Musipainter: A music-conditioned generative architecture for artistic image synthesis

Intelligent Systems with Applications 29 (2026) 200611

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Musipainter: A music-conditioned generative architecture for artistic image synthesis

Alfredo Baione [a,b], Giuseppe Rizzo [b], Luca Barco [b], Angelica Urbanelli [b], Luigi Di Biasi [a], Genoveffa Tortora [a,*]

[a] University of Salerno, Via Giovanni Paolo II, 132, Fisciano, 84084, Salerno, Italy
[b] LINKS Foundation, Via Pier Carlo Boggio, Torino, 10138, Italy

ARTICLE INFO

Keywords:
Generative art
Cross-modal learning
AI-driven creativity
Music-to-image

ABSTRACT

Generative art is a challenging area of research in deep generative modeling. Exploring AI's role in human–machine co-creative processes requires understanding machine learning's potential in the arts. Building on this premise, this paper presents Musipainter, a cross-modal generative framework adapted to create artistic images that are historically and stylistically aligned with 30-second musical inputs, with a focus on creative and semantic coherence. To support this goal, we introduce Museart, a dataset designed explicitly for this research, and GIILS, a creativity-oriented metric that enables us to assess both artistic-semantic consistency and diversity in the generated outputs. The results indicate that Musipainter, supported by the Museart dataset and the exploratory GIILS metric, can offer a foundation for further research on AI's role in artistic generation, while also highlighting the need for systematic validation and future refinements.

1. Introduction

A crucial reason for the automatic analysis of artistic content in data is the advancement of human-centric intelligent systems. In this respect, significant progress has been made in the highly challenging task of emotion recognition (Xing et al., 2015; You, Luo, Jin, & Yang, 2015, 2016), particularly through cross-modal approaches involving audio-visual contents (Li, Tao, Maybank, & Yuan, 2008; Wu, Qiao, Wang, & Tang, 2016; Xing, Zhang, Zhang, Wu, Dou, & Sun, 2019). In automatic learning, the powerful tools of multimodal applications in the field of audio and visual data (Lee, Roh, Byeon, Yoon, Kim, Kim, & Kim, 2022; Ruan et al., 2023; Sung-Bin, Senocak, Ha, Owens, & Oh, 2023; Verma, Dhekane, & Guha, 2019; Wu, Xu, Qiao, & Tang, 2012) represent a solid starting point for investigating the creativity potential of automatic systems in audio-visual domains. Also, generative modeling has recently been shown to be capable of creating original and fascinating works,[1] especially in the field of figurative art (Wang, Huang, Wang, & Roy, 2023).

An essential aspect of studying creativity within automated frameworks is related to the topic of emotional semantics. While it is worth emphasizing the research conducted in this direction through the study of image-music relationships (Verma et al., 2019; Xing et al., 2015, 2019), a fully quantifiable understanding of their underlying structure remains elusive. Not to mention the many disputable arguments in favor of a pure and human-based "artificial creativity", considered a highly counterproductive approach, able to put one at risk of not making use of the nearly unlimited possibilities provided by the computing power of artificial agents (Esling & Devis, 2020). All of the issues above demonstrate that generative models have the potential to understand and transform artistic production radically. To achieve that, the use of a Denoising Diffusion Probabilistic Model (DDPM, Ho, Jain, & Abbeel, 2020; Ho, Saharia, Chan, Fleet, Norouzi, & Salimans, 2022; Nichol & Dhariwal, 2021; Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022; Saharia et al., 2022) provides several advantages: these models generate high-resolution images at a level beyond the one reached by other generative models (Saharia et al., 2022); as shown in Nichol and Dhariwal (2021), DDPMs are extremely prone to conditioning, which can be very helpful in determining dependencies between inputs and outputs; lastly, they are efficiently scalable (Rombach et al., 2022).

This paper proposes Musipainter, an architecture that uses a latent diffusion model trained for text-to-image-generation (Rombach et al., 2022) and a pre-trained audio encoder (Chen et al., 2023) to generate artistic images historically and stylistically related to music inputs.

* Corresponding author.
E-mail addresses: abaione@unisa.it (A. Baione), giuseppe.rizzo@linksfoundation.com (G. Rizzo), luca.barco@linksfoundation.com (L. Barco), angelica.urbanelli@linksfoundation.com (A. Urbanelli), ldibiasi@unisa.it (L. Di Biasi), tortora@unisa.it (G. Tortora).

[1] https://www.artbreeder.com/, https://quasimondo.com/, https://refikanadol.com/

[2] https://www.kaggle.com/datasets/alfredobaione/museart

https://doi.org/10.1016/j.iswa.2025.200611
Received 30 July 2025; Received in revised form 9 November 2025; Accepted 16 November 2025
Available online 19 November 2025

Fig. 1. Teaser: An image generated from a 30-second audio fragment of the *Fantasia on a Theme by Thomas Tallis* by Ralph Vaughan Williams (1872–1958). The associated label is *20th century traditional open view*.

The created dataset, Museart,[2] is specifically designed to enhance this input–output relation. In fact, it contains 24216 instances of each type of datum (music and image), organized into 18 labeled categories according to semantic, cultural, and historical features. To our knowledge, prior works investigating the relations among artistic data (Qiu & Kataoka, 2018; Verma et al., 2019; Wu, Seetharaman, Kumar, & Bello, 2022) do not rely on such a uniquely structured database as Museart. For instance, they cannot provide specific historical input–output relations. The comparison of the Audio-Image Similarity (AIS) metric between Musipainter and Wav2CLIP (Wu et al., 2022) shows that Musipainter outperforms Wav2CLIP in its ability to associate music with images. Additionally, a music–image association survey based on Qiu and Kataoka's approach (Qiu & Kataoka, 2018) yields superior results, paving the way for further studies in this direction. Finally, a customized objective metric, named GIILS (Generated Image–Image Label Similarity), is proposed as a creativity metric that aims to evaluate the semantic similarity among generated images obtained from music inputs belonging to the same category. This paper provides also qualitative results obtained with music inputs taken inside and outside the proposed dataset (see Fig. 1). The structure of the paper is the following: in Section 2, the state-of-the-art is presented; then, in Section 3, the adopted research methodology is described; then analysis and results are discussed in Section 4; finally, in Section 6, conclusions are drawn together with potential directions of future works.

2. State of the art

The generative art field includes various works investigating models' potential on music data. In this regard, it is worth mentioning the survey realized by Ma et al. (2024), in which the state-of-the-art pre-trained models and foundation models for music are examined. In Lee et al. (2022), Qiu and Kataoka (2018), Saari, Eerola, and Lartillot (2011), Verma et al. (2019), Xing et al. (2015, 2019) authors focus on the possibility of establishing different kinds of correlations between audio and visual data. In particular, in Verma et al. (2019) the problem of learning affective correspondences between music and images is introduced. For this task, a music clip and an image are similar (having accurate correspondence) if they have identical emotional content. To estimate this crossmodal, emotion-centric similarity, authors propose a deep neural network architecture that learns to project the data from the two modalities into a common representation space. Then, the network performs a binary classification task to predict the affective correspondence (*true* or *false*). The proposed approach achieves 61.67% accuracy for the affective correspondence prediction task on a customized database containing more than 3500 music clips and 85000 images with three emotion classes (*positive*, *neutral* and *negative*).

