



LARP: Language Audio Relational Pre-training for Cold-Start Playlist Continuation

Rebecca Salganik*[†]
University of Rochester
Rochester, NY, United States
rsalgani@ur.rochester.edu

Xiaohao Liu*
National University of Singapore
Singapore, Singapore
xiaohao.liu@u.nus.edu

Yunshan Ma[‡]
National University of Singapore
Singapore, Singapore
yunshan.ma@u.nus.edu

Jian Kang
University of Rochester
Rochester, NY, United States
jian.kang@rochester.edu

Tat-Seng Chua
National University of Singapore
Singapore, Singapore
dcscts@nus.edu.sg

ABSTRACT

As online music consumption increasingly shifts towards playlist-based listening, the task of *playlist continuation*, in which an algorithm suggests songs to extend a playlist in a personalized and musically cohesive manner, has become vital to the success of music streaming services. Currently, many existing playlist continuation approaches rely on collaborative filtering methods to perform their recommendations. However, such methods will struggle to recommend songs that lack interaction data, an issue known as the *cold-start* problem. Current approaches to this challenge design complex mechanisms for extracting relational signals from sparse collaborative signals and integrating them into content representations. However, these approaches leave content representation learning out of scope and utilize frozen, pre-trained content models that may not be aligned with the distribution or format of a specific musical setting. Furthermore, even the musical state-of-the-art content modules are either (1) incompatible with the cold-start setting or (2) unable to effectively integrate cross-modal and relational signals. In this paper, we introduce LARP, a multi-modal cold-start playlist continuation model, to effectively overcome these limitations. LARP is a three-stage contrastive learning framework that integrates both multi-modal and relational signals into its learned representations. Our framework uses increasing stages of task-specific abstraction: within-track (language-audio) contrastive loss, track-track contrastive loss, and track-playlist contrastive loss. Experimental results on two publicly available datasets demonstrate the efficacy of LARP over uni-modal and multi-modal models for playlist continuation in a cold-start setting. Finally, this work pioneers the perspective of addressing cold-start recommendation via

relational representation learning. Code and dataset are released at: <https://github.com/Rsalganik1123/LARP>.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Applied computing** → **Sound and music computing**.

KEYWORDS

music playlist continuation, music representation learning, language-audio pre-training, cold-start problem

ACM Reference Format:

Rebecca Salganik, Xiaohao Liu, Yunshan Ma, Jian Kang, and Tat-Seng Chua. 2024. LARP: Language Audio Relational Pre-training for Cold-Start Playlist Continuation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, August 25–29, 2024, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3637528.3671772>

1 INTRODUCTION

In recent years, the rise of music streaming platforms has created significant changes in the orientation of music listening practices [47]. Most notably, the transition from physical to digitized music has manifested itself in the dominance of playlist-based listening over, the previously prevalent, album-based listening [4, 17, 28, 43]. In the general sense, a playlist can be considered as a collection of tracks (songs¹), intended to be listened to in sequential order and unified under some underlying theme such as genre, mood, activity, aesthetic, or other cohesive factor [5, 14, 63]. This radical shift in consumption habits has highlighted the significance of *playlist continuation*, a task which requires an algorithmic curatorial system to augment a list of seed tracks with musically cohesive suggestions to complete a playlist [4, 63].

To date, a large portion of approaches for playlist continuation rely on collaborative filtering (CF), which leverages user-item interactions to learn user and item representations. However, pure collaborative filtering hardly addresses the *cold-start problem*, where a lack of interactive data hinders the ability of a system to generate high-quality item or user representations and offer meaningful suggestions [3, 32, 46, 54, 56–58, 65, 66]. Broadly, solutions to

*Equal contribution.

[†]This work was mainly done while Rebecca was visiting the National University of Singapore.

[‡]Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '24, August 25–29, 2024, Barcelona, Spain

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0490-1/24/08

<https://doi.org/10.1145/3637528.3671772>

¹Track and song are used interchangeably in this paper.

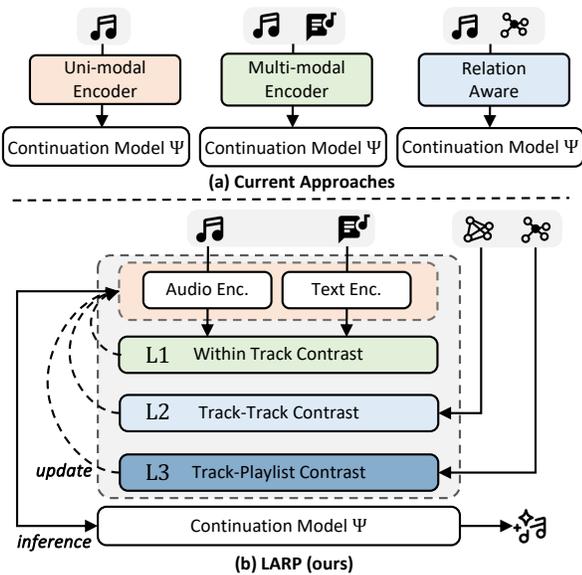


Figure 1: Comparison of cold-start approaches. Current cold-start learning methods (top) delegate content representation to a pre-trained model and focus on relational extraction. However, LARP (bottom) simultaneously focuses on relational and cross-modal learning, addressing both the limitations of previous content models and enhancing the performance of downstream cold-start frameworks.

the cold-start problem require methods which infuse a set of content representations with informative relational signals, enabling a model to later generalize to unseen users and/or items. Existing works approach the cold-start problem by designing complex systems that constrain [32, 46], align [61], or reconstruct [3, 18, 48, 56] relational signals and use them to enhance a set of pre-computed content representations. Thus, central to the success of these methods is the quality of the representations generated by their selected content module. Crucially, these approaches leave the generation of their content representations out of scope, delegating feature extraction to a frozen, pre-trained content extraction module. Meanwhile, music and audio representation learning has presented a wealth of content-based feature extractors which can serve as content-backbones in the aforementioned cold-start approaches. These methods encompass uni-modal [9, 29, 30, 38, 41, 49, 51, 59], language-audio [23, 26, 37, 60], and relational approaches [1, 21, 27].

The state-of-the-art (SOTA) in music representation learning have several key limitations with regard to the task of playlist continuation in a cold-start setting: either they (1) rely on collaborative filtering approaches to extract relational signals and are therefore incompatible with the cold-start setting due to a reliance on interactive data being present during inference or (2) they are unable to effectively integrate cross-modal and relational signals. Given our setting, in which inference is performed over a set of unseen playlists and tracks, there is an urgent need for comprehensive feature extraction methods that combine both relational and multi-modal learning but do not rely on collaborative filtering methods to do so.

In this work, we present a fundamental paradigm shift, transferring the focus of cold-start learning away from the complicated extraction of relational signals and towards the training of relational awareness *within* a multi-modal content module. We argue that the quality of the representations learned by a content module poses a significant bottleneck for cold-start learning. To this end, we introduce LARP, a Language Audio Relational Pre-training model, which effectively overcomes the limitations of previous SOTA music representation learning methods. LARP approaches the task of playlist continuation through a three-layered contrastive learning framework that integrates both multi-modal and relational signals into its learned representations with increasing layers of task-specific abstraction. We extend the canonical contrastive learning framework and introduce the concept of multi-stage contrastive learning by designing cross-modal contrastive loss functions for within-track, track-track and track-playlist pairings. Our approach centers around three key loss functions: (L1) *Within-Track Contrastive Loss*, (L2) *Track-Track Contrastive Loss*, and (L3) *Track-Playlist Contrastive Loss*, visualized in Figure 1. First (L1), individual tracks are passed through an language-audio encoder, which is trained by enforcing alignment between their respective textual and audio embeddings. Then (L2) these representations are further refined by aligning the representations between pairs of tracks which share a parent playlist. In the final layer (L3), we generate representations for playlists enforcing alignment between a subset of tracks within a playlist. In order to show the superiority of our methodology, we augment two publicly available music datasets to include audio samples and meta-data based textual annotations as multi-modal sources of information. Extensive experiments show that our method outperforms both uni-modal and multi-modal models to achieve impressive results on the cold-start playlist continuation task.

