all configurations from [19], the raw music is first pre-processed into a mel spectrogram with 128 mels, 25ms Hanning window and 10ms step size. Then, transformer encoders are used for learning the embeddings. Before returning the music embeddings, the tokens are reduced and averaged for finding the global mean. Recent research has shown that this approach is more effective than prepending `[CLS]` tokens [20]. The outputs are then normalised and linear-projected to the 128 dimension space [19].

2) Image Embedding Network: The network considers the architecture of the popular ViT [3]. In the configurations, a hidden dimension of 768 is used, targeting at images of resolution 224 x 224. Similar to the final processing of the music embeddings, the image embeddings are also reduced and averaged for finding the global mean before being returned (instead of prepending `[CLS]` token which is less effective). The outputs are then normalised and linear-projected to the 128 dimension space [19].

3) Overview: Putting them together, the image-music joint embedding model consists of 2 embedding towers as shown in Figure 8, inspired by the successful CLIP [18] and MULAN [19] models. The model can take either music input, image input, or both at the same time. The model encodes the data with AST for music input and ViT for image input in its forward call. These embeddings can then be used for classification, data retrieval, or other tasks by adding a final connected layer. The learning objective is presented next.

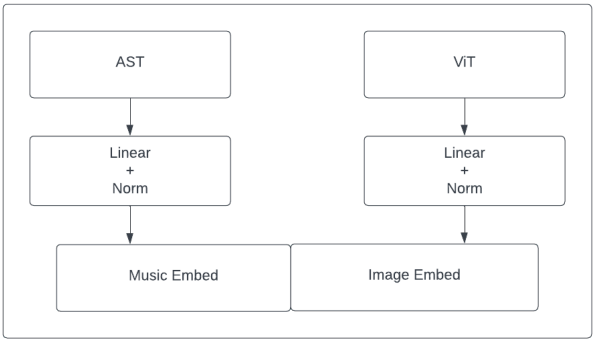

Figure 8. Architecture of the Music-Image Joint Embedding Model

4) Learning: For learning the relationship between the image and music embeddings, the contrastive loss [18, 19] is learnt so that the embeddings from the same image-music pair are brought closely and similar to each other. Given the definition of cross entropy loss (CE) in Equation 2, the contrastive loss (CL) is defined in Equation 3.

$$CE(x, y) = \{CE_1, ..., CE_N\}^T, CE_n = log \frac{exp(x_{n,y_n})}{\sum_{c=1}^{C} exp(x_{n,c})} \tag{2}$$

$$CL(music,\ image) = \frac{CE(music,\ y) + CE(image,\ y)}{2}, y = true\ labels \tag{3}$$

B. Results

For the AST and ViT encoders, their respective pretrained checkpoints on Hugging Face are loaded before training for speeding up the process. `MIT/ast-finetuned-audioset-10-10-0.4593` pretrained on AudioSet [21] is loaded for AST [2] where `google/vit-base-patch16-224-in21k` pretrained on ImageNet 2012 is loaded for ViT [3]. Training on top of the base checkpoints, the results for the experiments of few shot learning, image retrieval from music queries, and music retrieval from image queries are presented next.

Due to the limited number of effective benchmark/baseline for comparisons and validation, the performance of MuVis is compared against other multimodal joint embedding models such as CLIP [18] (text-image), MULAN [19] (text-audio) and CLAP [22] (text-audio) for understanding the typical performance of a successful joint embedding model. The discussion then revolves around the metrics below for evaluating the results in Table III and showing the feasibility of MuVis.

- Accuracy
  The percentage in decimals form in which the model correctly classify/tag the music-image data.
- AUROC Score
  Area under ROC curve. The score measures the efficiency of the model by analysing the area under the curve over true positive against false positive rates.
- Average Precision [23]
  Evaluates the model performance by combining recall and precision.

Table III
EXPERIMENT RESULTS OF MUVIS IN TERMS OF ACCURACY, AUROC, AVERAGE PRECISION.

