EMID: An Emotional Aligned Dataset in Audio-Visual Modality

Jialing Zou*
Jiahao Mei*
51255901103@stu.ecnu.edu.cn
10215102440@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Guangze Ye
East China Normal University
Shanghai, China
5226591017@stu.ecnu.edu.cn

Tianyu Huai
East China Normal University
Shanghai, China
tyhuai@stu.ecnu.edu.cn

Qiwei Shen
East China Normal University
Shanghai, China
10215101495@stu.ecnu.edu.cn

Daoguo Dong†
East China Normal University
Shanghai, China
dgdong@cs.ecnu.edu.cn

ABSTRACT

In this paper, we propose **E**motionally paired **M**usic and **I**mage **D**ataset (**EMID**), a novel dataset designed for the emotional matching of music and images, to facilitate auditory-visual cross-modal tasks such as generation and retrieval. Unlike existing approaches that primarily focus on semantic correlations or roughly divided emotional relations, EMID emphasizes the significance of emotional consistency between music and images using an advanced 13-dimension emotional model. By incorporating emotional alignment into the dataset, it aims to establish pairs that closely align with human perceptual understanding, thereby raising the performance of auditory-visual cross-modal tasks. We also design a supplemental module named EMI-Adapter to optimize existing cross-modal alignment methods. To validate the effectiveness of the EMID, we conduct a psychological experiment, which has demonstrated that considering the emotional relationship between the two modalities effectively improves the accuracy of matching in abstract perspective. This research lays the foundation for future cross-modal research in domains such as psychotherapy and contributes to advancing the understanding and utilization of emotions in cross-modal alignment. The EMID dataset is available at https://github.com/ecnu-aigc/EMID.

CCS CONCEPTS

• **Applied computing** → **Media arts**; • **Information systems** → **Multimedia streaming**.

KEYWORDS

Music-Image Dataset, Emotional Matching, Cross-modal Alignment

---

*Both authors contributed equally to this research.
†Corresponding author.

1 INTRODUCTION

Guided Imagery and Music (GIM) is defined as a comprehensive therapy integrating professional knowledge of psychology, biomedicine, and aesthetics [3, 4]. Through associations and metaphors between music and images, patients can gain deeper insight into and explore their inner world, which helps to foster emotional expression and improve therapeutic results [2, 37]. However, the current GIM scheme exhibits certain limitations, including imprecise targeting that lacks of pathological specificity, prolonged and labor-intensive procedures that result in high annotation costs, and so on.

Recently, the emergence of Artificial Intelligence Generated Content (AIGC) has instigated disruptive transformations in various cross-modal interaction scenarios. AIGC refers to a rising technology that utilizes advanced algorithms and deep learning methods to match or create a wide range of textual, visual, auditory, and multi-media content tailored to human demands [5]. From an intuitive standpoint, AIGC offers a viable avenue to mitigate the costs associated with music and image collection and annotation in GIM. Furthermore, it has the potential to provide patients with a seamless and captivating audio-visual experience, ultimately enhancing the effectiveness of the treatment. Nevertheless, to the best of our knowledge, little existing research or implementation has been found that applies AIGC technology specifically to the field of psychotherapy.

Actually, the aforementioned phenomenon can be attributed to the current cross-modal alignment approaches relied on by AIGC, which primarily focus on semantic relation, ignoring emotional and artistic level matching, as shown in Figure 1. An image of a laughing girl can easily correspond to an audio of giggling, but a photo of landscape without any sound clue may be challenging to access a suitable match at the semantic level. Meanwhile, it is challenging to find appropriate music that matches accordant artistic conception

of both them without any explicit labels or guidance. Therefore, this article aims to propose a music-image dataset that incorporates emotional information, in order to fulfill the following objectives: (1) update the original cross-modal alignment pattern to project emotional features into the shared latent space. This adjustment is helpful to address challenges related to data collection and annotation reliance as well, (2) provide data support for training existing cross-modal generation and retrieval models. This dataset goes beyond simplistic and superficial semantic binding and takes into account the emotional associations between visual and auditory modalities, enhancing the applicability of the proposed dataset in the field of psychotherapy.

To be more specific, our work makes the following key contributions:

- We introduce a high-quality **E**motionally paired **M**usic and **I**mage **D**ataset (**EMID**), comprising more than 30k data pairs accompanied by abundant emotional annotations.

- We pioneer the attempt to align two distinct modalities based on emotions. Our approach involves extracting semantics from source data to form raw image-music pairs, subsequently determining the optimal matches by projecting each candidate onto the emotional coordinate space, thereby modifying embedded representations from original cross-modal joint embedding space.