In Wu et al. (2016) and Wu et al. (2012), authors explore the topic of *cross matching* between music and image. In particular, in Wu

et al. (2016), they aim to understand whether machines can automatically match music and images. They construct a benchmark dataset of over 45000 music–image pairs and recruit human labelers to annotate whether these pairs are well-matched. Results show that labelers generally agree on the matching degree of music–image pairs. Secondly, a suitable semantic representation of music and image for this cross-modal matching task is investigated: authors adopt lyrics as a middle media to connect music and image and design a set of lyrics-based attributes for image representation. Ultimately, they propose a cross-modal ranking analysis (CMRA) to learn the semantic similarity between music and image, outperforming state-of-the-art cross-modal methods in the music–image matching task.

Works like Ge et al. (2022), Ruan et al. (2023) delve very deeply into the generative part of visual and musical data. Indeed, in Ruan et al. (2023), the authors propose an audio–video generation framework that simultaneously enhances the engagement of viewing and listening experiences toward high-quality realistic videos. To generate joint audio–video pairs, a novel Multi-Modal Diffusion model is leveraged with two-coupled denoising autoencoders. Extensive experiments show auspicious results in unconditional audio–video and zero-shot conditional tasks (e.g., video-to-audio). In this respect, another significant work dealing with zero-shot learning (ZSL) of music and images is the one by Wu et al. (2022). Here, the authors propose a methodology named Wav2CLIP, which can learn audio representations thanks to their projection into a shared embedding space with images and text. This approach enables multimodal applications such as zero-shot classification and cross-modal retrieval.

Even if not focused on the use of music data, the research conducted by Yariv et al. in Yariv, Gat, Wolf, Adi, and Schwartz (2023) proposes a framework for audio-to-image generation: given an audio sample containing an arbitrary sound, authors aim to generate a high-quality image representing the acoustic scene. Specifically, they propose learning a dedicated audio token to map the audio representation into an embedding vector. Such a vector is then forwarded into the network as a continuous representation, reflecting a new word embedding. In this way, authors can investigate the feasibility of directly encoding any audio signal into a dedicated representation, producing an additional token for text-conditioned image generation.

Finally, among many papers inspired by the work on contrastive language–image pre-training (CLIP, Radford et al. (2021)), it is worth mentioning the one carried out by Guzhov, Raue, Hees, and Dengel (2022), for its contribution to the improvement of sounds classification. In particular, here, authors present an extension of the CLIP model that handles audio in addition to text and images: a model that incorporates the ESResNeXt audio model (Guzhov, Raue, Hees, & Dengel, 2021) into the CLIP framework using the AudioSet dataset.[3]

3. The research methodology

This section includes: the data preparation, the description of the model, the audio encoding phase, the optimization choices, and the evaluation functions adopted.

3.1. Data preparation

First of all, we created a specific music–image dataset, named Museart,[4] in order to accomplish the artistic image generation task from music inputs. Images belonging to Museart are in JPEG format and come from five sources, i.e., *WikiArt*,[5] *A Refined WikiArt Dataset*,[6] *Best Artworks of All Time*,[7] *Classical Asian Art*,[8] and *Pinterest*.[9] The audio files inside Museart are 30 s long, in WAV format, and originate from high quality YouTube[10] music clips. Specifically, the dataset is organized into 18 labeled categories, each containing an equal number of images and audio files, classified according to expert-validated historical and stylistic criteria (see Fig. 2). Overall, 24216 potential music–image pairs have been collected in Museart, with the training, validation, and test sets consisting of 19366, 2425, and 2425 instances, respectively. An image–audio pair from the Museart dataset is created as follows: for each audio file of the dataset, its corresponding image is randomly sampled from those belonging to the same category of the audio file.

3.2. Description of the model

The model adopted to generate artistic images from music inputs relies on an audio-to-image generation method (Yariv et al., 2023) that leverages a pre-trained text-to-image diffusion model (Rombach et al., 2022), together with a pre-trained audio encoder (Chen et al., 2023). While the purpose of Yariv et al. (2023) is to generate images taken from video frames that match specific input sounds as closely as possible, i.e., a sound recognition task, here the goal is to produce artistic images that are coherent with music inputs, i.e., a co-creative artistic task for humans. This is why the construction of Museart became necessary in this context. Not to mention the clear difference in the image–audio pair construction: in Yariv et al. (2023), each video frame is sampled from a clip whose audio file is exactly the input to the method. Another substantial difference between Musipainter and Yariv et al. (2023) lies in the training pipeline. Unlike Yariv et al. (2023), Musipainter provides a contextual training–validation phase, allowing the selection of the most suitable minimum point with respect to both the training and validation loss.

The learning process of the pre-trained diffusion is realized on a latent representation of an encoder–decoder architecture (Rombach et al., 2022). The input of the method is a pair (i, a), where i represents an artistic image and a represents its corresponding music sample. The final purpose is to create a generative process that is audio-conditioned, i.e., that is able to learn $p(i|a)$. To achieve this, given that a text-conditioned generative model is being used, the audio signal a has to be associated with a text conditioning. This is accomplished by concatenating two embedded representations. Differently from Yariv et al. (2023), the first is obtained by encoding the prompt "An art image of" using a Transformer, resulting in a representation $e_{text} \in \mathbb{R}^{4\times d_a}$, where d_a is the embedding dimension of the input text. The second one is a latent representation of the audio signal, denoted as e_{audio}, where $e_{audio} \in \mathbb{R}^{d_a}$.

3.3. Audio encoding

To obtain e_{audio}, an Embedder, composed of a pre-trained audio encoding network (Chen et al., 2023) and a small projection network (see Fig. 3), is adopted. This results in $e_{audio} = \text{Embedder}(a)$. The audio encoder leverages a pre-trained audio classification network ϕ to represent the audio. In particular, to avoid the storage of redundant information, and to emphasize the discriminative ability of the network, following Chen et al. (2023), a concatenation of earlier layers and the last hidden layer (specifically the 4th, 8th, and 12th out of a total of 12) of this network is used to produce a temporal embedding of the audio $\phi(a) \in \mathbb{R}^{\hat{d}\times n_a}$. Here, n_a is the temporal audio dimension and $\hat{d}$ is the new dimension obtained from the concatenation. Then, to learn a projection into the textual embedding space, $\phi(a)$ is forwarded in two linear layers,

[3] https://research.google.com/audioset/dataset/index.html
[4] https://www.kaggle.com/datasets/alfredobaione/museart
[5] https://www.wikiart.org/
[6] https://www.kaggle.com/datasets/trungit/wikiart25k?rvi=1
[7] https://www.kaggle.com/datasets/ikarus777/best-artworks-of-all-time
[8] https://www.kaggle.com/datasets/ajrhee/classical-asian-art/data
[9] https://www.pinterest.it/
[10] https://www.youtube.com/

Fig. 2. Graphic representation of the Museart dataset labels.