In summary, the main contributions of this paper are as follows:

- **Problem Definition.** We argue that the quality of the representations learned by a content module poses a significant bottleneck for cold-start learning and address this challenge through relational pre-training.
- **Model Design.** We propose LARP, which is the first multi-modal relational pre-training framework that uses multi-stage contrastive learning to generate embeddings.
- **Experimental Results.** We augment two canonical music recommendation datasets with audio waveforms and their meta-data based textual annotation counterparts. Then, through extensive experimentation and analysis, we demonstrate the effectiveness of our method in addressing the needs of cold-start playlist continuation

2 PROBLEM FORMULATION

Given a set of tracks, $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$, each track s includes an audio input a_s and a text input t_s , which could be the description about its genre, artist, or title *etc.* (For the exact formulations of our own audio and textual inputs, please see Section 4.1). We define a playlist as a *set*² of tracks, denoted as $p =$

²Please note: in our setting we do not consider the sequential nature of playlists. Rather, we leave this characteristic for future work and simplify a playlist to be an unordered set of tracks.

$\{s_1, s_2, \dots, s_{|p|}\}$, where $|p|$ is the size of the playlist. Typically, we are given a collection of playlists as the training set, denoted as $\mathcal{P} = \{p_1, p_2, \dots, p_M\}$ and a collection of playlists for testing, denoted as $\bar{\mathcal{P}} = \{\bar{p}_{M+1}, \bar{p}_{M+2}, \dots, \bar{p}_{M+\bar{M}}\}$, where M is the size of the training set and \bar{M} is the size of the testing set. Associated with each playlist set is a collection of tracks that correspond to the playlists in the training set, $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$, and testing set, $\bar{\mathcal{S}} = \{\bar{s}_{N+1}, \bar{s}_{N+2}, \dots, \bar{s}_{N+\bar{N}}\}$. The interactions between these tracks and their parent playlists, are represented by a bipartite graph $\mathbf{R}_{M \times N} = \{r_{ps} | p \in \mathcal{P}, s \in \mathcal{S}\}$, where $r_{ps} = 1$ if $s \in p$ and $r_{ps} = 0$ otherwise. Based on the track-playlist affiliation relations in \mathbf{R} , we further define a track-track co-occurrence relation between every pair of tracks which share a parent playlist, forming a homogeneous graph $\mathbf{O}_{N \times N} = \{o_{s_i s_j} | s_i \in \mathcal{S}, s_j \in \mathcal{S}\}$, $o_{s_i s_j} = 1$ if the tracks s_i and s_j belong to the same playlist, and $o_{s_i s_j} = 0$ otherwise.

- **Playlist Continuation:** Given a test playlist, $\bar{p}_q \subset \bar{p} \in \bar{\mathcal{P}}$, consisting of q seed test tracks, $J = \{\bar{s}_j : \bar{s}_j \in \bar{p}_q\}$ such that $|J| < |\bar{p}|$, the task of playlist continuation aims to predict the missing tracks $\bar{s} \in \bar{p} \setminus J$.
- **Cold-start Playlist Continuation:** In this work, we define the cold-start setting to mean that there is no overlap between the tracks and playlists in the training and testing sets. Thus, $\mathcal{S} \cap \bar{\mathcal{S}} = \emptyset$ and $\mathcal{P} \cap \bar{\mathcal{P}} = \emptyset$ ³. Here, the set of training tracks, \mathcal{S} has size, N and the set of testing tracks, $\bar{\mathcal{S}} = \{\bar{s}_{N+1}, \bar{s}_{N+2}, \dots, \bar{s}_{N+\bar{N}}\}$, of size, \bar{N} , is associated with the testing playlists, $\bar{\mathcal{P}}$.

3 METHODOLOGY

We introduce LARP, a novel language audio relational pre-training method, targeted at extracting track representations that are effective and generalizable to various methods for cold-start playlist continuation. We first present the model architecture of LARP, followed by the multi-stage pre-training process, and finally we describe how to employ LARP for cold-start playlist continuation in tandem with several canonical cold-start frameworks.

3.1 Model Architecture

In order to generate audio and text embeddings for each track in our track dataset, we design a multi-modal architecture that is based on the BLIP framework [33]. As shown in Figure 2, our method consists of two uni-modal encoders: one for audio and the other for text. For more extensive details on the formatting of our audio and textual inputs, please see Section 4.1.

Following a similar architecture with [60], we select *HT-SAT* [9], a spectral transformer, for our audio encoding and Bert [12], a language transformer, for our textual encoding. In our experimentation, we tried several other audio backbones such as AST [24] and PANN [30] however each of these had specific drawbacks. First, similarly to [60], we found that the convolutional architecture of the PANN [30] model did not achieve high performance on the audio encoding task. Second, due to the computational complexity of the AST [24] model, we found that the number of parameters significantly limited the batch size that we were able to work with, thus deteriorating performance. As such, we use the HT-SAT, a lighter modified version of AST [24], to generate our audio representations and leave further

³Please note: that throughout this paper, we use $\bar{s} \in \bar{\mathcal{S}}$, $\bar{p} \in \bar{\mathcal{P}}$ to distinguish playlists and tracks in the test set.

backbone analysis for future works. Thus, given a track, s_i , each uni-modal encoder outputs a representation associated with the audio and textual caption of the track, $\mathbf{a}_{s_i} \in \mathbb{R}^{1 \times 768}$, $\mathbf{t}_{s_i} \in \mathbb{R}^{1 \times 768}$, respectively. These are then projected into a unified space where they are represented as $\mathbf{a}_{s_i} \in \mathbb{R}^{1 \times 256}$, $\mathbf{t}_{s_i} \in \mathbb{R}^{1 \times 256}$. In addition to the uni-modal encoders and their cross-modal alignment layers, our architecture includes a momentum encoder which is used to store track representations from previous epochs that can be used during the contrastive loss. This design choice is motivated by the substantial computational cost of generating and optimizing track representations, and we follow the previous practice in BLIP [33] for our implementation of the momentum encoder. This encoder can be seen as a queue which caches representations of tracks from the forward pass of each batch. Each time, a new representation is generated, it is integrated into the queue using a first-in first-out manner. As we explain in the subsequent Section 3.2, this cache is used for sampling negative examples during the optimization of the WTC and TTC losses.

3.2 Multi-stage Pre-training

In order to train our model, we design three contrastive loss variations WTC, TTC, and TPC, each associated with an increasing level of task abstraction. The three losses are integrated via three consecutive stages of training, in which the weights from a previous stage are transferred to the next. We begin by detailing them individually and then explain the paradigm used for their combination. For a visualization, please see Figure 2.