- A meticulously designed professional psychological experiment is conducted to assess the dataset's quality and affirm the pivotal role of emotions in modal alignment.

Overall, the EMID will serve as a solid foundation for AIGC's application in psychotherapy-related work. Its availability paves the way for novel advancements in music and image conditional generation and retrieval, taking a closer step to human cognition and unlocking new possibilities and opportunities.

2 RELATED WORK

2.1 Cross-modal Dataset

The latest advancements in cross-modal dataset development labor primarily focus on obtaining images, audio, or videos based on textual inputs (text-to-image [7, 26, 30, 34, 35]; text-to audio/videos [9, 11, 19]) consisting of labels, descriptive information, evaluation content, etc., which are available throughout the internet and can be collected by automated tools with ease. The majority of text sections from existing cross-modal datasets are composed of limited tags or monotonous sentences [29], hindering the establishment of strong sensory impacts through their collocation with images or audio. To overcome this limitation, efforts are being made to collect more impressive audio-visual combinations. Initially, videos and audio are often paired naturally, such as the huge dataset AudioSet [11], which is characterized as an object-oriented video dataset organized by audio events, using a meticulously constructed hierarchical ontology of 632 distinct audio classes, including at least 100 instances of 485 audio event classes. However, plain matching approach may lead to issues such as misalignment between video frames and audio, high data noise and classification errors [9].
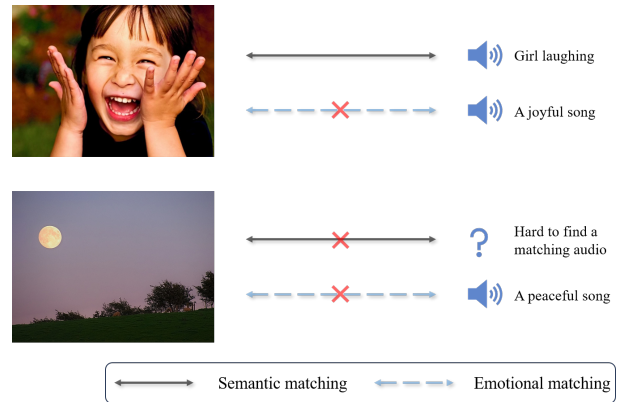

Figure 1: Defect of current cross-modal alignment approaches, ignoring the emotional relations of music and images.

Furthermore, in scenarios where audio is substituted with structured and creative music, the original semantic alignment scheme [40] becomes obsolete. Nonetheless, various solutions have been proposed to address this issue. For example, MusicCaps [1] manually annotated music segments of 5,521 pieces of AudioSet [11], each with an English text description written by ten professional musicians; Huang et al. [18] proposes a dataset called MuLaMCap, which leverages a pre-trained joint music-text model MULAN [17] to annotate music segments in AudioSet [11] based on music descriptions generated by LLM and 200k music descriptive sentences and labels from MusicCaps [1], retaining over 400k music-text pairs. The approach of emotional alignment, however, better conforms to human cognition [24, 41]. Nevertheless, there have been rare research on the emotional alignment between images and music, with one of the solely known music-image paired dataset based on artistic conception, the BOP dataset [8], which artificially annotates 3k+ images and 1.5k+ piano pieces based on creation background and emotions conveyed. Nonetheless, the manual annotation scheme obstructs data expansion, and the richness of source data is expected to be improved. Therefore, there is an urgent demand for a music-image matching dataset.

2.2 Cross-modal Task

Cross-modal tasks involving generation, retrieval, and editing are predominantly addressed using state-of-the-art frameworks such as GAN [13], VQ-VAE [31, 42], and Diffusion models [16, 32]. The recent rapid development of AIGC has garnered substantial attention from both academia and industry. Research efforts [22, 23, 25] have been directed towards visual editing guided by sound and vice versa, taking advantage of conditional guidance to facilitate tasks like style transfer, local fine-tuning, inpainting and super-resolution while preserving the integrity of the original content.