Experiment	Accuracy	AUROC	Average Precision
(a) **Tagging Task**			
MuVis - image/music (Ours)	0.8958	0.8833	0.8618
CLIP [18] - image/text	0.9350	0.8840	-
MULAN [19] - text/audio	-	0.8400	-
CLAP [22] - text/music	-	-	0.8260
(b) **Image Retrieval from Music Queries**			
MuVis - image/music (Ours)	0.8475	0.8838	0.8978
CLIP [18] - image/text	0.9060	-	-
(c) **Music Retrieval from Image Queries**			
MuVis - image/music (Ours)	0.8333	0.8634	0.8590
MULAN [19] - text/audio	-	0.9030	-
CLAP [22] - text/music	-	-	0.7020

1) Few Shot Learning: For investigating the music-image tagging task, the few shot learning experiment is setup for evaluating MuVis. Due to the many-to-many mapping relationship between the image and music data, they are not uniquely classify such that each pair belongs to a specific image or music category. This means that some hints in all possible image and music categories may be introduced during training. Therefore, few shot learning is more relevant in this case as deeper investigation in zero shot learning is required.

During training, MuVis learnt to generalise the seen image-music pairs. With these information, the model is able to predict and inference unseen data accurately even though limited data examples are introduced. From Table III, the accuracy score of 0.8958 is achieved. Even though there remains room of improvements if compared to a different cross-modal text-image joint embedding model - CLIP [18] (0.9350), achieving a score close to 0.9 has shown that the model has an ideal performance in few shot tagging since MuVis is the first music-image joint embedding.

When it comes to the AUROC score, MuVis recorded a value of 0.8833. This is slightly better than MULAN [19] (0.8400) and close to CLIP [18] (0.8840). This has shown that the model is capable of performing tagging tasks realistically. The average precision score of MuVis is 0.8618, which is performing better than CLAP [22] (0.8260), proving that the results of the model is reliable.

2) Image Retrieval from Music Queries: On top of few shot learning, image retrieval tasks are evaluated for analysing the performance of MuVis. This involves searching for the relevant image entry from the given collection, based on the music query. The experiment is carried out such that partial and coherent music snippets (shorter clips from the full music) are used as keywords for retrieving their relevant image entries from the collection/dataset.

The results are shown in Table III, where the accuracy hits 0.8475 in this task. While it has yet to achieved a comparable accuracy to a typical joint embedding model such as text-image CLIP [18], achieving a score within the range of 0.8 to 0.9 has suggested that the model is capable of performing the image retrieval task with minimal error. Moreover, the AUROC and average precision scores are 0.8838 and 0.8978 respectively. These numbers further suggest that the model prediction is up to good standard and realistic.

3) Music Retrieval from Image Queries: Lastly, the retrieval experiment is carried out on its counter-setting. Music retrieval tasks are evaluated for analysing the robustness of MuVis in performing tasks the other way round. The model is expected to match a relevant music entry from the collection, given an image query. To best represent the scenario and differentiate between the tagging task, image crops are used as queries/keywords for searching the matching music entry.

Based on Table III, the model has successfully retrieve music through image input up to an accuracy of 0.8333. Along with the AUROC score of 0.8634 and the average precision of 0.8590, the model succeeds in the task such that high true positive rate as well as good recall and precision level are achieved in

the retrieval task. Its AUROC score is comparable to those of text-audio MULAN [19] (0.9030) joint embedding model, whereas the average precision score outperforms another typical expected performance from CLAP [22] (0.7020).

V. Music-conditioned Image Generative AI Model - MusIm

Moving forward from CV and NLP, recent works on AI audio/music generation include AudioLM [24] (audio/speech generation from text), MusicLM [1] (music generation from text), and LP-MusicCaps [25] (text generation from music). However, music-to-image generation is a new research problem. This work takes a big step forward by introducing MusIm, a music-conditioned image generation model. With inspiration from MuVis and Stable Diffusion [11], the methodology and qualitative results are presented next.