3.2.1 (L1) WTC: Within-Track Contrastive Loss. We borrow the ITC, *Image-Text Contrast*, loss formulation from the original BLIP [33] framework and use it for *Within-Track Contrastive* loss. The motivation behind this loss is grounded in the improved generalization that contrastive loss can provide [1, 38, 60]. Thus, for our first layer of representation learning, we formally define this loss as:

$$\ell_{s_i}^{\text{WTC}} = \text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}), \quad (1)$$

where s_i is a track in the training set and $\mathbf{a}_{s_i}, \mathbf{t}_{s_i}$ are the audio and textual embeddings which are extracted by our multi-modal encoder $\Phi(\cdot)$. Here, the $\text{Contrast}(\cdot, \cdot)$ function follows the formulation proposed by [34]. More concretely, for an input track s_i , we calculate the softmax-normalized audio-to-text (a2t) and text-to-audio (t2a) similarity as:

$$\hat{y}_i^{\text{a2t}} = \frac{\exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_j}) / \tau)}, \quad (2)$$

$$\hat{y}_i^{\text{t2a}} = \frac{\exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_i}) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_j}) / \tau)}$$

where $\text{sim}(\cdot)$ is the cosine similarity function, τ is a temperature hyper-parameter, and s_j are the negative samples, sampled from the momentum queue. Then, given the one-hot ground-truth similarity vectors, $y_i^{\text{a2t}}, y_i^{\text{t2a}} \in \{0, 1\}$, the contrastive loss is defined as the cross entropy $\text{CE}(\cdot)$ between the predicted similarities $\hat{y}_i^{\text{a2t}}, \hat{y}_i^{\text{t2a}}$ and the ground truths $y_i^{\text{a2t}}, y_i^{\text{t2a}}$. Such that:

$$\text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}) = \frac{1}{2} [\text{CE}(\hat{y}_i^{\text{a2t}}, y_i^{\text{a2t}}) + \text{CE}(\hat{y}_i^{\text{t2a}}, y_i^{\text{t2a}})]. \quad (3)$$

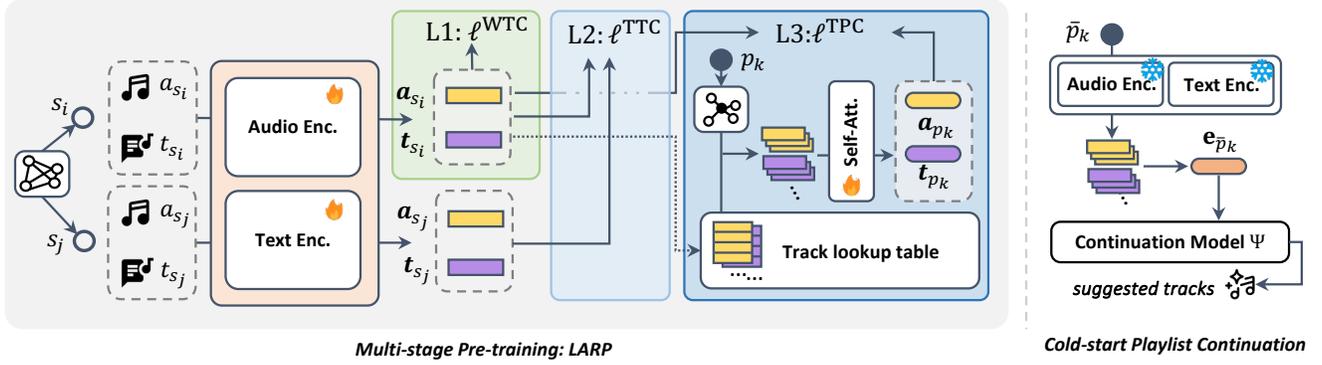


Figure 2: Overview of LARP. During training (left), LARP uses three stages of contrastive loss. In the first stage of training, LARP performs Within-Track Contrastive Loss (WTC) on the text, audio input from a single track. In the second stage of training, LARP performs Track-Track Contrastive Loss (TTC) on the text, audio pairs of co-occurring songs. In the final stage, LARP performs Track-Playlist Contrastive Loss (TPC) on the text, audio pair of a playlist and its child track. During testing (right), LARP is used to generate track embeddings which are pooled to create playlist representations and used as content-input to a downstream cold-start framework to perform playlist continuation.

3.2.2 (L2) TTC: Track-Track Contrastive Loss. In the second stage of training, we refine the representations learned from the WTC loss by applying contrastive loss to align the modalities between pairs of tracks. We refer to this loss as TTC, or *Track-Track Contrastive* loss. The purpose of this loss is to integrate relational signals into the learned representations of individual songs. This layer of contrast allows the model to learn essential information that can be used to unify songs in both the language and audio representational spaces. For two tracks s_i and s_j , which share a parent playlist p_k , we can define this loss as the following:

$$\ell_{s_i, s_j}^{\text{TTC}} = \frac{1}{2} [\text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_j}) + \text{Contrast}(\mathbf{t}_{s_i}, \mathbf{a}_{s_j})], \quad (4)$$

where the $\text{Contrast}(\cdot, \cdot)$ loss is similar with Equation 1, but, we change the contrast pair from single track to two different tracks. The negative samples are also looked up from the momentum encoder. For the full equation, please refer to Appendix A.

3.2.3 (L3) TPC: Track-Playlist Contrastive Loss. Finally, given that the ultimate goal of automatic playlist continuation requires matching between playlists and tracks, we build on the previous loss function to develop a contrastive loss between a track, s_i , and its parent playlist, p_k . The purpose of this loss is to create alignment between clusters of tracks that share a parent playlist, thus improving the neighbourhood awareness of a learned representation. In order to achieve this level of alignment, we design the TPC, short for *Track-Playlist Contrastive*, loss. Thus, given a central track, s_i , its parent playlist, p_k , and a set of neighbour tracks, $J = \{s_j : s_j \in p_k \setminus s_i\}$, we define the loss as follows:

$$\ell_{p_k, s_i}^{\text{TPC}} = \frac{1}{2} [\text{Contrast}(\mathbf{a}_{p_k}, \mathbf{t}_{s_i}) + \text{Contrast}(\mathbf{t}_{p_k}, \mathbf{a}_{s_i})], \quad (5)$$

where $\mathbf{a}_{p_k}, \mathbf{t}_{p_k} \in \mathbb{R}^{1 \times 256}$ are the playlist audio and text representations, respectively which are obtained by aggregating over a series of J seed tracks associated with each playlist using the Equation 6 below.

TPC-fusion. We experiment with several aggregation functions in order to generate a playlist embeddings, $(\mathbf{a}_{p_k}, \mathbf{t}_{p_k})$. While a simple

average would be the most computationally efficient, in practice, this method has some limitations. Primarily, it is unable to account for outliers in a track collection. For example, in a diverse playlist, it is difficult to guarantee that all of the selected tracks will be aligned within the same theme. Thus, using a simple average as an aggregator can cause the noise in the learned representation of a playlist, and significantly mislead the optimization of the TPC objective. To address this problem, we propose a *fusion* module to guide the playlist representation during training⁴. Specifically, we employ one simple self-attention layer on top of the set of track representations to automatically weigh the tracks within a playlist. Thereby, the outlier or noisy tracks would be assigned lower weights when forming the playlist representation. Formally, given a playlist p_k , its representation is calculated by:

$$\begin{aligned} \mathbf{a}_{p_k} &= \text{Self-Att}(\{\mathbf{a}'_{s_j} : s_j \in J\}), \\ \mathbf{t}_{p_k} &= \text{Self-Att}(\{\mathbf{t}'_{s_j} : s_j \in J\}), \end{aligned} \quad (6)$$

where $J = \{s_j : s_j \in p_k \setminus s_i\}$, $\text{Self-Att}(\{\cdot\})$ denotes a one-layer self-attention network, \mathbf{a}_{p_k} and \mathbf{t}_{p_k} are the audio and text representations for p_k , and \mathbf{a}'_{s_j} and \mathbf{t}'_{s_j} are the audio and text representations of tracks within p_k (excluding s_i), which are drawn from the non-differentiable track representation lookup dictionary $\mathbf{A}_{\tilde{s}}$ and $\mathbf{T}_{\tilde{s}}$. We note that this track representation lookup table differs from the momentum encoder used in the previous WTC and TTC losses. This is because the momentum encoder has limited capacity and cannot store all of the tracks, which could skew the results of the playlist aggregation. To address this issue, we leverage two non-differentiable lookup tables to cache the latest audio and text representations for all the tracks during training, denoted as $\mathbf{A}_{\tilde{s}} \in \mathbb{R}^{N \times 256}$ and $\mathbf{T}_{\tilde{s}} \in \mathbb{R}^{N \times 256}$. Both tables are updated during training such that when a track appears in a batch their representations in $\mathbf{A}_{\tilde{s}}$ and $\mathbf{T}_{\tilde{s}}$ will be replaced by the new representations generated by the

⁴Please note that the *fusion* module is not used to generate playlist representations during testing, i.e., playlist continuation, only training.

forward pass. For a more detailed breakdown to this loss function and its calculation procedures, please refer to Appendix A.