Regarding generative work, there is still ample room for advancement. Early generative works focus on restoring sound of silent instrument-playing videos based on the performer's body movements or gestures [10, 38]. Subsequent jobs attempt to generate corresponding audio based on images or videos of animals, water
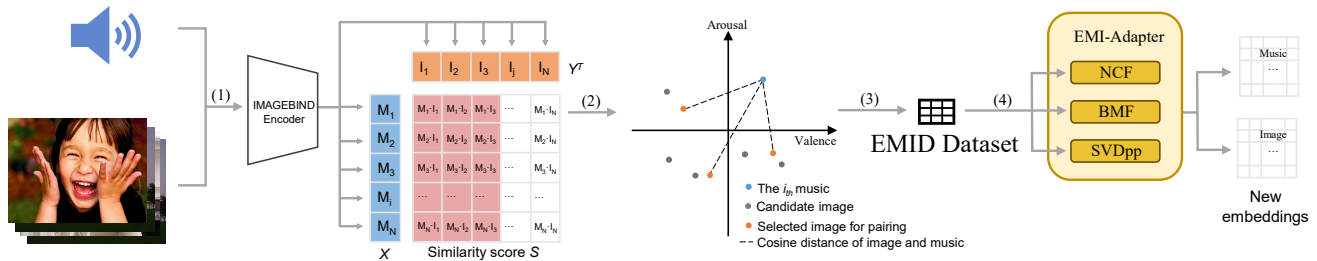
Figure 2: The construction process of EMID includes the following steps. (1) Use ImageBind to get the semantic embedding for images and music as matrix X and Y. Then we get the similarity score matrix through $S = Softmax(X \cdot Y^T)$, (2) select K images with the highest scores (in red color) from i_{th} row of S as the candidate image-set for the i_{th} music, then calculate the cosine distance of each music clip and its image-candidates in emotional valence-arousal space, (3) generate EMID through weighted summation of semantic scores and emotional scores, remaining m best-matching pairs, (4) design a lightweight component, EMI-Adapter, which can merge emotional features into the joint embedding space after training based on the EMID.

flow, knocking actions and more [33, 36, 45]. However, this type of audio often amounts to a semantic repetition of the image, lacking imaginative and artistic audio-visual elements. To overcome this constraint, recent works have explored novel manners to introduce more imaginative and artistic ideas into the generation process. Fan et al. [8] designs a multi-modal architecture named ConchShell based on GAN networks [13], which can generate piano music clips of around 8 seconds conditioned on input landscape images. Surís et al. [39] propose a new method for recommending music tracks that align with a given video (or vice versa), without extra reliance on manual annotation, establishing a connection between the audio-visual pairs at a rhythmic and artistic level.

Training multi-modal data representations that can be aligned in latent spaces has been proved to offer robust support for conditional generation tasks [12, 27, 40]. However, there is still a giant research gap in the domain of direct emotional alignment specifically focused on music and images. Therefore, the **EMID** proposed in this article aims to lay the foundation for this specific aspect of research and development.

3 THE PROPOSED DATASET: EMID

We construct the **EMID**, a high-quality collection of music-image pairs taking account of affective aspects. The EMID ensures a semantic closeness between music and images while evoking similar affective experiences and establishing a unified space of imagination. In the following sections, we will provide a comprehensive and detailed description of the emotional alignment process.

3.1 Emotional Alignment Process

The automated construction process of the EMID involves three distinct phases: (1) acquisition of emotional features, (2) formation of candidate pairs, and (3) injection of emotional information. The overall pipeline is illustrated in Figure 2.

3.1.1 Acquisition of Emotional Features. Both images and music have the ability to convey intricate emotions, although their methods of evoking feelings and intensities differ. As a first step, we aim to extract comparable emotional information from individual samples of music and images. Each music clip conveys a distinctive

emotion. For example, the combination of adagio tempo, deep tones, subdued harmonies, and gentle rhythms can evoke a poignant sense of sadness (e.g. the second movement of Dvorak's *New World symphony*). On the other hand, fast rhythms, strong percussion, and lively melodies can convey feelings of happiness and joy (e.g. the first movement of Bach's *Brandenburg Concerto*).

To acquire a diverse range of emotional data, we refer to the work conducted by Cowen et al. [6], from whose visualization web page[1] we crawl 1,841 music clips, each accompanied by rich emotional annotations. These music clips encompass various genres, including rock, folk, jazz, classical, wind, and heavy metal, covering a broad spectrum of emotional dimensions, such as amusing, annoying, anxious/tense, beautiful, calm/relaxing, dreamy, energizing/pump-up, erotic/desirous, indignant/defiant, joyful/cheerful , sad/depressing, scary/fearful and triumphant/heroic that contains 13 emotions in total, which are represented by letters *A* to *M* in Figure 5, Figure 6 and Table 2.

To establish a standardized representation of emotions, we adopt the Valence-Arousal space theory proposed by Hanjalic [14]. In this theory, valence is used to measure the positive or negative aspects of music emotions, while arousal describes the level of emotional stimulation. For images, we utilize the 8-dimensional emotion classification image data proposed by You et al. [44]. This classification system comprises eight categories: happiness, awe, contentment, surprise, anger, disgust, fear, and sadness. To establish an emotional connection between music and images, we exert the NRC-VAD-Lexicon [28] dictionary to select three semantically similar words for each image classification, which serve as emotional descriptions for that particular category. Subsequently, we query the dictionary to obtain the *<V, A>* coordinates corresponding to these eight classifications (shown in Figure 3a) and calculate their averages as the approximate *<V, A>* coordinates for the images within each category (shown in Figure 3b).