$W_1 \in \mathbb{R}^{\hat{d} \times \hat{d}}$ and $W_2 \in \mathbb{R}^{\hat{d} \times d_{audio}}$, with a GELU non-linear function σ between them, i.e., $\bar{e}_{audio} = W_2\sigma(W_1\phi(a))$. Finally, an attentive pooling layer is applied to a sequence of 248, reducing the temporal dimension of the audio signal, i.e., $e_{audio} = \text{Atten-Pooling}(\bar{e}_{audio})$. Fig. 4 shows the entire architecture adopted. The choice of using a text embedding space for projecting the audio representation is related to the semantic task of this research. In fact, while image embeddings capture visual style and texture, text embeddings allow for more abstract and categorical associations that are particularly relevant for describing the music effectively (e.g., melancholic, majestic, romantic, etc.). Nowadays, this approach is common and relies on a solid body of scientific literature (Doh, Won, Choi, & Nam, 2023; Liu, Hussain, Sun, & Shan, 2024; Melechovskỳ, Guo, Ghosal, Majumder, Herremans, & Poria, 2024; Schneider, Kamal, Jin, & Schölkopf, 2024).

3.4. Optimization

During the optimization process (i.e., the training phase), the loss adopted is an additional one that complements the typical loss adopted for a latent diffusion model (LDM),i.e.,

$$\mathcal{L}_{\text{LDM}} \triangleq \mathbb{E}_{x\sim S, t\sim U(0,1), \epsilon\sim N(0,\mathbb{I})}\left[\|\epsilon - \epsilon_\theta(f(x_t), t)\|_2^2\right]. \quad (1)$$

In this case $t \in [0, 1]$ is a timestamp defined over a specified reverse Markov process of length T applied to the data x, and $\epsilon_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is a denoising function that, targeting the noise ϵ added during the forward process, learns predicting a clean version of the perturbed x_t from the training distribution $S = \{x_1, \ldots, x_T\}$. Finally, f is the encoder where the latent diffusion operates. In any case, a complement of Eq. (1) is necessary to encode the label of the image, denoted by $l \in \mathbb{R}^{n_l \times d_a}$, with n_l representing the label's length. The label is encoded using the generative model's textual encoder, and then its spatial dimension is

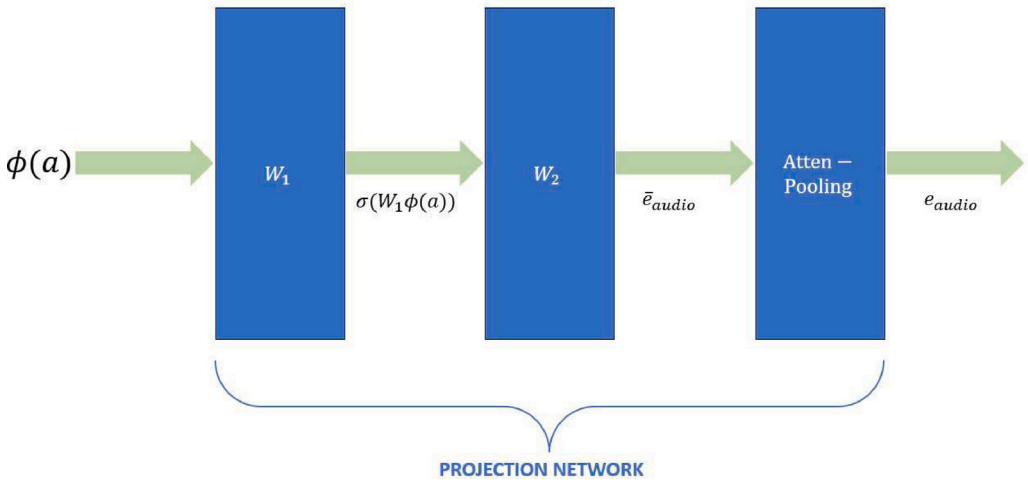

Fig. 3. Schematic representation of the projection network inside the Embedder.

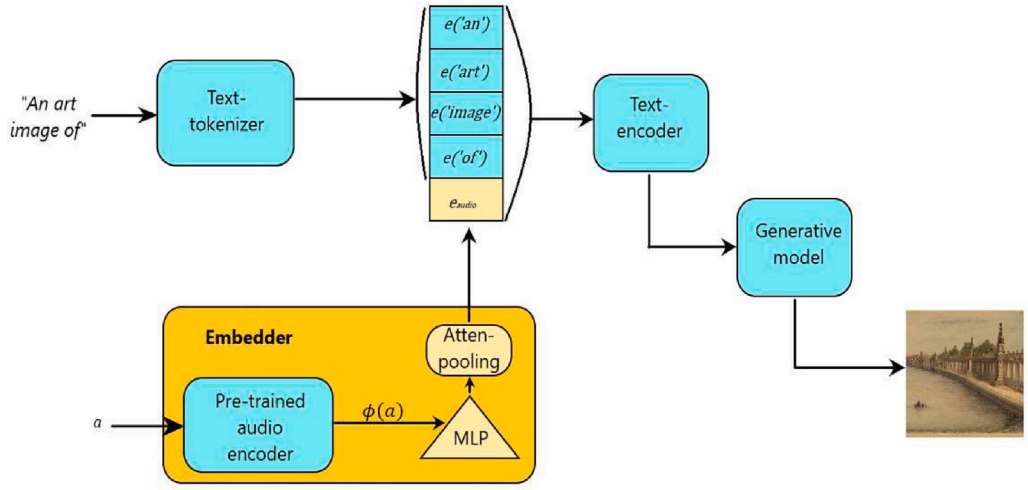

Fig. 4. Architecture overview: a 30-second music sample is forwarded through a pre-trained audio encoder and then through a projection network. A pre-trained text encoder extracts tokens from the vector representations of the tokenized prompt and the audio. Finally, the generative model is fed with the concatenated representations of these tokens.

reduced using average pooling, i.e., $\hat{l} = \text{Avg-Pooling}(l)$. Now, it is possible to introduce a classification loss, defined as follows:

$$\mathcal{L}_{\text{CL}} = \left(1 - \frac{\langle e_{audio}, \hat{l} \rangle}{\|e_{audio}\| \|\hat{l}\|}\right)^2. \tag{2}$$

Intuitively, this term ensures that the audio embedding remains close to the image's concept, facilitating faster and stabler convergence. Adding an l_1 norm regularization term for e_{audio} to encourage sparsity, the overall loss that is optimized is given by

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda_{l_1} \|e_{audio}\|_1 + \lambda_{\text{CL}} \mathcal{L}_{\text{CL}} \tag{3}$$

with λ_{l_1} and λ_{CL} tunable parameters.