A. Methodology

1) Music Embedding Network: For embedding the music with relation to its possible relevant image, MuVis - the image-music joint embedding is used for encoding the music input. MuVis consists of two major embedding towers: a) AST [2] for audio embedding, and b) ViT [3] for image embedding. In this model, the music embedding tower of MuVis is used specifically for the music encoding task, where a hidden dimension of 768 applies with input music targeted at 16kHz.

2) Image VAE Network: VAE [9] consists of an encoder and a decoder internally. These components are applied separately at two stages [11].

- Encoder: Extraction of image embeddings.
  During training, a target image is passed such that the model aims to learn the generation pattern. The VAE encoder is used for converting the 3 x 512 x 512 target image into its equivalent latents. Random noise are then added before passing the latents into UNet for the denoising phase.
- Decoder: Translating embeddings back into image data.
  When the UNet [8] model has successfully denoise the image latents, the VAE decoder is used for converting the latents back into its equivalent human-understandable image data.

3) Diffusion Network: For learning the noise reversal process, the UNet [8] model is integrated with inspirations from Stable Diffusion [11]. As the architecture of UNet resembles the letter "U" [8], the upwards decoder part (i.e., right-half of "U") is mainly involved in the tasks of denoising and upscaling images. As there are 8 layers of upscaling blocks in the UNet configuration within MusIm, the model accepts noisy latents of dimension 3 x 64 x 64. UNet then predicts the amount of noise that is added to the input latents at each step. Once the noise is removed, the clean latents are passed into the VAE decoder to be translated back into the generated 3 x 512 x 512 image data.

4) Overview: Joining up all the pieces above, the architecture of the music-conditioned image Generative AI (MusIm) is shown in the Figure 9. With the conditions of: a) music input compulsory for both training and generation, and b) image input compulsory during training and optional in generation, MuVis and the VAE encoder respectively encodes the music and image data into their equivalent latents/embeddings representation.

With both models froze during training, UNet then proceeds to learn the noise reversal process for predicting the clean latents. During generation, random noisy latents are generated as the image. The UNet model then aims to denoise the image into its equivalent clean latent form. Finally, the clean latents are reverted into their equivalent image format.

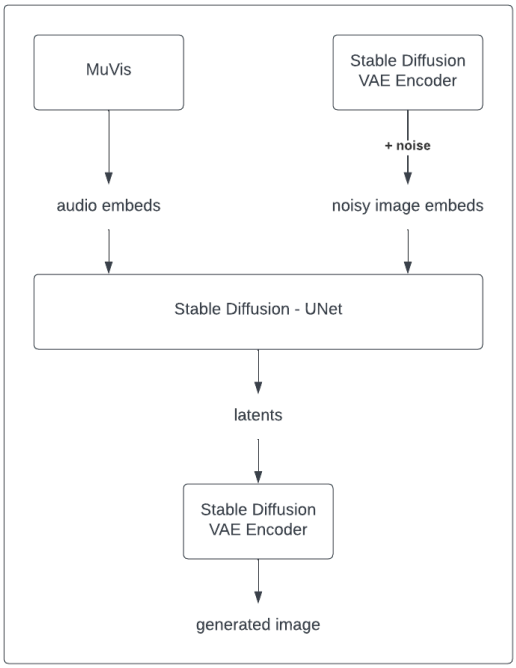

Figure 9. Architecture of the MusIm model

5) Learning: For reducing the errors in noise prediction, the mean squared error (MSE) loss is used as the learning objective of the model, which is defined in Equation 4 below. During training, random noise is added to the input images, where the UNet model learnt to reverse it. The noise prediction from UNet is compared through MSE so that the value is minimised to the actual noise.

$$MSE(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{4}$$

B. Results

Evaluating the output of the model, it is capable of generating relevant or matching images based on the input track. Some examples are shown below with qualitative discussion. As music cannot be displayed in the figure, a short text describing the style is shown instead.

