

Hybrid-Granularity Image-Music Retrieval Using Contrastive Learning between Images and Music

Xudong He¹, Li Wang^{1*}, Zhao Wang², and Jun Xiao¹

¹ Zhejiang University, Hangzhou, Zhejiang, China

² Ningbo Innovation Center, Zhejiang University, Ningbo, Zhejiang, China
{22360396 li.wang zhao_wang junx}@zju.edu.cn

Abstract. Cross-modal music retrieval is still a challenging task for current search engines. Existing search engines conduct music tracks matching via coarse-granularity retrieval of metadata, such as natural language queries including pre-defined tags and genres. However, such retrieval methods often encounter difficulties while handling fine-granularity queries on contexts. We aim to address fine-granularity music retrieval issue in this work. We construct a dataset with 66,048 image-music pairs for cross-modal music retrieval task. A modality-joint embedding space is learned, where hybrid-granularity context-alignment between images and music is considered via contrastive learning. Additionally, contrastive learning losses on hybrid-granularity contexts are designed to ensure image-music alignment in both inter-modal and intra-modal scenarios. The proposed approach is evaluated through experiments, which demonstrate that our method successfully aligns images and music, and outperforms previous methods in terms of cross-modal music retrieval tasks (image-to-music and music-to-image). Codes³ will be available for public.

Keywords: Multimodal Learning · Cross-Modal Retrieval · Contrastive Learning · Image-Music Alignment.

1 Introduction

Large-scale music websites, such as SoundCloud⁴ and Audiomack⁵, facilitate search engines based on cross-modal retrieval methods, which fetches music tracks by matching their metadata (e.g., song titles, artists' names, and music genres) with natural language queries. Though some offer more personalized query options (i.e., mood and theme), these retrieval methods still often fail to find soundtracks with implicit context aligned with films and their derivative works. And this is critical for creators to choose appropriate soundtracks.

Researchers dedicate to improve the cross-modal music retrieval systems. Manco *et al.* [21] and Doh *et al.* [7] make the attempts to bridge audio and

³ <https://blossomers.github.io/>

⁴ <https://soundcloud.com/>

⁵ <https://audiomack.com/>

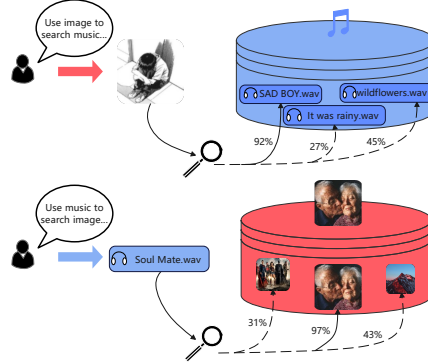


Fig. 1: An example that shows retrieval between image and music in searching engine.

text in music domain by learning a multimodal embedding to facilitate text-based music retrieval task on both tag-level and sentence-level. However, the increasing need to find soundtracks based on contexts is not yet considered, as such text-based music retrieval methods mainly focus on the metadata but not the contexts. To meet these needs, video-based music retrieval methods are developed recently. Yi et al [32] and Cheng *et al.* [2] propose micro-video based music retrieval systems from perspectives of cross-modal generation mechanism and labels noises reduction in datasets, respectively. These micro-videos (less than 10 seconds) usually display consecutive frames in a similar scene, which can be effectively compressed to one single key frame. Thus such video-based music retrieval can be simplified to image-base music retrieval.

Image-music retrieval has addressed lots of attention, since the images express the context information more effectively than text-based ones while holding retrieval efficiency than video-based approaches. In addition, image-based music retrieval are more preferred by users than text-based ones in terms of user experience and usability [25]. Nakatsuka *et al.* [24] utilize contrastive learning technique to learn a joint embedding space to align images and music based on the music genres and their cover art. Stewart *et al.* [27] further propose a cross-modal version of SupCon loss to better align images and music on emotion labels.