3.5. Evaluation functions

3.5.1. Objective evaluations

The objective evaluation functions used in this study include three metrics taken from the literature (AIS, IIS, and FID) (Yariv et al., 2023) and a new metric, named GIILS, proposed in this paper. Their descriptions are as follows:

- **Audio-Image Similarity (AIS)**, that provides a measure of the model's ability to semantically relate a generated image and its corresponding input audio. To rely on a common representation space, Wav2CLIP model (Wu et al., 2022) is employed. The Wav2CLIP model enables to measure the similarity between representations of an audio and image pair. This allows to quantify to which extent the generated image describes the audio. In this regard, the similarity between two embedded vectors, namely u and v, can be expressed as:

$$\cos(\theta) = \frac{\langle u, v \rangle}{\|u\| \|v\|} \tag{4}$$

where θ is the angle between the two vectors. This measure is also known as *cosine similarity*. The AIS measure is, then, obtained as the mean percentage of the distances between similarities computed with generated images and their respective audio inputs, and mean similarities computed with the same generated images and all audios in the test set. Formally:

$$\text{AIS} = \left(1 - \frac{\cos^{-1}\left(\sum_{i=1}^{G}\left(\frac{\text{sim}(i_g, a_i) - \sum_{j\neq i, j=1}^{T}\left(\frac{\text{sim}(i_g, a_j)}{T}\right)}{G}\right)\right)}{\pi}\right) \cdot 100 \tag{5}$$

where G and T are, respectively, the number of generated images and the number of audio file in the test set; $\text{sim}(i_g, a_i)$ is the similarity computed between the generated image i_g and its corresponding music input a_i, and $\text{sim}(i_g, a_j)$ is the similarity computed between the generated image i_g and a generic music input inside the test set;

- **Image–image similarity (IIS)**, that measures the semantic similarity between the generated image and the ground truth one (i.e., the one associated with the music input adopted to generate the image). Again, the same reference-based method used to compute the AIS metric is employed. The only difference is that, in this case, music inputs and all audios in the test set are replaced, respectively, with the ground truth images and all images in the test set. Formally:

$$\text{IIS} = \left(1 - \frac{\cos^{-1}\left(\sum_{i=1}^{G} \left(\frac{\text{sim}(i_g,i) - \sum_{j\neq i, j=1}^{T} \left(\frac{\text{sim}(i_g,j)}{T} \right)}{G} \right) \right)}{\pi} \right) \cdot 100 \tag{6}$$

where T, now, is the number of images in the test set, i is the ground truth image relative to the generated image i_g, and j is a generic image inside the test set;

- **Fréchet Inception Distance (FID)**, that compares the distribution of the generated images against the original ones using an internal representation obtained from a pre-trained model (Seitzer, 2020). It is a standard score, introduced for the first time by Heusel, Ramsauer, Unterthiner, Nessler, and Hochreiter (2017) that, essentially, measures the quality of the generated images. For any two probability distributions, μ and ν over $\mathbb{R}^n$, having finite mean and variance, their Fréchet distance is:

$$d_F(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2 d\gamma(x,y) \right)^{\frac{1}{2}} \tag{7}$$

where $\Gamma(\mu,\nu)$ is the set of all measures on $\mathbb{R}^n \times \mathbb{R}^n$ with marginal distributions μ and ν, respectively, on the first and second factors (in other words, the FID measure is the 2-Wasserstein distance on $\mathbb{R}^n$).

- **Generated Image–Image Label Similarity (GIILS)**, a customized metric that provides a measure of the model's ability to semantically relate generated images obtained from music inputs belonging to the same category. It relies on the same similarity space used to compute AIS and IIS, and applied to the generated images obtained from music inputs belonging to the same category. The internal mean, thus, is computed with respect to the generated images space. Formally:

$$\text{GIILS} = \left(1 - \frac{\cos^{-1}\left(\sum_{(i_{g_L}, j_{g_L})} \left(\frac{\text{sim}(i_{g_L}, j_{g_L}) - \sum_{j_g \neq i_{g_L}} \left(\frac{\text{sim}(i_{g_L}, j_g)}{G} \right)}{\#(i_{g_L}, j_{g_L})} \right) \right)}{\pi} \right) \cdot 100 \tag{8}$$

where $\#(i_{g_L}, j_{g_L})$ is the number of pairs of generated images from music inputs having the same label L and j_g is a generic generated image.

3.5.2. Human-centered evaluations

As a further evaluation function, the same experiment conducted in Qiu and Kataoka (2018) has been adopted. This approach allows to use results obtained in Qiu and Kataoka (2018) as a baseline for comparison (see Section 4).

In order to evaluate whether the proposed method can generate images coherent with the corresponding music labels, 12 participants were asked to listen to 4 audio samples, each belonging to a different category, and to examine the 4 corresponding generated images. Without knowing the actual input associated with each image, they were instructed to match each audio sample to one image, based on their personal perception. Results of this evaluation are shown in Section 4.2.

4. Experiments and results

In this section, two models' implementations, their comparison over objective and subjective metrics, and the measures obtained with the new creativity measure (GIILS) introduced are provided. The dataset used is the customized Museart, described in Section 3.1. As far as the generative part is concerned, *CompVis/stable-diffusion-v1-4* (Rombach et al., 2022) is adopted, resulting in a 8853507 parameters model.

4.1. Models' implementation

4.1.1. Musipainter conditional model 1

In the first experiment, only the projection network is optimized, i.e., the pre-trained audio encoder network and the pre-trained text-to-image generative network remain frozen. The model is trained on a Nvidia A100 for 25 epochs, with an initial learning rate of 8e−5, a training batch size equal to 8, a validation batch size equal to 1 and a *cosine* learning rate schedule. Fig. 5 shows the training and validation losses graph. The validation loss trend is irregular and does not decrease steadily. On the other hand, although faint, the training loss shows a decreasing profile. Results of the objective evaluation functions are computed after generating 335 images from music inputs sampled inside the test set. The weights adopted to run this inference are those learned at epoch 21 because, as shown by Fig. 5, in that circumstance losses are at their lowest values.

4.1.2. Musipainter conditional model 2

In the second experiment, the optimization phase involves the projection network, as well as the pre-trained audio encoder network and the pre-trained text-to-image generative network. This new configuration of trainable parameters is motivated by the prospect of achieving improved evaluation results compared to the Musipainter conditional model 1. The model is trained on a Nvidia A100 for 50 epochs, with an initial learning rate of 1e−5, a training batch size equal to 8, a validation batch size equal to 8, and a *cosine* learning rate schedule. Fig. 6 shows the training and validation losses graph. Not taking into account the first 8 epochs, the validation loss decreases irregularly at least until epoch 36, where it reaches its minimum. On the other hand, the training loss, apart from the first 10 epochs, decreases irregularly until epoch 41. Results of the objective evaluation functions are computed after generating 335 images from music inputs sampled inside the test set. The weights adopted to run this inference are those learned at epoch 36 because, as shown by Fig. 6, in that circumstance the validation loss is at its lowest value and the training loss is almost at its lowest value.