Example 1:
Description of Input Music: "The music is electrifying, pulsating, buoyant, punchy, groovy, youthful and high-spirited. This music is an outdoor live performance with ambient sounds."

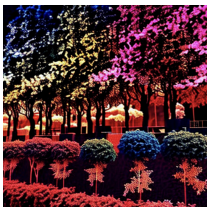

Figure 10. Image Generated for Example 1

Analysis: Figure 10 is the generated image. Bright neon colours can be seen in the generated image which matches the "electrifying", "punchy", "groovy", etc. aspects of the music. Furthermore, the garden-like setting in the generated image also corresponded to the "outdoor live performance" characteristic.

Example 2:
Description of Input Music: "It sounds relaxing and calming - like something that would put you in a zen mode."

Figure 11. Image Generated for Example 2

Analysis: Figure 11 is the generated image. When the music resembles more of a "relaxing and calming" environment, the generated image now consists of more pastel-like colours, providing soft visual effects. The Japanese style furniture and background also matches the "zen mode" stated in the description.

VI. Image-conditioned Music Generative AI Model - Imagic

Although there are many works in text-conditioned music or image generation such as StabilityAI's Stable Diffusion [11] for text-to-image generation and MusicGen [26] for text-to-music generation, there has not been a study in the feasibility of image-conditioned music generation. Therefore, Imagic is proposed in this thesis for image-to-music generation. Inspired by MusicLM's [1] and MusicGen's [26] transformer-based architectures, the methodology behind the architecture of Imagic and the respective qualitative results are described next.

A. Methodology

1) Image Embedding Network: For embedding the image with relation to its possible relevant music, MuVis - the image-music joint embedding model from Chapter 4 is used for encoding the image input. MuVis consists of two major embedding towers: a) AST [2] for audio embedding, and b) ViT [3] for image embedding. In this model, the image embedding tower of MuVis is used specifically for the image encoding task, where a hidden dimension of 768 applies with input image targeted at a resolution of 224 x 224.

2) Music Encoding Network: Inspired by the architecture of MusicGen [26], the Encodec [27] model is used for feature extraction and effective encoding of the music data. The produced music tokens are used in the training phase of the conditional transformer. Inheriting the configurations from [26] to ensure compatibility with the conditional transformer, the 32kHz variant of Encodec [27] is used for obtaining the music tokens.

3) Conditional Transformer: To learn the relationship between image embeddings and music tokens, the successful MusicGen [26] model is considered as the conditional transformer. All configurations are inherited from [26] to ensure compatibility of the different blocks in the architecture. The conditional transformer accepts image embeddings and music tokens for learning the input-output generation relationship.

4) Overview: To summarise, MuVis is first used to obtain the embeddings of the input image. To prepare for the training of Imagic, the target music is also passed into the Encodec encoder for getting the music tokens. With both components, the image embeddings are directly inputted into the MusicGen transformer encoder, where custom tokens to embeddings translation is guaranteed to avoid misinterpretation. Once the embeddings are encoded, these encoder outputs are passed into the MusicGen transformer decoder, with the target music tokens as condition. This allows the model to learn the mappings. For visualisation, the complete architecture is shown in Figure 12.

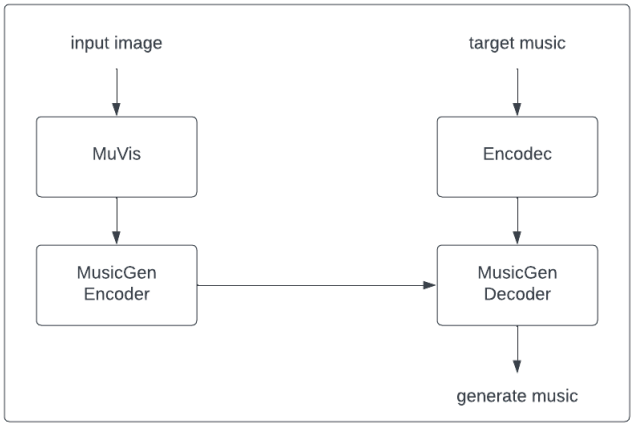

Figure 12. Architecture of the Imagic Model

5) Learning: For the task of music tokens generation, cross entropy (CE) loss is used as the learning objective for narrowing the distance between target and generated music tokens. The loss is defined in Equation 5 below. This objective is efficient for the music generation task because CE helps in optimising the logits (probabilities) prediction of each token, which is used for selecting the most suitable next-token in the generation process.