Aforementioned image-music retrieval methods are able to handle **coarse-granularity** retrieval as they align images and music based on explicit information like image classes and emotions. However, such methods probably match music with context unrelated images. As shown in the left of Fig 2, a music clip about happiness moment of couples is matched with an image of a dog smiling (same tender emotion), and an epic music clip of films is matched with images of team gathering as they share content-similarity with superheros gathering. We

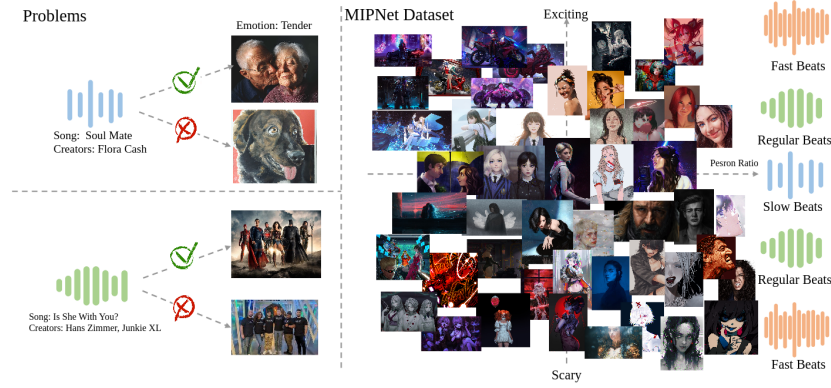


Fig. 2: Prior image-based music retrieval methods [24, 27] will mismatch music and images with unrelated contexts. The proposed MIPNet dataset contains over 66k image-music pairs, which share similar contexts, aligning with general human intuition.

refer these retrieval tasks on context information contained implicitly in queries as **fine-granularity** retrieval, which are not considered by methods [24, 27].

To achieve such fine-granularity retrieval, this work leverages a cross-modal version of Barlow Twins loss [34] to capture the implicit context information. Then we combine it with coarse-granularity retrieval to form a framework HG-CLIM for Hybrid-Granularity Context Alignment Contrastive Learning between Images and Music. It is capable of capturing context information in a hybrid manner. To conduct context-alignment on both coarse-granularity and fine-granularity, we construct a dataset consists of 66,048 image-music pairs, where both coarse-granularity (including the explicit contents in images, beats and rhythm with emotions in music) and fine-granularity (including the implicit context connections between images and music) are considered. As shown in Fig 2, music clips with slow beat tunes and emotional rhythm are paired with images containing raining weather, desolation landscapes and cold colors. Furthermore, to align these hybrid context information effectively, we propose hybrid-granularity contrastive learning losses for both inter-modal and intra-modal scenarios. Our work follows the modality-symmetric feature as [27] that is capable of image-to-music and music-to-image retrieval. The key contributions are summarised as follows:

- To the best of our knowledge, HG-CLIM is the first framework that performs image-music retrieval on queries about contexts which learns a hybrid context alignment between two modalities, which could benefit current metadata based music searching paradigm.
- To address the lack of datasets in the area of image-based music retrieval, we construct a private dataset termed MIPNet. It contains 66,048 image-music

pairs with alignment of hybrid context information, which is essential for further research in the area of image-based music retrieval.

- To capture the implicit contexts, we leverage a cross-modal version of Barlow Twins loss to propose fine-granularity contrastive learning losses for both inter- and intra-modal scenarios. Experimental results has demonstrated the effectiveness of our hybrid-granularity design.

2 Related Works

Cross-modal music retrieval methods [6, 30] utilize triplet loss to find items that close to the anchor queries by distance metrics. With the recent success of contrastive learning on cross-modal alignment [9, 14, 26, 35], it is naturally applied to align music modality with another modality (e.g., texts, videos and images) [7, 11, 12, 21, 32]. Intuitively, researchers connect text modality with music by learning a multi-modal embeddings [7, 21] to perform text-based music retrieval task. However, these methods limit the queries to pre-defined tags and sentences. To retrieve music with more personalized queries, video-based music retrieval approaches are developed rapidly as the micro-video platforms (e.g., Tiktok and Reels) show increasing needs for searching matched background music for micro-videos. And several methods(e.g., [17, 22, 32, 33] put efforts to learn an effective embedding space by leveraging extra information (e.g., optical flow and text) to perform music retrieval. There is a recent method [28] pioneer Control-MVR, which integrates both paradigms via semi-supervised contrastive and dynamically balance audiovisual alignment and genre-specific semantics during inference. More recently, several methods [5, 8, 23] attempt to integrate Large Language Models (LLM) into the frameworks, which involves interactive chat to further refine users’ queries and preferences. However, these video-based approaches are not practical for many music websites (e.g., SoundCloud). Furthermore, it can be simplified to music retrieval based on single key frame, as the micro-videos usually show consecutive frames of one similar scene.