It is worth noting that individual images are not specifically assigned with independent coordinates, as the presence of classification labels already constrain an estimated range of each sample within the Valence-Arousal space. By employing the classification

---

[1]https://www.ocf.berkeley.edu/~acowen/music.html
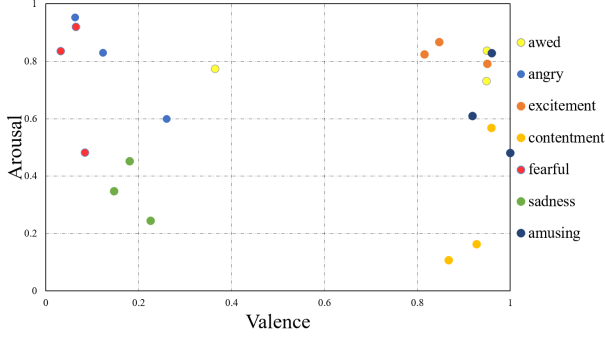
(a) For each of the eight emotional categories, we select three semantically similar words from the NRC-VAD-Lexicon dictionary as emotional descriptions, which are marked with the same color in the figure. Note that the range of Valence-Arousal values is $(0, 1)$.
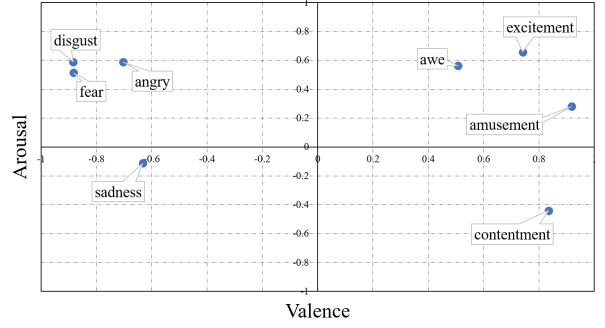
(b) To obtain the approximate Valence-Arousal coordinates for every image category, we calculate the mean values of the emotional description coordinates in Figure 3a for them and map these values to $(-1, 1)$.

Figure 3: Valence-Arousal coordinates of eight emotional categories from the NRC-VAD-Lexicon [28] dictionary and the final calculated values

coordinates to indicate the location of sample points, we dramatically reduce the computational expenses associated with measuring emotional distances in subsequent analysis, while also minimizing any unfavorable influence on the ultimate matching efficacy.

3.1.2 Formation of Candidate Pairs. Based on previous work, we initially pick several matched images for each music sample based on semantic similarities to create an initial dataset. Specifically, we employ the cross-modal retrieval model ImageBind [12] to calculate the semantic matrices for both music and image datasets, recording as $X \in R^{N_M * d}, Y \in R^{N_I * d}$ respectively, Where N_M, N_I refer to the number of music and image samples, and d represents the dimension of latent space that pertains to ImageBind. By performing matrix multiplication, we attain a similarity score between each music sample and image sample, resulting in a semantic similarity matrix S. The specific calculation process is as follow:

$$S = Softmax(X \cdot Y^T) \quad (1)$$

Each element $S_{ij} = \frac{M_i \cdot I_j}{\|M_i\| \cdot \|I_j\|}$ reflects the semantic distance between the i_{th} music segment and the j_{th} image. Utilizing this matrix, we select K (setting $K = 100$ in the EMID) images with the highest semantic relevance for each music sample, thereby setting up candidate music-image pairs. Note that this raw dataset solely relies on simple semantic associations, which may result in suboptimal matching performance. To further tackle this issue, we consider the integration of emotional information.

3.1.3 Injection of Emotional Information. Finally, we adopt a similar method as above to calculate the *<V, A>* coordinates of each music sample and the cosine distance from images in its candidate set in Valence-Arousal space so as to acquire the emotional similarity matrix E between music and alternative images. Then we add semantic and emotional similarity matrices by adjusting appropriate weight parameters α to generate the final score matrix

Table 1: Performance of different EMI-Adapters. The best results are viewed in bold.