Multi-stage Training. We train the model in three stages, using early stopping to achieve the optimal performance for each loss and before proceeding to the following stage. The weights from each intermediate stage are transferred between stages and the model remains unchanged. Only the loss function is modified to integrate each increasingly abstract loss. Thus, for a given track s_i and its affiliated playlist p_k , and *neighbour set*, $\{s_j : p_k = p(s_j), s_j \neq s_i\}$, we define the stages as:

$$\text{Stage 1: } \ell_1 = \ell_{s_i}^{\text{WTC}}; \quad (7)$$

$$\text{Stage 2: } \ell_2 = \ell_{s_i}^{\text{WTC}} + \ell_{s_i, s_j}^{\text{TTC}}; \quad (8)$$

$$\text{Stage 3: } \ell_3 = \ell_{s_i}^{\text{WTC}} + \ell_{s_i, s_j}^{\text{TTC}} + \ell_{p_k, s_i}^{\text{TPC}}. \quad (9)$$

Currently, we just use a simple average over the three losses, which already demonstrates excellent performance. We leave more complex balances between the three loss functions for future work.

3.3 Cold-start Playlist Continuation

After the pre-training stage, we obtain the model $\Phi(\cdot)$ for cold-start playlist continuation task. Specifically, we first generate the text and audio representations:

$$\mathbf{a}_{s_i}, \mathbf{t}_{s_i} = \Phi([a_{s_i}, t_{s_i}]), \quad (10)$$

where a_{s_i} and t_{s_i} stand for the audio and text input of the track s_i , respectively. Analogous to the pre-training, we employ a simple average pooling to obtain its unified representation $\mathbf{e}_{s_i} \in \mathbb{R}^{1 \times 256}$, formally represented as:

$$\mathbf{e}_{s_i} = \text{mean}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}). \quad (11)$$

Thus, we can obtain the embedding table for all tracks in training set \mathcal{S} as $\mathbf{E}_{\mathcal{S}} \in \mathbb{R}^{N \times 256}$, and testing set $\bar{\mathcal{S}}$ as $\mathbf{E}_{\bar{\mathcal{S}}} \in \mathbb{R}^{\bar{N} \times 256}$.

Following the typical setting in playlist continuation [47, 63], we generate the playlist representations \mathbf{e}_p and $\mathbf{e}_{\bar{p}_q}$ by simply average pooling its included tracks' representations, denoted as:

$$\mathbf{e}_p = \text{mean}(\{\mathbf{e}_{s_i} : s_i \in p\}), \quad \mathbf{e}_{\bar{p}_q} = \text{mean}(\{\mathbf{e}_{\bar{s}_i} : \bar{s}_i \in \bar{p}_q\}), \quad (12)$$

where \mathbf{e}_p and $\mathbf{e}_{\bar{p}_q}$ are used for training and testing, respectively. And the cold-start playlist continuation task aims to predict the missing tracks $\bar{s} \in \bar{p} \setminus \bar{p}_q$.

In order to showcase the ability of LARP to learn effective and generalizable representations, we integrate the representations generated by Equations 10 and 12 with three cold-start playlist continuation methods: ItemKNN [45], DropoutNet [56], and CLCRec [58]. It should be noted that these methods are originally designed for cold-start recommendation and we adapt them to the task of cold-start playlist continuation (please see B for more details).

- **ItemKNN** [45] is a parameter-free approach which relies on the cosine similarity between the input partial playlist $\mathbf{e}_{\bar{p}_q}$ and all the candidate tracks $\mathbf{E}_{\mathcal{S}}$, to find the K most similar tracks.
- **DropoutNet** [56] is designed to treat the cold-start as a robustness task in which sparse interactive signals are reconstructed using content features. It adopts WMF [25] to generate preference representations for playlists and tracks, and randomly applies dropout to a training set of preference-content concatenations, which are fed into a DNN model for score predictions.

- **CLCRec** [58] is a contrastive learning framework that reconstructs collaborative signals from the provided content embeddings. And it also uses inner product to compute the similarities, then to find the K possible tracks in the $\bar{p} \setminus \bar{p}_q$.

4 EXPERIMENTS

To guide our analysis, we center our experimentation around answering the following research questions:

- RQ1.** How does our proposed method perform in comparison with our selected benchmarks on the task of playlist continuation?
- RQ2.** How does each block of our proposed framework contribute to the final performance of our method?
- RQ3.** What are the key properties of our method that make it well suited to the task of cold-start playlist continuation?

4.1 Experimental Settings

4.1.1 Datasets. We evaluate LARP on The Million Playlist Dataset (MPD) [8] and the Last-FM (LFM) [40] datasets (Table 1). Our motivation for selecting these datasets is grounded in two key requirements of our setting: (1) direct access to audio samples and metadata, and (2) inclusion of track-based interactions. While there are several playlist datasets available [22, 39, 52, 64], none of them would give us access to the audio samples. And, even though the Melon dataset [22] has pre-computed spectrograms, the sampling frequency and time duration of these spectrograms are not compatible with the needs of our backbone audio extraction module.

We define our own procedure for augmenting both MPD and LFM to include the language-audio pairs necessary for our multi-modal pre-training methodology. For generating the audio samples, we scrape all the associated MP3s and truncate them into 10 second clips sampled randomly from the track (excluding the first and last 20 seconds). Then, to meet the size requirements of our backbone audio model, HTS-AT [9], we truncate each MP3 file into 10 second sound clips and upsample the audio quality to be 44.8kHz. To generate the captions, we use the associated Spotify IDs to collect the track name, artist name, and album name. The captions are then formulated using the following structure: The track <track name> by <artist name> on album <album name>. We leave the exploration of further captioning strategies for future work.

Finally, we acknowledge that LFM is not explicitly designed for playlist continuation, so we make several modifications to align it with our task. We draw inspiration from the notion of Weekly Discovery. This service, which is available on many popular streaming platforms [4, 50], involves the creation of a playlist on the basis of the aggregated listening habits over a user on a weekly basis. Unlike static playlist continuation, in which seed tracks are collected from the user inputs, Weekly Discovery uses implicit feedback from users to identify the seed tracks used for recommendation. Thus, to create a playlist continuation dataset from the user-track interaction data provided by LFM, we perform the following steps:

- (1) We randomly select a series of users for each split of training, validation, and testing.
- (2) We treat each user as a 'playlist' and aggregate all the songs that they have listened to into a single collection.

Table 1: Statistics of datasets with number of tracks (#S), number of playlists (#P), and Homogeneity (Homo.). Homogeneity is defined as the average pairwise cosine similarity among the audio embeddings of the tracks in a playlist.

Dataset	Training		Evaluation		Relevant Stats	
	#S	#P	#S	Avg #S/P	Avg #P/S	Homo.
MPD	98,552	100	6,978	98.12	1.41	0.60
LFM	87,558	1,000	17,439	91.04	4.63	0.56

- (3) We iteratively remove songs which overlap between the split pools (i.e., songs in the training set are removed from the subsequent validation and test sets).
- (4) To construct our test set, we select a contiguous subset of interactions for each user/playlist to contain more than 30 songs and less than 100 for standardization between MPD and LFM.

4.1.2 Baseline methods. We compare LARP with four state-of-the-art music representation learning baselines, including Jukebox [13], MULE [38], CO-Playlist [1], and CLAP [60]. Please note that, due to the cold-start nature of our dataset design, we do not include any collaborative filtering baselines, which require interactive data for representation learning.