In contrast, image-based music retrieval approaches are more straight-forward and practical since the images express more accurate and complex contexts than tags, and they are more effective than micro-videos on key context deliverance. With such advantages, Nakatsuka *et al.* [24] learn a joint embedding space to align images and music based on the music genres and their cover art. Stewart *et al.* [27] further propose to align images and music on emotion labels. However, these methods only perform image-based music retrieval as classification task (aka coarse-granularity retrieval) but not contexts alignment (aka fine-granularity retrieval). To address this problem, we propose a hybrid-granularity image-based music retrieval framework which takes both coarse-granularity retrieval and fine-granularity retrieval into account.

Large Language Models (LLMs) and diffusion models have advanced content generation across modalities. For example, liu *et al.* [19] leverages latent diffusion models for open-ended visual storytelling, demonstrating LLMs’ capability to generate narrative text from images. However, such generative approaches

prioritize creative content synthesis over precise cross-modal alignment required for retrieval tasks. Wang *et al.* [29] highlight challenges in AI-generated content (e.g., hallucination, consistency), underscoring the need for retrieval systems to complement generative paradigms.

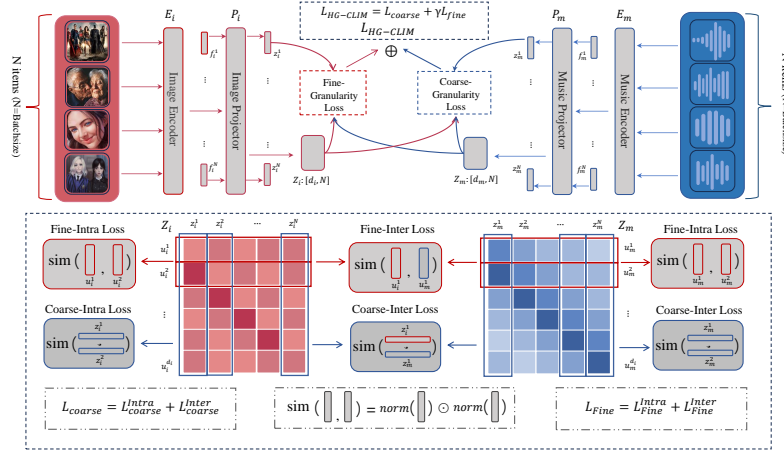


Fig. 3: The overview framework of HG-CLIM. The proposed loss term $\mathcal{L}_{HG-CLIM}$ is constructed with two losses: coarse-granularity \mathcal{L}_{coarse} and fine-granularity \mathcal{L}_{fine} , which compute contrastive loss along the batch(u_i, u_m) and feature dimensions(z_i, z_m), respectively. u_i and u_m are the normalized vectors along with the batch dimension, and z_i and z_m are the normalized vectors along with the feature dimension. The $\text{sim}()$ function computes cosine similarity which is employed by inter-modal loss and intra-modal loss.

3 Methodology

3.1 The MIPNet Dataset

To the best of our knowledge, there is no existing dataset that considers alignment of music and images on context information. Thus we construct a **Music-Image Pairing from Nets Dataset**, termed MIPNet. The MIPNet dataset consisting of 66,048 image-music pairs, which contains music clips (10 seconds per clip), images, and their emotional labels (in text format). There are seven emotion labels for both modalities: "Exciting", "Funny", "Happy", "Tender", "Sad", "Angry", "Scary" as [10]. The image-music pairs are collected by downloading music clips and thumbnails of music videos from online Social Media platforms Youtube ⁶ and Bilibili ⁷, and the emotion labels are pre-annotated by a LLM

⁶ <https://www.youtube.com/>

⁷ <https://www.bilibili.com/>

(e.g., Qwen-VL [1]) and refined by human annotations. In addition, the music clips and images are carefully paired which considers implicit context connections. For example, the images presenting story moments in films are paired with their background music clips, and the images presenting happy moments of couples are paired with music clips containing slow beats with chorus of male and female vocals with happy emotion. The images in the dataset are saved in JPG format, while the music clips are saved with a sampling rate of 32kHz in WAV format. The dataset is randomly split into train, valid, and test subsets in an 8:1:1 ratio.