	HR@10	NDGC@10
NCF	0.580	0.404
BMF	0.580	0.410
SVD	**0.602**	**0.427**

T for pairing as follows:

$$T = \alpha \cdot S + (1 - \alpha) \cdot E \quad (2)$$

This matrix determines the compatibility between music and image samples. To produce a set of $N_M * m$ music image pairs, we choose m images for each music sample based on the score ranking from highest to lowest.

To ensure that the newly matched samples are as close as possible in the embedding space, a tiny emotional fusion plugin, *EMI-Adapter*, is designed to regulate the encoding of ImageBind and incorporate emotional information into the shared latent space. We simultaneously train several lightweight network modules, comprising of NCF [15], BMF [21], and SVDpp [20], as adapters to promote cross-modal alignment based on existing music-image pairs. Results present in Table 1 and Figure 4 illustrate the performance of the different EMI-Adapter plugins.

Moreover, the EMI-Adapter has the capability to capture more comprehensive emotional information from the input modes leveraging the foundation of ImageBind, eliminating the reliance on additional emotional annotations provided by the source samples. This significant advancement will greatly facilitate the future expansion of the EMID, enabling it to encompass a broader range of emotional content.

(a) Curve of training loss
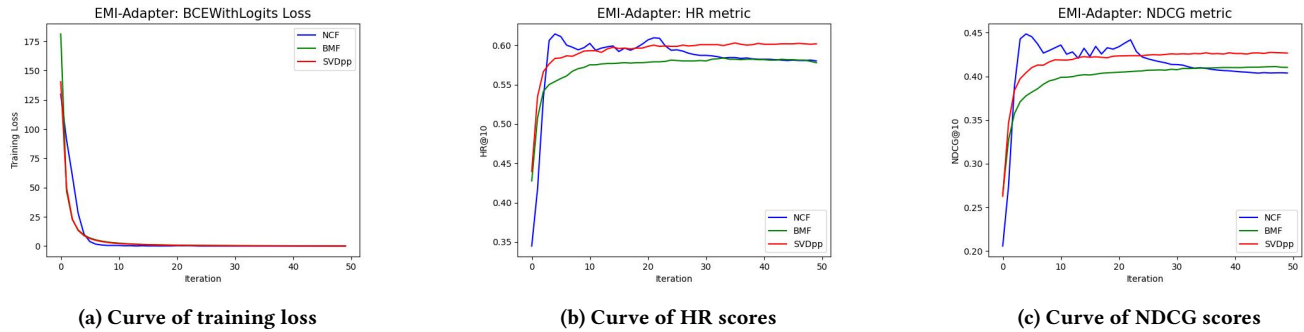
(b) Curve of HR scores

(c) Curve of NDCG scores

Figure 4: The above curves display the variation of training loss and evaluation results in 50 iterations for three types of networks, with @10 indicating that top-10 images will be recommended. At this point, the EMID version used for training is taken as $m = 3$ and $\alpha = 0.6$.

Table 2: The quantities of music dataset before/after expanding and the average of manually set thresholds on 13 emotions.

Emotion	A	B	C	D	E	F	G	H	I	J	K	L	M	Total
Before expanding	45	80	54	131	306	174	367	86	36	124	129	105	199	1836
Average threshold	0.611	0.606	0.559	0.620	0.707	0.632	0.751	0.661	0.610	0.669	0.603	0.570	0.628	-
After expanding	255	545	320	771	1531	889	1832	1036	323	1014	799	484	939	10738

3.2 Filtering and Expansion of Materials

3.2.1 Data Filtering. To keep data balance, we exclude several music samples with a duration of less than 3.7 seconds from the originally crawled music pieces, leaving behind a final dataset of 1,836 music samples with associated tags, among which 98% of the music clips held a duration of 4.8s to 5.2s, as the longest is 9.4s and the shortest is 3.7s. Regarding the image dataset, for the purpose of psychotherapy, we decide to exclude the *disgusting* category, which contains elements of threat, violence, and blood. In addition, to avoid the negative impact of image quality on pairing, we manually define thresholds for the mean and variance of image brightness. Images falling below these thresholds are considered invalid or of low quality and then filtered out. Furthermore, we employ the GIT model [43] to generate text descriptions for all images. Based on these captions, we remove images that are primarily characterized by textual information rather than visual content, as well as photos of bands or singers performing on stage as far as possible, for the sake of escaping from the impediment of strong semantics to emotional connection. After these filtering steps, the final retained image dataset consists of 19,901 pieces.