4.2. Comparisons

4.2.1. Objective metrics comparison

Results of the objective metrics are compared with those of Wav2CLIP (Wu et al., 2022) and AudioToken (Yariv et al., 2023). Here, in fact, the same embedding procedure adopted to measure the similarities introduced in 3.5.1 is implemented over several audio-visual datasets. Table 1 contains the AIS, IIS and FID values for the images generated with the two models described in 4.1, for Wav2CLIP, and AudioToken. Table 1 shows that the proposed method outperforms Wav2CLIP in the AIS metric, while remaining competitive in terms of IIS. The FID scores for the two proposed models are significantly higher

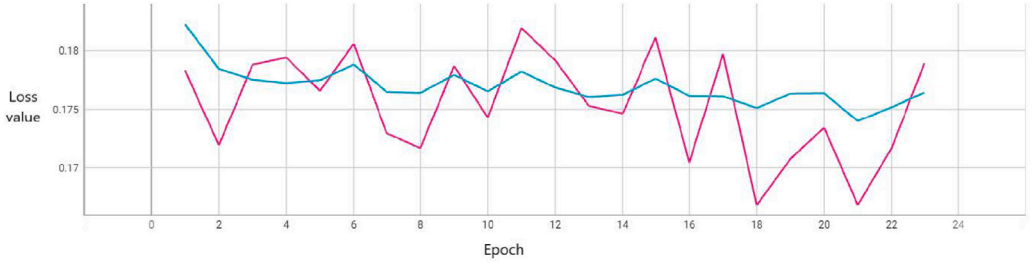

Fig. 5. Train (blue) and validation (purple) losses performance for the first experiment.

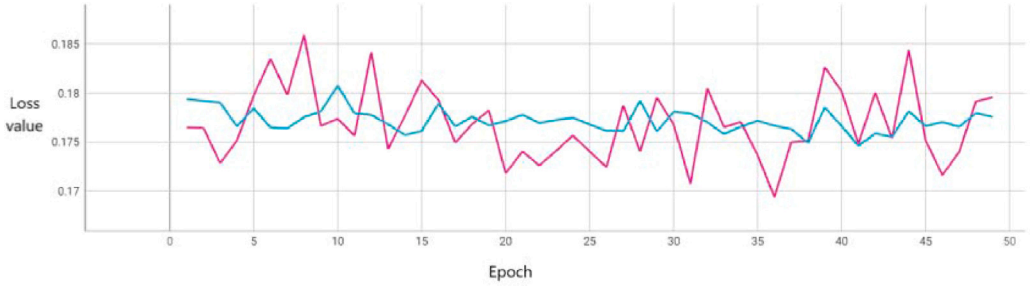

Fig. 6. Train (blue) and validation (purple) losses performance for the second experiment.

Table 1
AIS, IIS and FID values obtained with the two Musipainter conditional models, Wav2CLIP (Wu et al., 2022), and AUdioToken (Yariv et al., 2023).

MODEL	METRIC		
	AIS ↑	IIS ↑	FID ↓
Musipainter conditional model 1	50.15	49.62	179.16
Musipainter conditional model 2	49.91	50.03	191.73
Wav2CLIP (Wu et al., 2022)	47.76	51.11	99.89
AudioToken (Yariv et al., 2023)	**62.28**	**76.40**	**66.08**

than that of Wav2CLIP and AudioToken; this is somehow an expected result derived from the large variety of image and music samples belonging to the same categories inside Museart. In fact, this condition does not allow to guide the generative process toward the specific corrupted image selected together with the audio at the beginning of the process. Even if this might appear as a strong limitation of the study proposed, this metric does not represent a meaningful measure to optimize in the context of artistic associations (Jayasumana et al., 2024). The reason for this lies in the variety of semantic structures that, in an artistic context, can all be considered meaningful for explaining a "correct" association. Indeed, a piece of music will almost always evoke more than one image.

4.2.2. Human-centered comparison

The second comparison is obtained with the study conducted by Qiu and Kataoka (2018). This baseline was chosen because it is fully reproducible in the context of this research and, therefore, could effectively serve to demonstrate the relative advantages of our methodology. Table 2 contains the results of a subjective experiment conducted in Qiu and Kataoka (2018), following the methodology described by 3.5.2. On the diagonal of Table 2 there are the matches between the input audios and the corresponding generated images that the participants have correctly guessed. The other cells of Table 2 contains all the other wrong matches. In the same way, Tables 3 and 4 are constructed for the Musipainter conditional model 1 and the Musipainter conditional model 2, respectively. In this last two cases, input audios are selected inside the following 4 categories: *Baroque Classical symbolism*, *Electronic music*, *Heavy metal*, and *Asian traditional folklore*.

Table 2
Users evaluation results of the consistency of music and generated images as introduced in Qiu and Kataoka (2018).

	Sky	Water	Mountain	Desert
Sky	(9/12)	(3/12)	(0/12)	(0/12)
Water	(4/12)	(7/12)	(1/12)	(0/12)
Mountain	(1/12)	(2/12)	(9/12)	(0/12)
Desert	(1/12)	(0/12)	(0/12)	(11/12)

Table 3
Users evaluation results of the consistency of music and generated images for the Musipainter conditional model 1 with respect to 4 selected categories of Museart.

	Baroque Classical symbolism	Electronic music	Heavy metal	Asian traditional folklore
Baroque Classical symbolism	(12/12)	(0/12)	(0/12)	(0/12)
Electronic music	(0/12)	(12/12)	(0/12)	(0/12)
Heavy metal	(0/12)	(0/12)	(12/12)	(0/12)
Asian traditional folklore	(0/12)	(0/12)	(0/12)	(12/12)

Table 3 demonstrates that Musipainter conditional Model 1 achieves full alignment between its generated outputs and user selections across the categories of *Baroque Classical Symbolism*, *Electronic Music*, *Heavy Metal*, and *Asian Traditional Folklore*. In this regard, it clearly outperforms the model proposed in Qiu and Kataoka (2018), as illustrated in Table 2. Table 4 shows that Musipainter conditional Model 2 obtains optimal user alignment for *Baroque Classical Symbolism* and *Asian Traditional Folklore*. However, for the selected samples of *Electronic Music* and *Heavy Metal*, only the $58.\bar{3}\%$ of users selected the model's corresponding generation. Nevertheless, for at least two out of four categories, Musipainter conditional model 2 performs better than the baseline model from Qiu and Kataoka (2018) in the presented test.

Table 4
Users evaluation results of the consistency of music and generated images for the Musipainter conditional model 2 with respect to 4 selected categories of Museart.

	Baroque Classical symbolism	Electronic music	Heavy metal	Asian traditional folklore
Baroque Classical symbolism	(12/12)	(0/12)	(0/12)	(0/12)
Electronic music	(0/12)	(7/12)	(5/12)	(0/12)
Heavy metal	(0/12)	(5/12)	(7/12)	(0/12)
Asian traditional folklore	(0/12)	(0/12)	(0/12)	(12/12)

Table 5
GIILS values obtained with the Musipainter conditional model 1.

Music inputs label	GIILS ↑
Renaissance symbolism	58.70
Baroque Classical symbolism	57.91
Electronic music	56.12
Renaissance secular subject	54.77
20th century abstractionism	54.38
Romanticism open view	54.29
Heavy metal	53.90
African traditional folklore	53.75
20th century experimental open view	53.49
20th century traditional open view	53.41
20th century experimental figure	53.31
Asian traditional folklore	52.92
Romanticism figure	52.44
American traditional folklore	52.35
European traditional folklore	52.21
Modern instrumental music	51.87
Baroque Classical secular subject	51.56
20th century traditional figure	51.31

4.3. Creativity assessment

Tables 5 and 6 show ordered results of the GIILS measure computations with two different tuned models. The values are obtained over 10 generated images, sampled among those produced with inputs belonging to a specific category of Museart.