$$CE(x, y) = \{CE_1, ..., CE_N\}^T, CE_n = log \frac{exp(x_{n,y_n})}{\sum_{c=1}^{C} exp(x_{n,c})} \tag{5}$$

B. Results

Evaluating the output of the model, it is capable of generating relevant music based on the input image. An example is shown below with qualitative analysis. As music cannot be displayed in the paper, a short text describing the style is discussed instead.

Example 1:
Input Image: see Figure 13.

Figure 13. Example Input Image 1

Description of Generated Music: "The music is happy, perky, upbeat, enthusiastic, lively and spirited, which matches the positive aura of the nature."
Analysis: The theme can be described as colourful, naive and joyful, where the interaction between the birds is playful. This leads to a matching track generated since it is "upbeat, lively and spirited".

Example 2:
Input Image: see Figure 14.

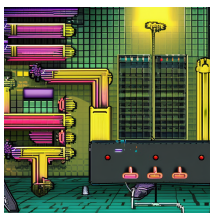

Figure 14. Example Input Image 2

Description of Generated Music: "The music features buzzy synth bass and 8 bit synth lead melody. It sounds groovy and weird, something expected as background music in games."
Analysis: The generated music resembles the environment in retro games. The 8-bit or 16-bit like image yields a generated music with "8-bit synth melody" and "buzzy synth bass", which is a match to the theme of the image.

VII. Discussion

The underlying contributions and limitations are discussed in this section. This will help in understanding the possible pros and cons for planning the related future work and exploration.

A. Image-Music Dataset - ImMuTe

1) Contributions:

i. New joint image-music-text dataset: When it comes to collecting real world data on music-image pairs, it can be difficult due to representation and interpretation differences. For example, the ambiguity of music understanding results in a many-to-many music to image mapping and vice versa. While there are many suitable images for a particular music clip, ImMuTe attempts to solve this with image Generative AI - Stable Diffusion 2. Being a partial pseudo-dataset, the data collection process is eased while still able to provide a suitable dataset for training and inference.

ii. Drive music-image related Generative AI research: As the focus and popularity of music generation grows, ImMuTe can be a starting point for music-image multimodal training. This could help in the related Generative AI research as it provides input for training and inference. For producing a proof-of-concept, ImMuTe could be used on the related SOTA models research and production.

2) Limitations:

i. High resemblance: Similar photos may be generated across several music-text pairs due to the some descriptions being semantically similar or equivalent. For example, when a caption has several similar keywords, the generation might not show large and significant differences.

Example 1:
Text: "This clip features a male voice narrating about guitars. In the background, an acoustic guitar is strumming chords and picking notes used as fills. The mood of the song is happy and up-tempo. There are no other instruments in this song. This song can be played in a romantic mood. This audio can be used in a guitar promotional video."
Images Generated: see Figure 15 below.

Figure 15. High Resemblance Example 1

Example 2:
Text: "The Rock song features a passionate male vocalist singing over wide electric guitar melodies, groovy bass guitar, shimmering cymbals, groovy toms, punchy snare and kick hits. It sounds energetic, passionate, emotional and loud."
Images Generated: see Figure 16 below.

Figure 16. High Resemblance Example 2

ii. Subjective interpretation: Differences in human knowledge and understanding [28] may also lead to certain level of bias. For instance, a professional artist may think that the paired photos is not up to par with the caption whereas an art enthusiast may feel that the photo style is relevant to the attached captions. In some cases, the generated image may be relevant to the text input in an abstract way, which depends on personal interpretation.