3.2.2 Data Expansion. To gather more available data, we expand the music samples by leveraging the API of the music listening and recognition platform[2]. This allows us to obtain 1,361 integral music tracks corresponding to fragments in our original dataset. Next, using the *librosa* library[3], we compute the cross-similarity matrix between the music clips in the existing dataset and original music tracks by employing a sliding window approach with a length of 5

seconds and a step size of 1 second. Based on the idea of dynamic programming, we obtain the similarity score (ranged from 0 to 1) from the cross-similarity matrix. Subsequently, we determine the selection of new music clips by setting a manually defined threshold to the similarity score of each music, primarily aiming to differ from the original ones in terms of auditory content while reserving similar affectivity. These new clips are assigned the same labels as the corresponding source samples. The quantities of the music dataset before and after expansion are presented in Table 2, and the ultimately expanded music dataset involves 10,738 music clips.

3.3 Dataset Statistics

The ultimate dataset comprises 32,214 pairs of images and music (setting $m = 3$). The EMID is accompanied with rich annotations as the music samples are equipped with three types of labels: genre, emotion, and sentiment, among which the genre labels classify the music into 13 emotional dimensions, while the sentiment labels provide a detailed breakdown of the percentage distribution of each music sample across different emotional dimensions, and the emotion labels embody another 11 sensory dimension values that quantify the subjective experience evoked by the music. On the other hand, the image samples are marked with category labels and brief content descriptions. This information assists in further understanding the intricate connections between the two modalities. The location of pairs from the EMID in the cross-emotional space of music and images before and after considering emotional alignment is shown in the following Figure 5. It can be observed that the distribution of data is more balanced on different emotional dimensions after fusing effective information.
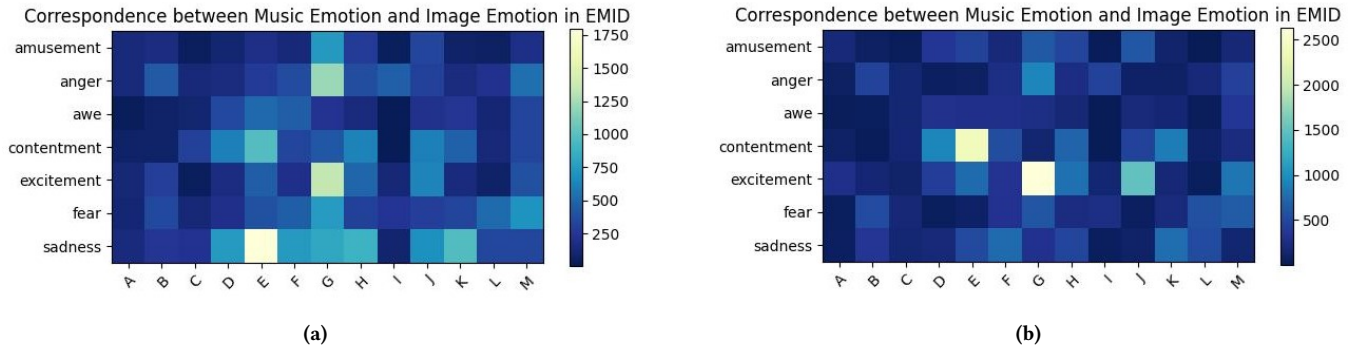
---

[2]https://audd.io/

[3]https://github.com/librosa/librosa

(a)

(b)

Figure 5: Distribution of music-image pairs in the EMID without ($\alpha = 1$) and with ($\alpha = 0.6$) emotional alignment.

Table 3: Accuracy of Music-Image Matching Study. The best results are viewed in bold.

	Max	Min	Mean	Variance
	Random Match			
Music-to-Image	0.363	0.176	0.252	0.002
Image-to-Music	0.341	0.132	0.247	0.002
Overall	0.330	0.176	0.250	0.001
	Before Emotional Aligned			
Music-to-Image	0.604	**0.429**	0.516	0.005
Image-to-Music	**0.725**	**0.462**	**0.575**	0.012
Overall	0.621	0.445	0.546	0.005
	After Emotional Aligned (Ours)			
Music-to-Image	**0.648**	0.407	**0.556**	0.003
Image-to-Music	0.692	0.352	0.563	0.011
Overall	**0.659**	**0.462**	**0.560**	0.005

4 ASSESSMENT EXPERIMENT

4.1 Experimental Protocol

Preview. To assess the efficacy of the EMID, we conduct two types of psychological experiments: music-to-image validation and image-to-music validation. In order to explore whether intensifying emotional connections can effectively support cross-modal alignment research, subjects are instructed to determine one optimal match with the assigned example from the other modality according to their subjective intuition,

4.1.1 Subjects. The study involves 50 subjects from East China Normal University, comprising 29 males and 21 females. The average age of the subjects is approximately 21 years old. All subjects are right-handed, with normal naked or corrected vision.