Figs. 7, 8, 9, 10, and 11 contain images generated with the Musipainter conditional model 1 and allow to visualize the semantic relation among outputs generated from music inputs having the same label (already quantified by the GIILS measure) and between these outputs and their input label. Fig. 12 is generated with the Musipainter conditional model 1 with a music input taken out of the Museart dataset. In this case, a brief description of the input is provided in the caption.

Figs. 13, 14, 15, 16, and 17 contain images generated with the Musipainter conditional model 2. Fig. 18 is generated with the Musipainter conditional model 2 using a music input taken out of the Museart dataset.

5. Discussion

The GIILS measure introduced in Section 3.5.1 provides insight into the creativity of the Musipainter Conditional Models 1 and 2. Following Schmidhuber's idea (Schmidhuber, 2012), according to which creativity is defined as the ability to discover new patterns in a familiar context, the similarity among outcomes generated from inputs within the same known category becomes a key factor in evaluating the creative tendencies of a deep generative model.

As shown in Tables 5 and 6, GIILS scores are significantly higher than the AIS and IIS values reported in Table 1 for almost every label. This suggests that the two models effectively capture the input music context (i.e., its label) and adapt accordingly during the inference

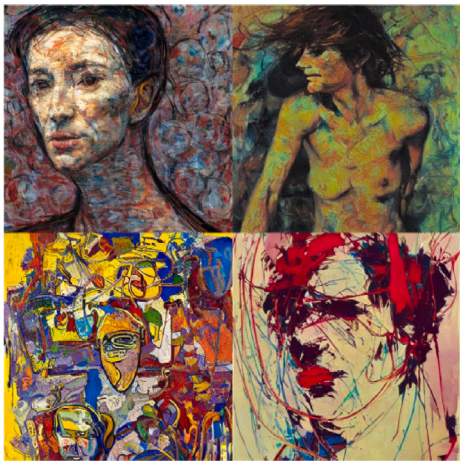

Fig. 7. Four images generated from music inputs with label *Electronic music*, using the Musipainter conditional model 1.

Fig. 8. Four images generated from music inputs with label *Renaissance secular subject*, using the Musipainter conditional model 1.

Fig. 9. Four images generated from music inputs with label *African traditional folklore*, using the Musipainter conditional model 1.

Fig. 10. Four images generated from music inputs with label *American traditional folklore*, using the Musipainter conditional model 1.

Fig. 11. Four images generated from music inputs with label *Heavy metal*, using the Musipainter conditional model 1.

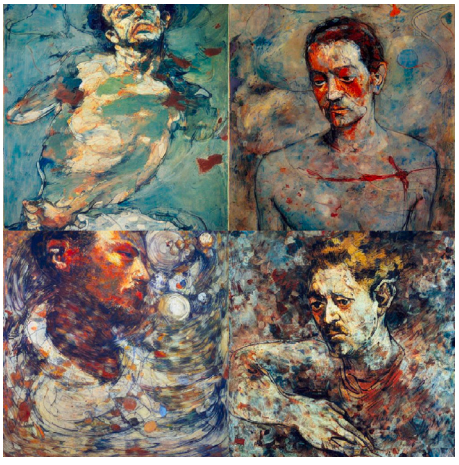

Fig. 12. Four images generated from a fragment of a nostalgic piano improvisation in jazz style, using the Musipainter conditional model 1.

Fig. 13. Four images generated from music inputs with label *Romanticism open view*, using the Musipainter conditional model 2.

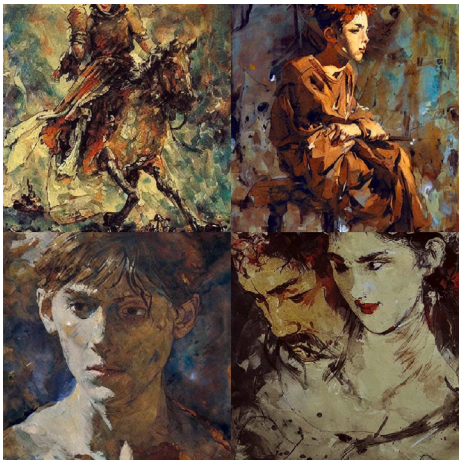

Fig. 14. Four images generated from music inputs with label *European traditional folklore*, using the Musipainter conditional model 2.

Fig. 15. Four images generated from music inputs with label *Modern instrumental music*, using the Musipainter conditional model 2.

Fig. 16. Four images generated from music inputs with label *Baroque Classical secular subject*, using the Musipainter conditional model 2.

Fig. 17. Four images generated from music inputs with label *Asian traditional folklore*, using the Musipainter conditional model 2.

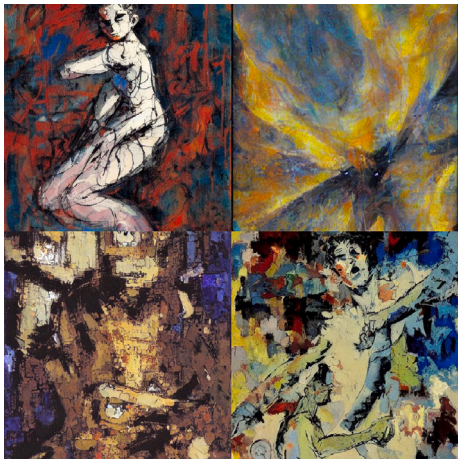

Fig. 18. Four images generated from an elaborately manipulated improvisation for electric piano and alto saxophone. The manipulation involves primarily the temporal inversion of the sounds and their cumulative distortion. The model adopted to obtain these generations is the Musipainter conditional model 2.

Table 6
GIILS values obtained with the Musipainter conditional model 2.

Music inputs label	GIILS ↑
Renaissance symbolism	58.28
Baroque Classical symbolism	57.16
Renaissance secular subject	54.85
Electronic music	53.83
Heavy metal	53.80
Romanticism figure	53.72
20th century abstractionism	53.48
20th century experimental open view	53.32
20th century experimental figure	53.12
American traditional folklore	52.86
African traditional folklore	52.75
European traditional folklore	52.00
Romanticism open view	51.97
20th century traditional figure	51.76
Asian traditional folklore	51.44
Modern instrumental music	51.35
Baroque Classical secular subject	51.29
20th century traditional open view	50.96

phase. This behavior can also be observed in several figures presented in Section 4.3. Another significative consequence of the GIILS values being in the [50.96, 58.70] interval is that, given the definition of GIILS, the two models are able to diversity the outputs generated from the same context, i.e., to show a creative tendency, quantifiable by the complementary percentage range of the GIILS interval values, i.e. 42%–50%. On the other hand, although the GIILS values in Table 6 are all lower than the corresponding values (in terms of ranking) in Table 5, the negligible differences between them do not suggest a stronger creative tendency for Musipainter Conditional Model 2 compared to Model 1. This means that no significant superior tendency to diversify outputs from inputs with the same label can be observed by Musipainter conditional model 2 with respect to model 1. This aspect, however, deserves further investigation.