- **Jukebox** [13] is a generative music model with state-of-the-art performance in various music understanding tasks [7].
- **MULE** [38] is an audio representation learning model based on convolutional neural network (CNN) [30].
- **CO-Playlist** [1] is a contrastive learning model, in which positive pairs are selected through playlist-track interaction data.
- **CLAP-PT/CLAP-FT** [60]⁵ is a multi-modal model. For the pre-trained (PT) setting, we use the pre-trained model to generate representations. For the fine-tuned (FT) setting, we fine-tune before generating representations.

4.1.3 Hyper-parameter settings and reproducibility. We train LARP in a distributed setting with two Nvidia A40 GPUs, each with 48GB of memory. We train for 45 epochs with early stopping that terminates after 5 epochs without improvement. We use a batch size of 50 and use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Our learning rate of $1e^{-4}$ is controlled by a cosine scheduler after a 3000 step warm up period. Our final results are achieved with an embedding size of 256. For our temperature parameter we use $\tau = 0.07$. Our momentum encoder has a queue of size 57600 and a momentum parameter of $m = 0.995$. For the hyperparameters used in our cold-start frameworks, please see Appendix B.

4.2 Overall Performance Comparison (RQ1)

First, we compare our method with the various encoder benchmarks across all the downstream recommendation frameworks. As we can clearly see in Table 2, our method consistently outperforms all the benchmarks across all metrics and their k values on both LFM and MPD. This performance shows the consistency of our approach over a variety of canonical metrics and cold-start music datasets. Furthermore, we can see that LARP significantly outperforms the other relational representation learning benchmark, CO-Playlist [1].

⁵There is a concurrent work [19] that has the same name and similar idea with CLAP [60] that we implemented in this paper.

This indicates the superiority of our relational learning paradigm and the value of adding textual inputs (which CO-Playlist does not use). Also, comparing with one of the two strongest benchmarks, MULE, we can see that even though MULE was trained on 1.8M examples of high-quality, proprietary music data [38], our method still performs better.

Second, we compare the performance of LARP, our method, across the various recommendation frameworks. Here, we can see that the difference between using a simple, non-parametrized recommendation model such as ItemKNN and a SOTA method such as DropoutNet or CLCRec is relatively insignificant. This indicates that LARP’s exceptional performance can be achieved without reliance on a complex downstream recommendation model. This important finding highlights the benefits that the integration of cross-modal and relational signals can have in the cold-start setting. Furthermore, the lack of significant gains between various frameworks indicates that the key differentiating factor between these methods lies in the quality of the learned content representations. To us, this highlights the importance of considering the quality of content module as part of a cold-start recommendation task as compared to the downstream recommendation framework.

4.3 Ablation Study (RQ2)

To assess the effectiveness of each element in our proposed LARP architecture, we iteratively remove each component and evaluate the model’s performance. In Table 3 we present the results from each of the unimodal backbone encoders, followed by iterations of LARP in which we remove first the losses and then the *fusion* layer.

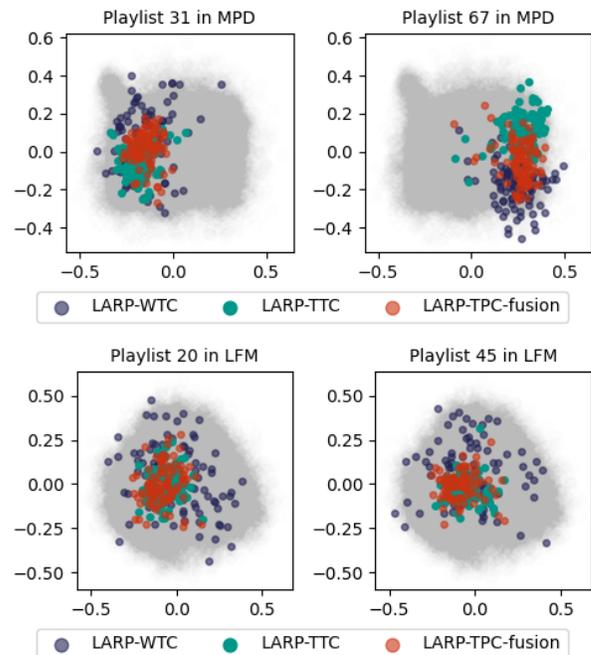


Figure 3: Within playlist alignment of tracks. Visualization of tracks associated with individual playlists. Note, gray cloud indicates the entire embedding set.

Table 2: Effectiveness results on Recall@K (R@K) and NDCG@K (N@K). Best results in bold and second best with underscore.

Rec. Model	Encoder	MPD Dataset						LFM Dataset					
		R@10	N@10	R@20	N@20	R@40	N@40	R@10	N@10	R@20	N@20	R@40	N@40
ItemKNN [45]	Jukebox [13]	0.0012	0.0045	0.0025	0.0051	0.0055	0.0059	0.0005	0.0017	0.0008	0.0015	0.0021	0.0019
	MULE [38]	0.0156	0.0798	0.0286	0.0738	0.0457	0.0609	<u>0.0036</u>	<u>0.0140</u>	0.0066	0.0131	0.0121	0.0133
	CO-Playlist [1]	0.0012	0.0058	0.0022	0.0056	0.0067	0.0076	0.0008	0.0036	0.0016	0.0036	0.0024	0.0029
	CLAP-PT [60]	0.0020	0.0099	0.0041	0.0099	0.0065	0.0084	0.0013	0.0061	0.0024	0.0054	0.0038	0.0047
	CLAP-FT [60]	<u>0.0217</u>	<u>0.1047</u>	<u>0.0386</u>	<u>0.0967</u>	<u>0.0682</u>	<u>0.0877</u>	0.0033	0.0125	<u>0.0067</u>	<u>0.0131</u>	<u>0.0132</u>	<u>0.0137</u>
	LARP (Ours)	0.0406	0.1964	0.0753	0.1868	0.1306	0.1669	0.0137	0.0536	0.0233	0.0479	0.0389	0.0446
DropoutNet [56]	Jukebox [13]	0.0022	0.0106	0.0034	0.0090	0.0063	0.0082	0.0008	0.0040	0.0014	0.0035	0.0023	0.0030
	MULE [38]	<u>0.0286</u>	<u>0.1464</u>	<u>0.0534</u>	<u>0.1377</u>	<u>0.0861</u>	<u>0.1166</u>	0.0066	<u>0.0261</u>	<u>0.0131</u>	<u>0.0259</u>	<u>0.0236</u>	<u>0.0257</u>
	CO-Playlist [1]	0.0016	0.0087	0.0034	0.0088	0.0065	0.0083	0.0011	0.0053	0.0017	0.0043	0.0030	0.0038
	CLAP-PT [60]	0.0022	0.0114	0.0047	0.0117	0.0081	0.0105	0.0006	0.0031	0.0016	0.0034	0.0031	0.0034
	CLAP-FT [60]	0.0283	0.1396	0.0531	0.1325	0.0888	0.1158	<u>0.0070</u>	0.0259	0.0128	0.0247	0.0210	0.0232
	LARP (Ours)	0.0418	0.2171	0.0784	0.2028	0.1375	0.1815	0.0146	0.0558	0.0245	0.0495	0.0417	0.0471
CLCRec [58]	Jukebox [13]	0.0014	0.0078	0.0038	0.0092	0.0071	0.0088	0.0010	0.0036	0.0013	0.0029	0.0021	0.0025
	MULE [38]	<u>0.0294</u>	<u>0.1546</u>	<u>0.0521</u>	<u>0.1391</u>	<u>0.0878</u>	<u>0.1201</u>	<u>0.0087</u>	<u>0.0324</u>	<u>0.0168</u>	<u>0.0321</u>	<u>0.0310</u>	<u>0.0326</u>
	CO-Playlist [1]	0.0026	0.0121	0.0041	0.0103	0.0069	0.0091	0.0009	0.0060	0.0015	0.0048	0.0023	0.0036
	CLAP-PT [60]	0.0024	0.0111	0.0042	0.0103	0.0071	0.0091	0.0012	0.0053	0.0017	0.0042	0.0030	0.0038
	CLAP-FT [60]	0.0240	0.1137	0.0446	0.1088	0.0814	0.1020	0.0076	0.0289	0.0141	0.0275	0.0241	0.0264
	LARP (Ours)	0.0416	0.2069	0.0779	0.1962	0.1330	0.1737	0.0151	0.0581	0.0259	0.0527	0.0423	0.0487

Table 3: Ablation study with ItemKNN [45] as the recommender. The last row is our final method, LARP-TPC-*fusion*. Each row before is an ablated version of LARP without (w/o) one of the contrastive modules.