4.1.2 Design. Music-to-Image Verification. Each subject is supposed to complete 91 *(7 × 13)* trials as 7 music clips from 13 emotional categories are selected respectively as stimulus materials. In each trial corresponding to one music clip, subjects are informed to choose the most suitable image from the four options based on
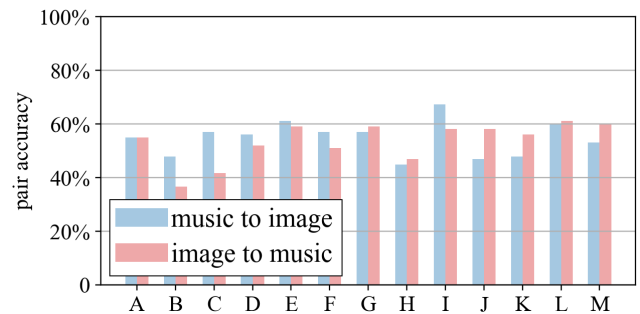

Figure 6: Pair accuracy of Image-to-Music Verification and Music-to-Image Verification on the 13 emotions, letter A to M stands for different emotions.

their personal feeling of music expression after listening to the music clip. Only one of the options is the correct answer determined by the EMID (i.e. the best match estimated from the score matrix in **section 3.1.3**), and the remaining three images are considered as interference items. All options belong to different emotional categories.

Image-to-Music Verification. Each subject is supposed to complete 91 *(7 × 13)* trials as 7 sets of matched images from 13 emotional categories are selected respectively as stimulus materials. In each trial corresponding to one image set, subjects will be presented with three images used for describing the same music clip (i.e. m images that ultimately pair with the music clip in **section 3.1.3**, setting $m = 3$). After observing three images, subjects are informed to choose the most suitable music clip from the four options based on their personal feeling of image expression. The settings for correct answers and interference items are the same as above.

Ablation Experiment. To further probe the effect of emotions on alignment work, we also design an additional small-scale ablation experiment for comparison. To be more specific, we build a variant of the EMID by setting the $\alpha = 1$ in formula 2, thus we attain a dataset merely considering ImageBind semantic matching. By adopting the same partitioning scheme utilized in the previous experiment, we can compare the impact of image pairing before and after emotional alignment on final accuracy while maintaining

consistent music coverage. A total of 15 subjects participated in this experiment.

Additionally, in order to verify that semantic alignment also plays a crucial role as the first step of our work, we simulate random matching scenarios for two types of validations (actually achieved by randomly generating user responses). The interference term settings in the ablation experiment remain consistent with the previous experiment.

4.1.3 Evaluation. Based on the response data of different subjects, we calculate the separate accuracy and overall accuracy of music-to-image and image-to-music experiments. Additionally, we conduct an analysis to investigate the distribution of accuracy across different music emotional categories. Note that accuracy refers to the ratio of the number of consistent situations between the subjects' answers and pairs in the EMID to the total number of validation questions.

4.2 Performance Analysis

The statistical information of the subject data collected is presented in Table 3. The results show that the music-image matching with emotional alignment outperforms other schemes in overall validation. To be more specific, simulated random matching conforms to our conjecture, with an overall mean accuracy of 0.25. Semantic matching based on ImageBind performs significantly better than random pairing, with a variance of 0.005, the maximum accuracy is 0.621, the minimum is 0.445, and the mean is 0.546. After emotional alignment, the matching accuracy improves further in both music-to-image validation and overall experiment, as the general assessment has a variance of 0.005, with a maximum accuracy of 0.659, a minimum of 0.462, and a mean of 0.560. However, in image-to-music verification, the effect after emotional alignment is slightly inferior to before alignment. This may be due to the strong semantics presented in the three images selected through ImageBind, which can definitely correspond to specific background sounds or instruments in the music. But in music-to-image validation which picks one suitable image based on a more abstract form of music, both semantic and emotional factors will be taken into account as images and music used in psychotherapy are supposed to evoke strong emotions rather than having explicit semantics.

Figure 6 presents that the accuracy distribution of music-image matching in the EMID tends to average as a whole in different music emotion categories. In some categories, there are notable differences in accuracy between different validation directions, particularly in categories *C* (anxious/tense) and *J* (joyful/cheerful). In the case of the former category, the accuracy of music-to-image verification is significantly higher compared to the other direction. This difference may be attributed to the removal of the *disgusting* classification in the source image dataset that is closest to category *C* at the sensory level. On the other hand, the latter is exactly the opposite, possibly due to the large scale of *amusement* classification in the source image dataset, which possesses similar feelings to category *J*.