6. Conclusions and further developments

This paper introduces a method for implementing a text-conditioned generative model based on music conditioning. Thanks to a structured categorization of a newly customized dataset, named Museart,[11] the method produces artistic images historically and stylistically related to the music inputs given. The proposed methodology is evaluated through a comprehensive framework that, including both subjective and objective metrics, is able to position Musipainter at the frontiers of Audio-to-Image and Image-to-Image artistic similarity detection tasks, while also aligning closely with human preferences in music–image artistic associations. From this perspective, this work reveals several interesting aspects that warrant further investigation.

First, given the complexity of the architecture, the results strongly suggest that defining an appropriate validation setup is crucial for efficiently handling hyperparameter changes during the model training phase. In this respect, a limitation of this work is the insufficient reporting on the impact of introducing an l_1 norm regularization term and a classification loss during the experimental phase. Moreover, while pretrained models provide stable outcomes and help saving computational resources, they may hinder efforts to reduce losses, especially when the trainable layers are significantly fewer than the frozen ones. Future approaches should consider a systematic investigation of trainable vs frozen parameters, as well as comparisons among different inner procedures, e.g., text vs image embedding space for projecting the audio representation, or attentive pooling vs average pooling for encoding the audio signal.

[11] https://www.kaggle.com/datasets/alfredobaione/museart

Secondly, performance analysis indicates that traditional machine learning validation techniques, such as cross-validation and bootstrap, cannot be applied to large datasets (as in generative modeling) due to computational constraints. Therefore, dimensionality reduction methods, such as feature selection, random search, and Bayesian optimization, are essential in scenarios like the one explored in this study.

Third, most critical, the depth and diversity of the semantic content in a piece of art are critical, particularly for multimodal generative learning. Our results strongly suggest that it is fundamental to structure the connections between inputs and outputs in advance. Based on these connections, subjective and objective evaluation metrics should be chosen. According to Theis, van den Oord, and Bethge (2016) and Benny, Galanti, Benaim, and Wolf (2021), it is necessary to introduce new metrics for assessing generative model performance in a class-conditional image generation context. In this regard, the proposed GILLS metric aims to serve as a cornerstone for incorporating a *creativity measure* into generative frameworks such as the one presented in this paper. Specifically, the similarity among generated images from inputs with the same label could act as a penalization factor, assessing the model's ability to diversify outputs for a specific task within predefined thresholds.

CRediT authorship contribution statement

Alfredo Baione: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Giuseppe Rizzo:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration. **Luca Barco:** Software, Validation, Resources, Writing – review & editing, Supervision. **Angelica Urbanelli:** Software, Validation, Resources, Writing – review & editing, Supervision. **Luigi Di Biasi:** Software, Resources, Writing – review & editing. **Genoveffa Tortora:** Resources, Writing – reviewing & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

This section contains some images generated with the methodology described in the Sections 3.2, 3.3, and 3.4, using music inputs taken out of the Museart dataset. The models adopted are the two Musipainter conditional models described in Section 4.1 and other two Musipainter conditional models named, respectively, Preliminary model 1 and Preliminary model 2. A brief description of the input is provided for every image shown, while the training hyperparameter configuration adopted to achieve the inference is given for Preliminary model 1 and Preliminary model 2.

Preliminary model 1

The model is trained on a Nvidia GeForce RTX 3060 for 8 epochs, with an initial learning rate of 8e−5, a training batch size equal to 1, gradient accumulation step of 4, a validation batch size equal to 1, and a *constant* learning rate schedule. Weights used for the generation are taken at epoch 8. Figs. 19 and 20 are generated with this model and show, respectively, the generations obtained from a fragment of an experimental electronic music, based on a mixture of traditional and concrete sounds, and a fragment of the *Suite hellenique*, a piece of music composed by Pedro Iturralde (Falces, July 13 1929–Madrid, November 1 2020), for piano and alto saxophone.

Fig. 19. An electronic music sample. The music is essentially experimental, obtained with a mixture of pre-sampled traditional and concrete sounds. The model adopted for the generation is the Preliminary model 1.

Fig. 20. A melodic fragment of a suite for piano and alto saxophone, named *Suite hellenique* and composed by Pedro Iturralde (Falces, July 13 1929–Madrid, November 1 2020). The model adopted for the generation is the Preliminary model 1.

Preliminary model 2

The model is trained on a Nvidia GeForce RTX 3060 for 13 epochs, with an initial learning rate of 8e−5, a training batch size equal to 1, gradient accumulation step of 4, a validation batch size equal to 1, and a *cosine* learning rate schedule. An l_2 norm regularization term equal to 0.01 for the encoded audio is added to the loss function, that becomes:

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda_{l_1} \|e_{audio}\|_1 + \lambda_{l_2} \|e_{audio}\|_2 + \lambda_{\text{CL}} \mathcal{L}_{\text{CL}}. \tag{9}$$

Weights used for the generation are taken at epoch 13. Figs. 21 and 22 are generated with this model and show, respectively, a fragment of a symphony created entirely with digital sounds, and a fragment of a digital noise/industrial music sample.

Musipainter conditional model 1

Figs. 23 and 24 are generated with the Musipainter conditional model 1 and show, respectively, four images generated from a 90's dance music fragment, and four images generated from a fragment of a canon for two violins in classical style.

Fig. 21. A fragment of a digital symphony. The model adopted for the generation is the Preliminary model 2.

Fig. 22. A digital noise/industrial music sample. The model adopted for the generation is the Preliminary model 2.

Fig. 23. Four images generated from a 90's dance music sample. The model adopted for the generations is the Musipainter conditional model 1.

Fig. 24. Four images generated from a two violins canon sample. The music style is mainly classical, and the imitation is realized in unison. The model adopted for the generations is the Musipainter conditional model 1.

Fig. 25. Four images generated from an experimental electronic piece of music titled *The Witchfinder*, composed by a duo called Amorphous Androgynous. The model adopted for the generations is the Musipainter conditional model 2.

Musipainter conditional model 2

Figs. 25 and 26 are generated with the Musipainter conditional model 2 and show, respectively, four images generated from a fragment of *The Witchfinder*, a piece of music composed by the duo Amorphous Androgynous, and four images generated from a fragment of *The Caterpillar Song*, a piece of music played in the famous Disney movie *Alice in Wonderland*.

Data availability

I have shared the data link into the manuscript.

Fig. 26. Four images generated from a fragment of *The Caterpillar Song*, a piece of music played in the famous Disney movie *Alice in Wonderland*. The model adopted for the generations is the Musipainter conditional model 2.

References

Benny, Yaniv, Galanti, Tomer, Benaim, Sagie, & Wolf, Lior (2021). Evaluation metrics for conditional image generation. *International Journal of Computer Vision*, *129*(5), 1712–1731.

Chen, Sanyuan, Wu, Yu, Wang, Chengyi, Liu, Shujie, Tompkins, Daniel, Chen, Zhuo, et al. (2023). BEATs: Audio pre-training with acoustic tokenizers. In *International conference on machine learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA*: *Vol. 202*, (pp. 5178–5193). Proceedings of machine learning research.

Doh, SeungHeon, Won, Minz, Choi, Keunwoo, & Nam, Juhan (2023). Toward universal text-to-music retrieval. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Esling, Philippe, & Devis, Ninon (2020). Creativity in the era of artificial intelligence. *CoRR*, arXiv:2008.05959.