To reduce the noise in image domain, several filtering methods are applied, for example, duplication removal (based on MD5 hashes and CNN-based model [13]), quality filtering (CNN-based model [18] and human judgement for image quality assessment), textual filtering in images (crop out textual contents), and ratio filtering that filters out abnormal aspect ratio (e.g., greater than 4:1). To reduce the noise in music domain, we perform filtering based on sound quality. The audio files are removed if it contains unclear vocals, low quality of music instruments and concert lives. For the openness of our dataset, we will make the dataset **available** through an application process following Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.

Compared to other datasets, our dataset offers more accurate and rigorous image-music matching logic, aligning with our research focus: fine-grained one-to-one image-music matching. The advantages of our dataset lie in its fine-grained alignment between image and music modalities across multiple levels, coupled with its substantial scale. For further details, please refer to the table below:

Table 1: Comparison of Different Datasets

	Emotion Match	Scene Match	Rhythm Match	Style Match	Key Segment	Over 50k
MIPNet	✓	✓	✓	✓	✓	✓
EMO	✓	×	×	×	×	×
MCA [24]	×	✓	×	✓	×	✓
IMSA [31]	✓	×	×	×	×	✓

3.2 The Proposed HG-CLIM Framework

As shown in Figure 3, the HG-CLIM framework consists of three main components: encoders for feature extraction, projectors for feature projection into a joint embedding space, and the proposed contrastive learning loss term: $\mathcal{L}_{HG-CLIM}$.

Given an image x_i and a music clip x_m , HG-CLIM computes an image feature z_i and an music feature z_m as follows:

$$z_i = \mathbf{P}_i(\mathbf{E}_i(x_i)); z_m = \mathbf{P}_m(\mathbf{E}_m(x_m))$$

Feature Extraction In this work, ConvNext [20] is employed as the image feature encoder, which is shown with both superior efficiency and effectiveness of image classification on ImageNet [4]. The music encoder utilizes PaSST (Patchout Spectrogram Transformer) Hear21 model [16], which performs state-of-the-arts performances of music classification on Audioset [10]. These encoders are capable of capturing robust and informative representations for cross-modal alignment due to their effective extraction of global and local patterns for images and audios.

Projectors for Joint Embedding Space To project extracted features of image and music into a joint embedding space, we employ a projector for one modality to follow each encoder. Each projector is a multi-layer perception (MLP) network. To be specific, the image projector consists of two linear layers, each followed by one BatchNorm and one ReLU activation layer, and one single linear projector. Similarly, the music projector has the same structure as the image projector, and their output dimensions are set to 8192, which is crucial for reconciling the substantial differences between the music and image modalities [34].

3.3 Loss terms for Aligning Modalities

In this paper, $z_i \in \mathbb{R}^{d_i}$ and $z_m \in \mathbb{R}^{d_m}$ denote the **normalized embedding vector** from image and music projectors, where d_i and d_m denote their dimensions. And $Z_i \in \mathbb{R}^{d_i \times N}$ and $Z_m \in \mathbb{R}^{d_m \times N}$ denote the matrix formed by z_i and z_m in a training batch, where N is the batch size. Also, $u_i \in \mathbb{R}^N$ and $u_m \in \mathbb{R}^N$ indicate the normalized embedding vector extracted along with the batch dimension of Z_i and Z_m , and ω_p indicates the modalities of p , with $p \in \{1, 2\}$.

Coarse-Granularity Contrastive Learning Loss for Inter-Intra Modality The coarse-granularity contrastive loss along the batch dimension for cross-modal is defined as :

$$\mathcal{L}_{\text{coarse}}^{\omega_1 \rightarrow 2} = -\frac{1}{N} \sum_{k=1}^N \log \frac{\text{sim}(z_{\omega_1}^k, z_{\omega_2}^k)}{\lambda_c \sum_{\substack{j=1 \\ j \neq k}}^N \text{sim}(z_{\omega_1}^k, z_{\omega_2}^j)} \quad (1)$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \exp(\frac{\mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|}{\tau})$. $z_{\omega_1}^k$ is defined as the k -th embedding vector of modality ω_1 . τ is the temperature factor that adjusts the distribution of the logits. We define intra-modal loss as $\mathcal{L}_{\text{coarse}}^{\text{Intra}}$ and inter-modal as $\mathcal{L}_{\text{coarse}}^{\text{Inter}}$. The coarse-granularity contrastive learning loss is defined as the weighted sum of

both inter-modal and intra-modal losses:

$$\mathcal{L}_{coarse}^{Intra} = \mathcal{L}_{coarse}^I + \lambda^M \mathcal{L}_{coarse}^M \quad (2)$$

$$\mathcal{L}_{coarse}^{Inter} = \mathcal{L}_{coarse}^{I \rightarrow M} + \lambda^{M \rightarrow I} \mathcal{L}_{coarse}^{M \rightarrow I} \quad (3)$$

$$\mathcal{L}_{coarse} = \mathcal{L}_{coarse}^{Inter} + \alpha \mathcal{L}_{coarse}^{Intra} \quad (4)$$

where we set $\alpha = 1$, $\lambda^M = 1$, and $\lambda^{M \rightarrow I} = 1$ empirically.