Baseline	Encoder	MPD Dataset						LFM Dataset					
		R@10	N@10	R@20	N@20	R@40	N@40	R@10	N@10	R@20	N@20	R@40	N@40
Text	Bert [12]	0.0075	0.0415	0.0122	0.0351	0.0191	0.0285	0.0032	0.0128	0.0052	0.0109	0.0089	0.0104
Audio	HTS-AT [9]	0.0012	0.0072	0.0031	0.0079	0.0071	0.0086	0.0006	0.0025	0.0011	0.0024	0.0023	0.0025
w/o ℓ^{TTC}	LARP-WTC	0.0256	0.1278	0.0444	0.1168	0.0779	0.1039	0.0057	0.0232	0.0104	0.0218	0.0201	0.0218
w/o ℓ^{TPC}	LARP-TTC	0.0373	0.1968	0.0724	0.1833	0.1243	0.1631	0.0141	0.0587	0.0256	0.0539	0.0425	0.0496
w/o <i>fusion</i>	LARP-TPC	0.0357	0.1865	0.0641	0.1699	0.1132	0.1517	0.0094	0.0354	0.0171	0.0333	0.0286	0.0320
LARP- <i>fusion</i>	LARP-TPC- <i>fusion</i>	0.0406	0.1964	0.0753	0.1868	0.1306	0.1669	0.0137	0.0536	0.0233	0.0479	0.0389	0.0446

First, we notice that each of the individual encoders has significantly lower performance than the overall architecture, motivating the need for the cross-attention layers and the WTC contrastive loss that is used to unify the uni-modal encoders.

Second, we consider the fundamental architecture of LARP. Contrasting the performance of LARP-WTC with respect to its benchmark counterpart, ItemKNN/CLAP-FT in Table 2, we can see that even without the remaining two losses, our method outperforms this comparable multi-modal benchmark by a significant margin.

Third, we analyze the performance gains achieved by the integration of each loss. We can see that LARP-TTC, which begins to integrate relational content via track-track alignment, already shows noticeable improvement over the LARP-WTC which has no relational awareness. This is because the ℓ^{TTC} loss enables the model to account for acoustic or linguistic heterogeneity within the tracks of a playlist. For example, if we consider two tracks from different genres (e.g. Jazz and Country) that are both titled "Love", a method which is unaware of their relation in a playlist would be unaware of the unifying element (title) between them.

Fourth, we explore the importance of our fusion layer and its complementary ℓ^{TPC} loss. Simply the integration of a ℓ^{TPC} does not improve the performance and actually achieves worse performance than the ℓ^{TTC} loss. This is because the effectiveness of the ℓ^{TPC} loss hinges on the quality of the playlist representation that is used to anchor the children tracks. Without the self-attention layer, the model is unable to properly integrate information from a diverse set of seed tracks into a meaningful playlist representation. However, in adding this trained self-attention layer we are able to generate expressive playlist encodings that contribute to the final success of our LARP-TPC-*fusion* method.

Notably, we can see that there is a difference in which method performs best between the two datasets. For MPD, we can see that LARP-*fusion* clearly outperforms LARP-TTC. We attribute this to the clear alignment between the MPD dataset setting (which was clearly designed for the playlist continuation task) and the formulation of the TPC loss that uses the *fusion* encoder layer. Meanwhile, on LFM, we can see that the performance of LARP-TTC supersedes that of LARP-*fusion*. This is due to the potential misalignment between the LFM dataset and the goals of the TPC

Table 4: Generalization results. We apply LARP, pre-trained on MPD, and test its performance on LFM as compared with best performing benchmarks (both trained and tested on LFM). Best results in bold.

Encoder	Training Dataset	LFM Results					
		R@10	N@10	R@20	N@20	R@40	N@40
MULE	LFM	0.0036	0.0140	0.0066	0.0131	0.0121	0.0133
CLAP-FT	LFM	0.0033	0.0125	0.0067	0.0131	0.0132	0.0137
LARP-TPC- <i>fusion</i>	MPD	0.0102	0.1311	0.0165	0.1178	0.0276	0.1079

loss. As we show in Table 1, LFM shows noticeably higher levels of diversity in its tracks, thus it is possible that there is too much heterogeneity among the representational spaces of the tracks, making it difficult for a single attention layer to correctly capture a meaningful playlist representation. However, the flexibility of our approach allows for the selection of different variations that can be tailored to the needs of a dataset.

Finally, we provide a case study in which we visualize the embeddings of various tracks in a playlist (see Figure 3) that are generated by each method presented in the ablation table results above (Table 3). To create these plots, we randomly select an example playlist, p , and its associated tracks, $S_p = \{s_i : s_i \in p\}$. We generate embeddings for the playlist, e_p , and tracks, $e_{S_p} = \{e_{s_i} : s_i \in S_p\}$ by Eqs. (11) and (12). We then apply 2-dimensional PCA on the set of embeddings and plot the outcomes. This visualization is significant in the context of the industrial setting where online recommendations are served by selecting nearest neighbours from a large scale embedding space representing an extremely large catalog. In this setting, efficiency can be gained from having similar songs tightly clustered in the overall embedding space. We show that the LARP-TPC-*fusion* generates song embeddings that are most closely clustered together in the embedding space which is aligned with the purpose of the ℓ^{TPC} loss. By creating a playlist-track relational awareness in our representation learning module, we are able to identify clusters in unseen, held out evaluation tracks. Thus, LARP’s fine-grained representation learning framework facilitates high performance from even a simple k-means recommendation framework like ItemKNN.

4.4 Model Study (RQ3)

4.4.1 Generalization Performance. In addition to our evaluation of unseen playlists and items from within the same dataset, we also evaluate the generalization capabilities of LARP on completely different datasets. In Table 4, we showcase the performance of the best performing methods on the LFM dataset and evaluate against a LARP model which has been trained on the MPD dataset. Crucially, we can see that this generalization performance is incredibly strong, even going so far as to outperform the best benchmark performance on the LFM dataset. This is extremely important for the cold-start setting, because it shows that our model is robust to the dataset distribution - a gold standard in this setting.

4.4.2 Parameter Sensitivity. As we show in the ablation study, one of the key components of our method’s impressive performance is the integration of the TPC, or Track-Playlist Contrast, loss. Critical to the success of this loss function is a comprehensive methodology

for generating meaningful playlist representations that can be used to properly anchor the various tracks within a playlist’s collection.

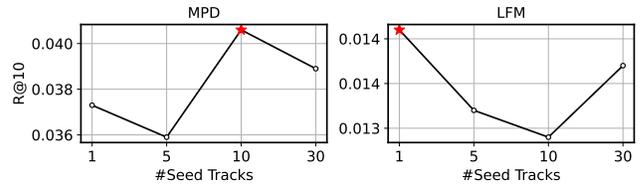


Figure 4: Sensitivity analysis. We show relationship between performance and number of seed tracks, J , for generating playlist representations. Note: the red star signifies the best results reported in the ablation tables above.