Furthermore, we notice that the validation accuracy of category *I* (indignant/defiant) is relatively higher, while category *B* (annoying) and category *H* (erotic/desirous) tend to have lower accuracy. This pattern may be influenced by the incomplete equilibrium distribution of the source dataset. Additionally, the interference between

similar emotions may also affect the accuracy of classification. To gain broader insights into these issues, our future research will further explore the reasons behind these accuracy differences and addressing these challenges.

5 CONCLUSION

In this paper, we proposed an emotional aligned music-image paired dataset named EMID, which considers the sentimental connections of different modalities in addition to semantic relations. Moreover, we enhanced the existing cross-modal alignment approaches through a simple module called EMI-adapter, which contributes to effectively capturing and integrating emotional information from the indicated samples and projecting it into the joint embedding space. In the end, the psychological experiment demonstrated that the matching pattern in the EMID is more in line with human cognition compared to traditional schemes, and also more suitable for cross-modal tasks in the field of psychotherapy. Currently, our dataset still suffers issues like insufficient scale, weak music structure, and limited data coverage. These challenges will be further discussed and addressed in our future research.

REFERENCES

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).
[2] Kat R Agres, Rebecca S Schaefer, Anja Volk, Susan van Hooren, Andre Holzapfel, Simone Dalla Bella, Meinard Müller, Martina De Witte, Dorien Herremans, Rafael Ramirez Melendez, et al. 2021. Music, computing, and health: a roadmap for the current and future roles of music technology for health care and well-being. *Music & Science* 4 (2021), 2059204321997709.
[3] Helen Lindquist Bonny. 1989. Sound as symbol: Guided imagery and music in clinical practice. *Music Therapy Perspectives* 6, 1 (1989), 7–10.
[4] Helen L Bonny. 2010. Music psychotherapy: Guided imagery and music. In *Voices: A World Forum for Music Therapy*, Vol. 10.
[5] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).
[6] Alan S Cowen, Xia Fang, Disa Sauter, and Dacher Keltner. 2020. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences* 117, 4 (2020), 1924–1934.
[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
[8] Wanpeng Fan, Yuanzhi Su, and Yuxin Huang. 2022. ConchShell: A Generative Adversarial Networks that Turns Pictures into Piano Music. *arXiv preprint arXiv:2210.05076* (2022).
[9] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 829–852. https://doi.org/10.1109/TASLP.2021.3133208
[10] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. 2020. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 758–775.
[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.
[12] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.
[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
[14] Alan Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *IEEE transactions on multimedia* 7, 1 (2005), 143–154.

[15] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[17] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. MuLan: A Joint Embedding of Music Audio and Natural Language. In *Ismir 2022 Hybrid Conference*.

[18] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. 2023. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917* (2023).

[19] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 119–132. https://doi.org/10.18653/v1/N19-1011

[20] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.

[21] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[22] Seung Hyun Lee, Sieun Kim, Innfarn Yoo, Feng Yang, Donghyeon Cho, Youngseo Kim, Huiwen Chang, Jinkyu Kim, and Sangpil Kim. 2023. Soundini: Sound-Guided Diffusion for Natural Video Editing. *arXiv preprint arXiv:2304.06818* (2023).

[23] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. 2022. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3377–3386.

[24] Bochen Li and Aparna Kumar. 2019. Query by Video: Cross-modal Music Retrieval.. In *ISMIR*. 604–611.

[25] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. 2022. Learning visual styles from audio-visual associations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*. Springer, 235–252.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[27] Alex Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2022. Cross-Modal Discrete Representation Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3013–3035.

[28] Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 174–184.

[29] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 407–426.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[31] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[33] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10219–10228.

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

[35] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*. Jülich Supercomputing Center.

[36] Roy Sheffer and Yossi Adi. 2023. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[37] Maria Eleni Smyrnioti, Chrysa Arvaniti, Georgia Kostopanagiotou, and Chrysanthi Batistaki. 2023. Guided Imagery and Music in Patients With Chronic Daily Headache: A Pilot Study. *Music Therapy Perspectives* 41, 1 (2023), e13–e20.

[38] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems* 33 (2020), 3325–3337.

[39] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. 2022. It's Time for Artistic Correspondence in Music and Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10564–10574.

[40] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-Any Generation via Composable Diffusion. *arXiv preprint arXiv:2305.11846* (2023).

[41] Ha Thi Phuong Thao, Gemma Roig, and Dorien Herremans. 2023. EmoMV: Affective music-video correspondence learning datasets for classification and retrieval. *Information Fusion* 91 (2023), 64–79.

[42] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[43] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research* (2022).

[44] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[45] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3550–3558.