Fine-granularity Contrastive Learning Loss for Inter-Intra Modality
Inspired by Barlow Twins loss [34], the fine-granularity contrastive learning loss for cross-modal is formulated as:

$$\mathcal{L}_{fine}^{\omega_1 \rightarrow 2} = \sum_k \left(1 - C_{k,k}^{\omega_1 \rightarrow 2}\right)^2 \quad (5)$$

$$+ \lambda_f \sum_k \sum_{j \neq k} (C_{k,j}^{\omega_1 \rightarrow 2})^2$$

$$\mathcal{L}_{fine} = \mathcal{L}_{fine}^{Inter} + \beta \mathcal{L}_{fine}^{Intra} \quad (6)$$

where λ_f is introduced as a balance factor, $\beta = 1$ empirically, and $C_{k,j}^{\omega_1 \rightarrow \omega_2}$ is the cross-correlation matrix computed between the embeddings $u_{\omega_1}^k$ and $u_{\omega_2}^j$, which is defined as:

$$C_{k,j}^{\omega_1 \rightarrow 2} = \frac{u_{\omega_1}^k u_{\omega_2}^j}{\sqrt{(u_{\omega_1}^k)^2} \sqrt{(u_{\omega_2}^j)^2}} \quad (7)$$

Follow Equation (4), we design two components: the inter-modal loss and the intra-modal loss, while the fine-granularity contrastive learning loss contains these two parts.

The proposed HG-CLIM loss is a weighted sum of Equation (4) and Equation (6), defined as:

$$\mathcal{L}_{HG-CLIM} = \mathcal{L}_{coarse} + \gamma \mathcal{L}_{fine} \quad (8)$$

where γ denotes the weight factor between two loss terms. The loss term \mathcal{L}_{coarse} "unite" embeddings with similar explicit information (e.g., emotions, rhythms in audios and contents in images) and "separate" embeddings without such information. The loss term \mathcal{L}_{fine} "unite" embeddings with same implicit contexts (e.g., vocals in audios and styles in images) and "separate" embeddings with different contexts.

The theoretical background behind the fine-granularity loss term lies on the alignment between two **normalized vectors sharing a same dimension**. Specifically, in a sharing embedding space, the distance metric for measuring two vectors is using cosine similarity, which calculates the angles between two vectors. However, since the values in vectors contain information on both the scale and direction, thus the cosine similarity only take the coarse-grained direction

into account but without fine-grained scale information. To address this issue, the values of each element in the vectors should be considered. In this work, we firstly normalize every vector into a unit vector, then calculate the cross-correlation matrix between any two vectors, as the correlation matrix should be close to the identity matrix when their values in each element are close enough. In this way, each value in every element of vectors (presenting fine-grained information) are taken into account. By leveraging this with coarse-grained information on direction, we achieve more accurate alignment than regular contrastive learning methods.

4 Experiments

4.1 Implementation Details

We use the Adam [15] optimizer with a learning rate of 8×10^{-5} and a weight decay of 0.1 with 400 training epochs. The dimensions of image features and music features extracted from their pre-trained encoders [16, 20] are fixed to 2048 and 768, respectively, then they are all projected to the same dimension 8192. During the training procedure, the image encoder and music encoder is frozen, and the projectors are trained from scratch. To balance the hyperparameters in Equation (1), (6) and (8), we set temperature-scaling $\tau = 0.2$, $\lambda_c = 1$, $\lambda_f = 0.0061$, $\gamma = 0.01$.

4.2 Experimental Results

Since the proposed MIPNet dataset is composed of one-to-one pairs, evaluation on this dataset is a pair-wise retrieval task. In order to conduct further comparisons with other methods, we also apply our method to other type of retrieval tasks (e.g., emotion-based music retrieval) and assess the generalizability of our approach.