In our experimentation, we found that our model had some sensitivity to J , or the number of seed tracks which are used to compute the playlist representations during the training stage. And, most significantly, this parameter differs between datasets. As we show in Figure 4, the best performance on the MPD dataset is achieved by using 10 seed tracks for the generation of the playlist representation by the *fusion* playlist encoder (defined in Equation 6). Meanwhile, in the LFM dataset, the best performance is achieved by using 1 or 30 seed tracks. We believe that this can be attributed to the difference in overall track diversity between the two datasets. As shown in Table 1, LFM is generally less homogeneous than MPD in its overall track collection. This is expected since the LFM dataset is constructed over aggregate listening patterns while the MPD dataset is localized to specific playlists. Thus, these results can be attributed to the complexity of unifying a diverse set of tracks into a single playlist representation. In LFM, we can either control this complexity by significantly decreasing the number of seed tracks or providing a large enough subset that our method can find commonalities to anchor representations.

4.4.3 Convergence Analysis. The LARP model is trained in three consecutive stages, where each stage adds an additional loss to the training framework. In Figure 5, we showcase the trajectory of performance that is yielded with the addition of each individual loss. The purpose of this plot is to showcase the significant improvements that are achieved with the integration of relational pre-training. Notably, we showcase the performance from the 5 epochs which follow the best performance to highlight that the optimal performance has, indeed been achieved. As we can see from these plots, on both datasets, there is a significant advantage to adding relational awareness. Interestingly, we can see that in the initial epochs of the ℓ^{TPC} training, there is a drop in performance. This can be attributed to the training of the self-attention layers, which are initialized with random values, and slowly tuned over the subsequent epochs of LARP-*fusion* training. We note, that as mentioned in earlier sections, we can see that for MPD, which is specifically designed for the playlist continuation task, there are more advantages to the integration of TPC loss than for LFM. However, we believe this is expected given the high diversity of the LFM dataset.

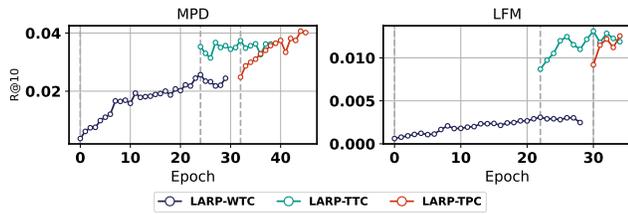


Figure 5: Convergence analysis of LARP. Grey, vertical dashed line signifies the end of one training stage.

5 RELATED WORK

In this section, we review related works to the *cold start problem* and *music understanding*.

5.1 Cold-Start Recommendation

The issue of cold-start has been an important focus within the recommendation domain [3, 32, 46, 54, 56–58, 65, 66]. Many approaches in the field borrow from the domain of context based learning for handling partial or sparse inputs [3, 65]. Alternatively, many works randomly mask interactive signals [48, 56, 66]. Finally, another branch of work uses content-based signals as a form of regularization or constraints during training [18, 46]. More recently, recent works have begun to integrate innovations from other tasks into the cold-start challenge such as contrastive learning [58], meta-learning [16, 31, 61], and zero-shot or multimodal learning [10, 15, 20, 32, 35]. Crucially, these methods differ from our approach in that they are either using pre-trained representation learning methods, single modality content representations, or limit their cold-start formulation to include only cold-start users (while maintaining a shared item pool).

Within the specific field of music recommendation, there have been many works that consider the topic of cold/warm start recommendation and playlist continuation tasks [2, 6, 44, 53, 55, 62, 63]. Notably, these works differ from ours in that they focus on either cold-start tracks [11, 53], users [6], or playlists [2, 44, 55, 62, 63] but not both playlists and tracks (as our setting suggests). Furthermore, to our knowledge, there have yet to be any approaches that design methods that use both cross-modal and relational information to achieve their results.

5.2 Music Understanding

In recent years, the music information retrieval community has used a variety of approaches to design music representation learning models. It has become a very standard practice within to apply state of the art vision models to images of the spectral content of audio samples using either convolutional neural networks [30, 37, 41, 42] or transformer-based architectures [59, 60] to learn representations. In the last two years, innovations in vision-language tasks, have also found success when applied to audio-language tasks. Most notably, [23, 26, 36, 37, 60] have all borrowed architecture from contrastive learning methods used in vision-language tasks and applied them to audio by replacing the visual encoder with an audio backbone model. Several works have also proposed the integration of collaborative signals into their training paradigms [1, 27, 41] by using various collaboration-informed heuristics for selecting

positive and negative pairs in their contrastive loss frameworks. Notably, our work differs from those presented in that none of these works integrate cross-modal *and* relational signals without the use of collaborative filtering.

6 CONCLUSION & FUTURE WORK

In this work, we proposed LARP, a multi-modal relation-aware representation learning model for cold-start playlist continuation. At the core of our approach, we argued that the true bottleneck of cold-start learning lies in the quality of the content module. To this end, LARP integrates multi-stage contrastive losses to learn multimodal and relation-aware item embeddings. First, we compared between the performance of LARP and a series of representation learning baseline methods, showing its superiority as an encoder. Second, we evaluated LARP in tandem with several canonical cold-start frameworks, showing that the quality of the embeddings produced by LARP significantly narrowed the gap between complex, parametrized recommendation frameworks (DropOutNet, CLCRec) and simple, unparameterized ones (ItemKNN). Finally, we demonstrated our method’s generalization capabilities by performing cross-dataset evaluation.

We see many opportunities for future work. From the music representation perspective, we envision many expansions of the playlist encoding capacity of our model. Currently, the playlist encoder is not used during inference. With some refinement, it could become an independent playlist encoder during inference. Additionally, our current track-playlist loss function does not integrate sequential awareness, which, if included, could potentially improve performance. Finally, while the application of this work is centered around the music domain, we believe that the contributions and outcomes of this work have significance for the broader domains of multimodal representation learning and cold start recommendation. Here, we extract relational signal from the playlists, because they are formed via explicit user feedback and thus represent reliable relational signals. However, the intuition behind these loss functions is applicable to many other domains such as books, podcasts, or video and is thus a major contribution. The flexibility of our relational training paradigm has potential for expansion into a multi-headed encoder that generates album-level, playlist-level, or even user-level representations each with its own dimension of relational awareness. From the recommendation perspective, we envision further exploration into the warm-start problem, where we imagine that sparse interaction can be used to guide our relational training paradigm to further refine its representational qualities. Ultimately, we hope that our work demonstrates both the importance of content representations in recommendation and the power of encoding relational awareness into content modules.