Ge, Songwei, Hayes, Thomas, Yang, Harry, Yin, Xi, Pang, Guan, Jacobs, David, et al. (2022). Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *Computer vision - ECCV 2022 - 17th European conference, Tel Aviv, Israel, October 23–27, 2022, proceedings, part XVII*: *Vol. 13677*, (pp. 102–118). Lecture notes in computer science.

Guzhov, Andrey, Raue, Federico, Hees, Jörn, & Dengel, Andreas (2021). ESResNe(X)t-fbsp: Learning robust time-frequency transformation of audio. In *International joint conference on neural networks, IJCNN 2021, Shenzhen, China, July 18–22, 2021* (pp. 1–8).

Guzhov, Andrey, Raue, Federico, Hees, Jörn, & Dengel, Andreas (2022). Audioclip: Extending clip to image, text and audio. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2022, virtual and Singapore, 23–27 May 2022* (pp. 976–980).

Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, & Hochreiter, Sepp (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA* (pp. 6626–6637).

Ho, Jonathan, Jain, Ajay, & Abbeel, Pieter (2020). Denoising diffusion probabilistic models. In *Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, neurIPS 2020, December 6–12, 2020, virtual*.

Ho, Jonathan, Saharia, Chitwan, Chan, William, Fleet, David J., Norouzi, Mohammad, & Salimans, Tim (2022). Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, *23*, 47:1–47:33.

Jayasumana, Sadeep, Ramalingam, Srikumar, Veit, Andreas, Glasner, Daniel, Chakrabarti, Ayan, & Kumar, Sanjiv (2024). Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9307–9315).

Lee, Seung Hyun, Roh, Wonseok, Byeon, Wonmin, Yoon, Sang Ho, Kim, Chanyoung, Kim, Jinkyu, et al. (2022). Sound-guided semantic image manipulation. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022* (pp. 3367–3376).

Li, Xuelong, Tao, Dacheng, Maybank, Stephen J., & Yuan, Yuan (2008). Visual music and musical vision. *Neurocomputing*, *71*(10–12), 2023–2028.

Liu, Shansong, Hussain, Atin Sakkeer, Sun, Chenshuo, & Shan, Ying (2024). Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing* (pp. 286–290). IEEE.

Ma, Yinghao, Øland, Anders, Ragni, Anton, Sette, Bleiz Macsen Del, Saitis, Charalampos, Donahue, Chris, et al. (2024). Foundation models for music: A survey. *CoRR*.

Melechovskỳ, Jan, Guo, Zixun, Ghosal, Deepanway, Majumder, Navonil, Herremans, Dorien, & Poria, Soujanya (2024). Mustango: Toward controllable text-to-music generation. In *NAACL-HLT*.

Nichol, Alexander Quinn, & Dhariwal, Prafulla (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th international conference on machine learning, ICML 2021, 18–24 July 2021, virtual event*: *Vol. 139*, (pp. 8162–8171). Proceedings of machine learning research.

Qiu, Yue, & Kataoka, Hirokatsu (2018). Image generation associated with music data. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th international conference on machine learning, ICML 2021, 18–24 July 2021, virtual event*: *Vol. 139*, (pp. 8748–8763). Proceedings of machine learning research.

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, & Ommer, Björn (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022* (pp. 10674–10685).

Ruan, Ludan, Ma, Yiyang, Yang, Huan, He, Huiguo, Liu, Bei, Fu, Jianlong, et al. (2023). MM-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023* (pp. 10219–10228).

Saari, Pasi, Eerola, Tuomas, & Lartillot, Olivier (2011). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Speech and Audio Processing*, *19*(6), 1802–1812.

Saharia, Chitwan, Chan, William, Saxena, Saurabh, Li, Lala, Whang, Jay, Denton, Emily L., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in neural information processing systems 35: annual conference on neural information processing systems 2022, neurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022*.

Schmidhuber, Jürgen (2012). A formal theory of creativity to model the creation of art. In *Computers and creativity* (pp. 323–337). Springer.

Schneider, Flavio, Kamal, Ojasv, Jin, Zhijing, & Schölkopf, Bernhard (2024). Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 8050–8068).

Seitzer, Maximilian (2020). pytorch-fid: FID Score for PyTorch. Version 0.3.0, https://github.com/mseitzer/pytorch-fid.

Sung-Bin, Kim, Senocak, Arda, Ha, Hyunwoo, Owens, Andrew, & Oh, Tae-Hyun (2023). Sound to visual scene generation by audio-to-visual latent alignment. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023* (pp. 6430–6440).

Theis, Lucas, van den Oord, Aäron, & Bethge, Matthias (2016). A note on the evaluation of generative models. In *4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings*.

Verma, Gaurav, Dhekane, Eeshan Gunesh, & Guha, Tanaya (2019). Learning affective correspondence between music and image. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2019, Brighton, United Kingdom, May 12–17, 2019* (pp. 3975–3979).

Wang, Wentao, Huang, Xuanyao, Wang, Tianyang, & Roy, Swalpa Kumar (2023). DeepArt: A benchmark to advance fidelity research in AI-generated content. *CoRR*.

Wu, Xixuan, Qiao, Yu, Wang, Xiaogang, & Tang, Xiaoou (2016). Bridging music and image via cross-modal ranking analysis. *IEEE Transactions on Multimedia*, *18*(7), 1305–1318.

Wu, Ho-Hsiang, Seetharaman, Prem, Kumar, Kundan, & Bello, Juan Pablo (2022). Wav2CLIP: Learning robust audio representations from clip. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2022, virtual and Singapore, 23–27 May 2022* (pp. 4563–4567).

Wu, Xixuan, Xu, Bing, Qiao, Yu, & Tang, Xiaoou (2012). Automatic music video generation: cross matching of music and image. In *Proceedings of the 20th ACM multimedia conference, MM '12, Nara, Japan, October 29 – November 02, 2012* (pp. 1381–1382).

Xing, Baixi, Zhang, Kejun, Sun, Shouqian, Zhang, Lekai, Gao, Zenggui, Wang, Jiaxi, et al. (2015). Emotion-driven Chinese folk music-image retrieval based on DE-SVM. *Neurocomputing*, *148*, 619–627.

Xing, Baixi, Zhang, Kejun, Zhang, Lekai, Wu, Xinda, Dou, Jian, & Sun, Shouqian (2019). Image-music synesthesia-aware learning based on emotional similarity recognition. *IEEE Access*, *7*, 136378–136390.

Yariv, Guy, Gat, Itai, Wolf, Lior, Adi, Yossi, & Schwartz, Idan (2023). AUDIOTOKEN: Adaptation of text-conditioned diffusion models for audio-to-image generation. In *Proceedings of the annual conference of the international speech communication association, INTERSPEECH* (pp. 5446–5450). International Speech Communication Association.

You, Quanzeng, Luo, Jiebo, Jin, Hailin, & Yang, Jianchao (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence, January 25-30, 2015, Austin, Texas, USA* (pp. 381–388).

You, Quanzeng, Luo, Jiebo, Jin, Hailin, & Yang, Jianchao (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the thirtieth AAAI conference on artificial intelligence, February 12–17, 2016, Phoenix, Arizona, USA* (pp. 308–314).