Table 2: Cross-modal Retrieval performance comparisons between methods on MIPNet Dataset. $I \rightarrow M$ and $M \rightarrow I$ denote image-to-music and music-to-image retrieval respectively.

Method	$I \rightarrow M$			$M \rightarrow I$		
	MRR	R@10	P@1	MRR	R@10	P@1
EMO-CLIM	0.0804	0.1592	0.0831	0.0812	0.1633	0.0791
VM-NET	0.3279	0.6463	0.2001	0.3165	0.6258	0.2057
HG-CLIM (ours)	0.5124	0.8080	0.2931	0.5104	0.8082	0.2910

Results on MIPNet Dataset To evaluate the effectiveness of our proposed method, we conduct cross-modal retrieval tasks on the proposed MIPNet dataset.

Both image-to-music and music-to-image retrieval tasks are conducted. Inspired by [24] [27], evaluation metrics including Mean Reciprocal Rank (MRR [3]), Recall@10 ($R@10$), and Precision@1 ($P@1$) are employed to assess the performance of the retrieval tasks.

Since there are no previous methods available for direct comparison on the same task, we selected two models that have demonstrated strong performance in similar tasks. The EMO-CLIM [27] model matches images and music based on emotion labels, while VM-NET [11] pairs videos with music. We extracted the respective feature encoders from these models and reconstructed their loss functions to align with the task of pair-wise matching. We then train all three models: EMO-CLIM, VM-NET, and our method on the MIPNet dataset, and compare them with same evaluation metrics for fair comparisons. As shown in Table 2, our model consistently outperforms the others across all metrics.

Table 3: Cross-modal emotion-based music retrieval comparison among MMTS* [30], EMO-CLIM [27] and our method on MIPNet and EMO Dataset. * denotes MMTS is text-based music retrieval on emotion labels. MRR, $R@10$, $P@1$ represent the performance metrics for both $I \rightarrow M$ (left side) and $M \rightarrow I$ (right side). Best results are shown in underline.

Dataset	Method	MRR	$R@10$	$P@1$
MIPNET	MMTS	0.4575/0.4807	0.6887/0.7123	0.4070/0.4188
	EMO-CLIM	0.4619/0.5072	<u>0.8237/0.7986</u>	0.4917/0.4935
	HG-CLIM	<u>0.4765/0.5123</u>	0.8215/0.7921	<u>0.5033/0.5094</u>
EMO	MMTS	0.6616/0.6843	0.6013/0.6125	0.3904/0.3988
	EMO-CLIM	0.7859/0.7400	0.6533/0.6238	0.4125/0.4234
	HG-CLIM	<u>0.8012/0.7485</u>	<u>0.6908/0.6574</u>	<u>0.4577/0.4396</u>

The table compares Top1 retrieval results of three methods (EMO-CLIM, VM-NET, and HG-CLIM) on bidirectional tasks: Music \rightarrow Image and Image \rightarrow Music retrieval. Our method (HG-CLIM) consistently achieves superior alignment accuracy by capturing both explicit and implicit contexts.



















Method \ Query	Music to Image Query Top1 result	Image to Music Query Top1 result
EMO-CLIM	Sad boy.wav → 	 → Fade.wav
VMNET	Sad boy.wav → 	 → Shine.wav
HG-CLIM (ours)	Sad boy.wav → 	 → Wildflower.wav
EMO-CLIM	State light.wav → 	 → Tell me.wav
VMNET	State light.wav → 	 → Singur.wav
HG-CLIM (ours)	State light.wav → 	 → cozy winter.wav
EMO-CLIM	Wildflower.wav → 	 → Dinner.wav
VMNET	Wildflower.wav → 	 → Soul.wav
HG-CLIM (ours)	Wildflower.wav → 	 → Heroes' Day Off.wav

Fig. 4: Top1 retrieval results of HG-CLIM compared to baselines (EMO-CLIM, VM-NET) on bidirectional cross-modal retrieval tasks.

Also, as illustrated in Fig 4, the HG-CLIM model successfully aligns the song "sad boy.wav" with an image depicting a boy bowing his head in tears and accurately pairs the image of a girl holding flowers with the song "Wildflower.wav". In contrast, baseline methods (e.g., EMO-CLIM and VM-NET) exhibit mismatches in emotional or contextual alignment under the same queries. These results validate HG-CLIM's capability to capture fine-granularity semantic relationships (e.g., scene-specific textures and implicit stylistic cues), demonstrating its superiority in context-aware cross-modal retrieval tasks.