Example:
Text: "This song contains a bit-rate lo-fi melody in a poor audio-quality. This song may be playing in an old 2d video-game."
Images Generated: see Figure 17 below.

Figure 17. Subjective Interpretation Example

3) Ethics:

i. Copyright: Due to images being generated from the Stable Diffusion model, copyright is identical to the statement, where any intellectual property rights are waived. All data can be used by anyone for any purpose.

ii. Privacy: Due to the base dataset being MusicCaps, all data are music related and should not contain any sensitive information other than song or audio related descriptions.

iii. License: MIT License applies on the dataset, where it is free to be used, copied, modified, published, distributed, etc.

B. Image-Music Joint Embedding Model - MuVis

1) Contributions:

i. First Image-Music Joint Embedding Model: There have been previous research in text-image (e.g., CLIP [18]) and text-audio (e.g., MULAN [19]) embedding models, but the image-music joint embedding model - MuVis introduced in this work is a new exploration. This work studies the original research in multimodal image-music joint embedding for effective learning of their representation and relationship.

ii. Insights to Image-Music Cross-Modal Capabilities: From the experiments described in the last section, it is shown that MuVis is capable of performing music-image cross tagging, music retrieval, and image retrieval tasks well. These showed that given minimal information and hints, the model is able to related image-music relationship, and further utilise the knowledge for achieving different objectives. For instance, future study on image-music generative tasks are now possible with the joint embedding model for encoding the cross-modal inputs realistically and efficiently.

2) Limitations:

i. Zero Shot Learning: Due to the many-to-many mappings between music and image pairs as described in Section 4.4.1, it is believed that some hints are introduced to all available categories or patterns of music and image data during training. Although it is shown that few shot learning is successful in MuVis, there still lacks evidence to conclude that MuVis is also truly capable of zero shot learning. This is an experiment to be investigated further in the future for concluding that the model is capable of predicting data which are completely unseen before.

ii. Small Training Dataset: As ImMuTe is the only dataset with image-music pairs, this is the only dataset used for training and evaluation in this study. The data pairs present in the dataset is a total of 5521 data pairs, which is relatively small if compared to other music-related dataset such as AudioSet [6] or ArtCap [29]. Although results in Section 4.4 has proven the capability of the model, further experiments and training could be done for pushing the limits of the model and achieving the maximum performance.

C. Music-conditioned Image Generative AI Model - MusIm

1) Contributions:

i. Open Source Music-to-Image Generative AI: There have been previous research in text-to-image (e.g., CLIPGen [14], DALL-E [13], etc.), text-to-audio (e.g., AudioLM [24]) and text-to-music (e.g., MusicLM [30], MusicGen [26], etc.) generations, but the music-to-image generative model - MusIm introduced in this work is a new study. This work investigates the original research in multimodal music-to-image generations for effective learning of their relationship and generate high resolution image of high relevance to its input music.

ii. Diverse Styles of Generated Image: With a reasonable FID score of 26.4 achieved, a good diversity is present in MusIm such that it is able to generate images of various themes for accommodating different music categories. This is advantageous such that it is unlikely that a "rock" music will have a similar output to a "melancholic" music, where the first should have darker colours and the latter should have blue or emotional tones. More examples of generated images are also shown in Figures 19 and 20.

Figure 18. Generated image for music described by: "A passionate melody with backup singers in vocal harmony. It has the vibe of a retro disco and/or R&B love song."

Figure 19. Generated image for music described by: "A simple hip hop groove along with e-bass and e-guitars, the song may be played while having a BBQ in the garden with your friends."

2) Limitations:

i. Slow Generation Speed: As a diffusion-based approach is applied for the architecture of MusIm, there exists some limitations in the speed of music generation although the generation quality is good. Due to the multi-step noise reversal process for image generation, latency is introduced. Although the issue is minimised through the approach of denoising a low resolution image then up-sampling it to become high resolution, the slow-down in the generation process remains. Therefore, relevant future work will be insightful for overcoming the limitation.

ii. Non-Support of Optional Image-guided Generation: At the current phase, MusIm only supports music-conditioned image generation. If a user wished to condition their desired image generation with an extra image sample on top of the compulsory music sample, the feature is not available in MusIm. However, this could be a useful feature such that the user can obtain feedback and/or inspiration if they already have a ready-made image and wish to improve it.