Results on Emotion-aligned Music Retrieval We recognize that the emotion-based music retrieval task is also a valuable research task. Thus we compare our method with the EMO-CLIM and the MMTS [30] framework in this context. Music and images with same emotion labels are regarded as positive pairs, while

the others are treated as negative pairs. Besides, we identified a dataset designed for emotion-based matching. Following the EMO dataset construction approach for AudioSet described in [27], we constructed a dataset and performed comparative experiments, with the results presented in Table 1. Based on this matching criterion, we conduct experiments and report the same retrieval metrics on Table 3.

The results demonstrate that the proposed method (HG-CLIM) continues to perform strongly in this context, further validating the generalizability of the HG-CLIM approach.

4.3 Ablation studies

We conduct in-depth ablation studies to systematically evaluate the effectiveness of different components in proposed HG-CLIM framework. The experimental results are shown in Table 4. The baseline method consists of only the coarse-granularity contrastive learning loss. We apply the fine-granularity loss term to the intra- and inter-modality respectively to evaluate its impact on retrieval performance with MRR metric. In Table 4, it is noticeable that the fine-granularity loss on intra-modal (second row) contributes a slight improvement in task performance compared to the baseline. While applied on inter-modal (third row), the fine-granularity loss contributes significant improvement for model performance. This demonstrates that the fine-granularity loss enables the model to learn more robust and informative representations of images and music for retrieval tasks, which indicates contribution on learning implicit context-alignment information on MIPNet Dataset.

Table 4: Ablation studies on different losses under $\text{MRR}(I \rightarrow M)$ and $\text{MRR}(M \rightarrow I)$ metrics. Baseline method is trained only with loss \mathcal{L}_{coarse} , and next two following methods are trained with extra losses $\mathcal{L}_{fine}^{intra}$, and $\mathcal{L}_{fine}^{inter}$, respectively.

Loss	$\text{MRR}(I \rightarrow M)$	$\text{MRR}(M \rightarrow I)$
Baseline	0.3164	0.2662
Baseline + $\mathcal{L}_{fine}^{intra}$	0.3272	0.3196
Baseline + $\mathcal{L}_{fine}^{inter}$	0.4793	0.4761
Baseline + $\mathcal{L}_{fine}^{intra} + \mathcal{L}_{fine}^{inter}$	0.5124	0.5104

5 Conclusion

In this work, we propose HG-CLIM, a novel image-based music retrieval framework, which aims to align images and music on contexts. To address the lack of

dataset for such purpose, we construct a private dataset MIPNet, which contains 66,048 image-music pairs and their emotion labels. With our proposed hybrid-granularity contrastive learning loss, HG-CLIM is capable of learning an image-music joint embedding space, which considers context alignment on both coarse-granularity and fine-granularity. Experiments on our MIPNet dataset demonstrate this embedding space is effective for cross-modal music retrieval task. Our approach shows a promising direction for image-based music retrieval on context queries. Beyond cross-modal retrieval, our framework’s ability to align implicit contexts between images and music has broader implications for AI-driven creative applications. For instance, in AI storytelling, dynamically matching music to narrative scenes (e.g., pairing suspenseful music with a thriller plot) can enhance emotional engagement. Similarly, in AI music generation, retrieval-based context alignment can guide models to synthesize music that aligns with visual themes (e.g., generating orchestral scores for fantasy landscapes). Our work bridges multimodal understanding and generative AI, offering a foundation for context-aware applications in virtual reality (VR), interactive media, and automated content creation.

Acknowledgments. This research has been supported by Natural Key Research and Development Project of Zhejiang Province (Grant No. 2023C01043), and in part by the Major Program of The National Social Science Fund of China (Grant No. 24&ZD070), Ningbo Natural Science Foundation (Grant No. 2024Z234)

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
2. Cheng, X., Zhu, Z., Li, H., Li, Y., Zou, Y.: Ssvm: Saliency-based self-training for video-music retrieval. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
3. Craswell, N.: Mean reciprocal rank. Encyclopedia of database systems pp. 1703–1703 (2009)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Doh, S., Lee, M., Jeong, D., Nam, J.: Enriching music descriptions with a finetuned-llm and metadata for text-to-music retrieval. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 826–830. IEEE (2024)
6. Doh, S., Won, M., Choi, K., Nam, J.: Textless speech-to-music retrieval using emotion similarity. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
7. Doh, S., Won, M., Choi, K., Nam, J.: Toward universal text-to-music retrieval. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