3) Ethics: MusIm aims to contribute in inspiring and adding creativity to the relevant fields. However, the misuse of Generative AIs remains an open ethical issue in the ML/AI field. With proper dataset and training, there is a high possibility such that the model is fine-tuned and used for generating inappropriate

images and spreading false information [4, 31]. More ethical concerns in Generative AIs are discussed in detailed in [31].

D. Image-conditioned Music Generative AI Model - Imagic

1) Contributions:

i. Open Source Image-to-Music Generative AI: There have been previous research in text-to-image (e.g., CLIPGen [14], DALL-E [13], etc.) and text-to-music (e.g., MusicLM [30], MusicGen [26], etc.) generations, but the image-to-music generative model - Imagic introduced in this work is a new study. This work investigates the original research in multimodal image-to-music generations for effective learning of their relationship and generate music of high musical coherence and relevance to its input image.

ii. Diverse Themes and Good Coherency of Generated Music: With a reasonable FAD score of 4.6 achieved, a good diversity is present in Imagic such that it is able to generate music of various styles and categories for accommodating different image themes. This is a significant contribution such that this image-conditioned Generative AI is a first to be able to model the relationship between image input and music output for generation purposes.

2) Limitations:

i. Trade-offs in Generation Speed vs. Quality: As the transformer-based approach is taken for the architecture design of Imagic, there exists some limitation in the quality of the generated music even though the generation speed is quick. Although good musical coherence is achieved such that the generated music sounds natural and logical, there remains room for improvements for the music quality and resolution. A possible solution may be approaching through diffusion-based architectures [32].

ii. Non-Support of Optional Music-guided Generation: At the current phase, Imagic only supports image-conditioned music generation. If a user wished to condition their desired music generation with an extra music sample on top of the compulsory image sample, the feature is not available yet. However, this could be a useful feature such that the user can get inspiration if they already have a ready-made instrumental and wish to improve it or add creativity.

3) Ethics: Although Generative AIs have opened opportunities to better creativity and production costs, they have also introduced challenges to the involved industry [26]. With Imagic being completely open-source, equal access is available to all users and developers. Instead of being a competition, Imagic is hoped to give inspirations instead through the free license applied to all generated music.

VIII. Conclusion and Future Work

To sum up, although several limitations exist, the outcomes of the research (i.e., ImMuTe, MuVis, MusIm and Imagic) can help in bridging the gap for researching better multimodal music-to-image or image-to-music generation. When the progress of the field is more mature, the dataset can be further extended or updated to be more accurate and reliable. Some of the challenges and possible approach of overcoming them are listed below.

1) Pretraining / Fine-tuning Phase: Currently, the Stable Diffusion 2 model is used directly with the pretrained checkpoints, i.e., `stable-diffusion-2-1` and `stable-diffusion-2-1-base`. This is due to limited high-quality dataset for fine tuning it to be targeted at music-related captions. Thus, the proven successful official checkpoints are used for the generation. In the future, when there are more music text-image dataset, the base checkpoint could be further fine-tuned and trained for more specific purposes, for instance, targeted at art generation or cartoon generation. This could then be used for fine-tuning for more specific types of learning of the embedding model. Then, comparisons and regeneration could be done for minimising the disadvantages and limitations discussed in the previous section.

2) Further Investigation into Zero Shot Learning: As explained in the discussion of the MuVis image-music joint embedding model, it is believed that the many-to-many relationship between image and music data may give out information on all possible combinations and representations during training. It remains unclear if MuVis is truly outperforming in zero shot learning. Thus, future work can dive deeper into this aspect and gather more evidence between concluding the statement.

3) Speed-Up in Model Performance of MusIm: Due to the nature and working principles of diffusion-based models, the speed of model training and music generation is considered slow without hardware optimisation (e.g., performing tasks on GPU or TPU). Even though the quality of the generated image is outperforming other neural network architectures (e.g., GANs [33, 15, 12], transformers [5, 14], etc.), future work related to exploring speed up techniques, which is the major disadvantage in diffusion model [4, 11], will be a great advancement in the field.

4) Explore Support for Optional Music Sample as an Additional Condition in Imagic: As pointed out in Chapter 6, even though image-conditioned music generation is well supported, an extra feature of music sample as an optional condition could be useful for the user to explore possibilities based on their existing work. This could be an insightful and useful future work with references to some existing multimodal Generative AI such as MusicGen [26] and MusicLM [1] which supports optional conditioning of the output data type.

REFERENCES

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

[2] Yuan Gong, Yu-An Chung, and James Glass. "Ast: Audio spectrogram transformer". In: *arXiv preprint arXiv:2104.01778* (2021).

[3] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[4] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[5] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[6] Jort F Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

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

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

[9] George Lawton. *Generative models: Vaes, gans, diffusion, transformers, nerfs*. Apr. 2023. URL: https://www.techtarget.com/searchenterpriseai/tip/Generative-models-VAEs-GANs-diffusion-transformers-NeRFs#:~:text=Diffusion%20models%20add%20and%20then,text%20summarization%20and%20image%20creation..

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

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

[12] Shengyu Zhao et al. "Large scale image completion via co-modulated generative adversarial networks". In: *arXiv preprint arXiv:2103.10428* (2021).

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

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

[15] Tao Xu et al. "Attngan: Fine-grained text to image generation with attentional generative adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1316–1324.

[16] Ming Ding et al. "Cogview: Mastering text-to-image generation via transformers". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 19822–19835.

[17] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

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

[19] Qingqing Huang et al. “Mulan: A joint embedding of music audio and natural language”. In: *arXiv preprint arXiv:2208.12415* (2022).
[20] Adam Ek and Nikolai Ilinykh. “Vector Norms as an Approximation of Syntactic Complexity”. In: *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*. 2023, pp. 121–131.
[21] *AudioSet: A large-scale dataset of manually annotated audio events*. 2017. URL: https://research.google.com/audioset/.
[22] 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.
[23] Ethan Zhang and Yi Zhang. “Average precision”. In: *Encyclopedia of database systems* (2009), pp. 192–193.
[24] Zalán Borsos et al. “Audiolm: a language modeling approach to audio generation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).
[25] SeungHeon Doh et al. “Lp-musiccaps: Llm-based pseudo music captioning”. In: *arXiv preprint arXiv:2307.16372* (2023).
[26] Jade Copet et al. “Simple and Controllable Music Generation”. In: *arXiv preprint arXiv:2306.05284* (2023).
[27] Alexandre Défossez et al. “High fidelity neural audio compression”. In: *arXiv preprint arXiv:2210.13438* (2022).
[28] Michelle C Baddeley, Andrew Curtis, and Rachel Wood. “An introduction to prior information derived from probabilistic judgements: elicitation of knowledge, cognitive bias and herding”. In: *Geological Society, London, Special Publications* 239.1 (2004), pp. 15–27.
[29] Yue Lu et al. “Artcap: A dataset for image captioning of fine art paintings”. In: *IEEE Transactions on Computational Social Systems* (2022).
[30] *MusicLM: Generating Music from Text*. 2023. URL: https://google-research.github.io/seanet/musiclm/examples/.
[31] Emily Denton. “Ethical considerations of generative ai”. In: *AI for Content Creation Workshop, CVPR*. Vol. 27. 2021, p. 17.
[32] Haohe Liu et al. “Audioldm: Text-to-audio generation with latent diffusion models”. In: *arXiv preprint arXiv:2301.12503* (2023).
[33] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

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. 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.