8. Dong, Z., Liu, X., Chen, B., Polak, P., Zhang, P.: Musechat: A conversational music recommendation system for videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12775–12785 (2024)
9. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
10. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
11. Hong, S., Im, W., Yang, H.S.: Cbvmr: content-based video-music retrieval using soft intra-modal structure constraint. In: Proceedings of the 2018 ACM on international conference on multimedia retrieval. pp. 353–361 (2018)
12. Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J.Y., Ellis, D.P.: Mulan: A joint embedding of music audio and natural language. arXiv preprint arXiv:2208.12415 (2022)
13. Jain, T., Lennan, C., John, Z., Tran, D.: Imagededup. <https://github.com/idealo/imagededup> (2019)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. *Advances in neural information processing systems* **33**, 18661–18673 (2020)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Koutini, K., Schlüter, J., Eghbal-Zadeh, H., Widmer, G.: Efficient training of audio transformers with patchout. arXiv preprint arXiv:2110.05069 (2021)
17. Lee, Y.S., Tseng, W.C., Wang, F.E., Sun, M.: Vmcml: Video and music matching via cross-modality lifting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2060–2069 (2024)
18. Lennan, C., Nguyen, H., Tran, D.: Image quality assessment. <https://github.com/idealo/image-quality-assessment> (2018)
19. Liu, C., Wu, H., Zhong, Y., Zhang, X., Wang, Y., Xie, W.: Intelligent grimm-open-ended visual storytelling via latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6190–6200 (2024)
20. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
21. Manco, I., Benetos, E., Quinton, E., Fazekas, G.: Contrastive audio-language learning for music. arXiv preprint arXiv:2208.12208 (2022)
22. Mao, T., Liu, S., Zhang, Y., Li, D., Shan, Y.: Unified pretraining target based video-music retrieval with music rhythm and video optical flow information. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7890–7894. IEEE (2024)
23. McKee, D., Salamon, J., Sivic, J., Russell, B.: Language-guided music recommendation for video via prompt analogies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14784–14793 (2023)
24. Nakatsuka, T., Hamasaki, M., Goto, M.: Content-based music-image retrieval using self-and cross-modal feature embedding memory. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2174–2184 (2023)

25. Park, J., Shin, H., Oh, C., Kim, H.Y.: “is text-based music search enough to satisfy your needs?” a new way to discover music with images. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–21 (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
27. Stewart, S., Avramidis, K., Feng, T., Narayanan, S.: Emotion-aligned contrastive learning between images and music. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8135–8139. IEEE (2024)
28. Stewart, S., KV, G., Lu, L., Fanelli, A.: Semi-supervised contrastive learning for controllable video-to-music retrieval. arXiv preprint arXiv:2412.05831 (2024)
29. Wang, Y., Pan, Y., Yan, M., Su, Z., Luan, T.H.: A survey on chatgpt: Ai-generated contents, challenges, and solutions. IEEE Open Journal of the Computer Society **4**, 280–302 (2023)
30. Won, M., Salamon, J., Bryan, N.J., Mysore, G.J., Serra, X.: Emotion embedding spaces for matching music to stories. arXiv preprint arXiv:2111.13468 (2021)
31. Xing, B., Zhang, K., Zhang, L., Wu, X., Dou, J., Sun, S.: Image–music synesthesia-aware learning based on emotional similarity recognition. IEEE Access **7**, 136378–136390 (2019). <https://doi.org/10.1109/ACCESS.2019.2942073>
32. Yi, J., Zhu, Y., Xie, J., Chen, Z.: Cross-modal variational auto-encoder for content-based micro-video background music recommendation. IEEE Transactions on Multimedia **25**, 515–528 (2021)
33. Yi, J., Zhu, Y., Xie, J., Chen, Z.: Cross-modal variational auto-encoder for content-based micro-video background music recommendation. IEEE Transactions on Multimedia **25**, 515–528 (2023). <https://doi.org/10.1109/TMM.2021.3128254>
34. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International conference on machine learning. pp. 12310–12320. PMLR (2021)
35. Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C.: Weakly supervised contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10042–10051 (2021)