ACKNOWLEDGEMENTS

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

- [1] Pablo Alonso-Jiménez, Xavier Favory, Hadrien Foughmand, Grigoris Bourdalas, Xavier Serra, Thomas Lidy, and Dmitry Bogdanov. 2023. Pre-Training Strategies Using Contrastive Learning and Playlist Information for Music Classification and Similarity. In *ICASSP*. IEEE, 1–5.
- [2] Sebastiano Antenucci, Simone Boglio, Emanuele Chioso, Ervin Dervishaj, Shuwen Kang, Tommaso Scarlatti, and Maurizio Ferrari Dacrema. 2018. Artist-driven layering and user’s behaviour impact on recommendations in a playlist continuation scenario. In *RecSys Challenge*. ACM, 4:1–4:6.
- [3] Oren Barkan, Noam Koenigstein, Eylon Yogev, and Ori Katz. 2019. CB2CF: a neural multiview content-to-collaborative filtering model for completely cold item recommendations. In *RecSys*. ACM, 228–236.
- [4] Geoffray Bonnin and Dietmar Jannach. 2014. Automated Generation of Music Playlists: Survey and Experiments. *ACM Comput. Surv.* 47, 2 (2014), 26:1–26:35.
- [5] Théo Bontempelli, Benjamin Chapus, François Rigaud, Mathieu Morlon, Marin Lorant, and Guillaume Salha-Galvan. 2022. Flow Moods: Recommending Music by Moods on Deezer. In *RecSys*. ACM, 452–455.
- [6] Léa Briand, Guillaume Salha-Galvan, Walid Bendada, Mathieu Morlon, and Viet-Anh Tran. 2021. A Semi-Personalized System for User Cold Start Recommendation on Music Streaming Apps. In *KDD*. ACM, 2601–2609.
- [7] Rodrigo Castellon, Chris Donahue, and Percy Liang. 2021. Codified audio language modeling learns useful representations for music information retrieval. In *ISMIR*. 88–96.
- [8] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys challenge 2018: automatic music playlist continuation. In *RecSys*. ACM, 527–528.
- [9] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. In *ICASSP*. IEEE, 646–650.
- [10] Szu-Yu Chou, Yi-Hsuan Yang, Jyh-Shing Roger Jang, and Yu-Ching Lin. 2016. Addressing Cold Start for Next-song Recommendation. In *RecSys*. ACM, 115–118.
- [11] Szu-Yu Chou, Yi-Hsuan Yang, Jyh-Shing Roger Jang, and Yu-Ching Lin. 2016. Addressing Cold Start for Next-song Recommendation. In *RecSys*. ACM, 115–118.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.
- [13] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. *CoRR abs/2005.00341* (2020).
- [14] Ricardo Dias, Manuel J. Fonseca, and Ricardo Cunha. 2014. A User-centered Music Recommendation Approach for Daily Activities. In *CBRecSys@RecSys (CEUR Workshop Proceedings, Vol. 1245)*. CEUR-WS.org, 26–33.
- [15] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-Shot Recommender Systems. *CoRR abs/2105.08318* (2021).
- [16] Manqing Dong, Feng Yuan, Lina Yao, Xiwei Xu, and Liming Zhu. 2020. MAMO: Memory-Augmented Meta-Optimization for Cold-start Recommendation. In *KDD*. ACM, 688–697.
- [17] Eric Drott. 2018. Why the Next Song Matters: Streaming, Recommendation, Scarcity. *Twentieth-Century Music* 15 (10 2018), 325–357. <https://doi.org/10.1017/S1478572218000245>
- [18] Xiaoyu Du, Xiang Wang, Xiangnan He, Zechao Li, Jinhui Tang, and Tat-Seng Chua. 2020. How to Learn Item Representation for Cold-Start Multimedia Recommendation?. In *ACM Multimedia*. ACM, 3469–3477.
- [19] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP*. IEEE, 1–5.
- [20] Philip J. Feng, Pingjun Pan, Tingting Zhou, Hongxiang Chen, and Chuanjiang Luo. 2021. Zero Shot on the Cold-Start Problem: Model-Agnostic Interest Learning for Recommender Systems. In *CIKM*. ACM, 474–483.
- [21] Andres Ferraro, Jaehun Kim, Sergio Oramas, Andreas F. Ehmann, and Fabien Gouyon. 2023. Contrastive Learning for Cross-Modal Artist Retrieval. In *ISMIR*. 375–382.
- [22] Andres Ferraro, Yuntae Kim, Soohyeon Lee, Biho Kim, Namjun Jo, Semi Lim, Suyon Lim, Jungtaek Jang, Sehwan Kim, Xavier Serra, and Dmitry Bogdanov. 2021. Melon Playlist Dataset: A Public Dataset for Audio-Based Playlist Generation and Music Tagging. In *ICASSP*. IEEE, 536–540.
- [23] Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. 2023. LLark: A Multimodal Foundation Model for Music. *CoRR abs/2310.07160* (2023).
- [24] Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: Audio Spectrogram Transformer. In *Interspeech*. ISCA, 571–575.
- [25] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *ICDM*. IEEE Computer Society, 263–272.
- [26] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. MuLan: A Joint Embedding of Music Audio and Natural Language. In *ISMIR*. 559–566.
- [27] Qingqing Huang, Aren Jansen, Li Zhang, Daniel P. W. Ellis, Rif A. Saurous, and John R. Anderson. 2020. Large-Scale Weakly-Supervised Content Embeddings for Music Recommendation and Tagging. In *ICASSP*. IEEE, 8364–8368.
- [28] Laurie Jakobsen. 2016. Playlists Overtake Albums in Listenership, says Loop Study. <https://musicbiz.org/news/playlists-overtake-albums-listenership-says-loop-study/>
- [29] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2021. Slow-Fast Auditory Streams for Audio Recognition. In *ICASSP*. IEEE, 855–859.
- [30] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE ACM Trans. Audio Speech Lang. Process.* 28 (2020), 2880–2894.
- [31] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *KDD*. ACM, 1073–1082.
- [32] Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, Yang Yang, and Zi Huang. 2019. From Zero-Shot Learning to Cold-Start Recommendation. In *AAAI*. AAAI Press, 4189–4196.
- [33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 12888–12900.
- [34] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*. 9694–9705.
- [35] Yunshan Ma, Yingzhi He, Wenjun Zhong, Xiang Wang, Roger Zimmermann, and Tat-Seng Chua. 2024. CIRP: Cross-Item Relational Pre-training for Multimodal Product Bundling.
- [36] Yunshan Ma, Xiaohao Liu, Yinwei Wei, Zhulin Tao, Xiang Wang, and Tat-Seng Chua. 2024. Leveraging Multimodal Features and Item-level User Feedback for Bundle Construction. In *WSDM*. ACM, 510–519.
- [37] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Learning Music Audio Representations Via Weak Language Supervision. In *ICASSP*. IEEE, 456–460.
- [38] Matthew C. McCallum, Filip Korzeniowski, Sergio Oramas, Fabien Gouyon, and Andreas F. Ehmann. 2022. Supervised and Unsupervised Learning of Audio Representations for Music Understanding. In *ISMIR*. 256–263.
- [39] Brian McFee and Gert R. G. Lanckriet. 2012. Hypergraph Models of Playlist Dialects. In *ISMIR*. FEUP Edições, 343–348.
- [40] Alessandro B. Melchiorre, Navid Rekasaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Inf. Process. Manag.* 58, 5 (2021), 102666.
- [41] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. 2018. Representation Learning of Music Using Artist Labels. In *ISMIR*. 717–724.
- [42] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. 2018. End-to-end Learning for Music Audio Tagging at Scale. In *ISMIR*. 637–644.
- [43] Robert Prey. 2020. Locating Power in Platformization: Music Streaming Playlists and Curatorial Power. *Social Media + Society* 6 (04 2020), 205630512093329. <https://doi.org/10.1177/2056305120933291>
- [44] Vasily Rubtsov, Mikhail Kamenshchikov, Ilya Valyaev, Vasily Leksin, and Dmitry I. Ignatov. 2018. A hybrid two-stage recommender system for automatic playlist continuation. In *RecSys Challenge*. ACM, 16:1–16:4.
- [45] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. ACM, 285–295.
- [46] Martin Saveski and Amin Mantrach. 2014. Item cold-start recommendations: learning local collective embeddings. In *RecSys*. ACM, 89–96.
- [47] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *Int. J. Multim. Inf. Retr.* 7, 2 (2018), 95–116.
- [48] Shaoyun Shi, Min Zhang, Xinxing Yu, Yongfeng Zhang, Bin Hao, Yiqun Liu, and Shaoping Ma. 2019. Adaptive Feature Sampling for Recommendation with Missing Content Feature Values. In *CIKM*. ACM, 1451–1460.
- [49] Janne Spijkervet and John Ashley Burgoyne. 2021. Contrastive Learning of Musical Representations. In *ISMIR*. 673–681.
- [50] Darko Stanisljevic. 2020. The impact of Spotify features on music discovery in the streaming platform age.
- [51] Marco Tagliasacchi, Beat Gfeller, Felix de Chaumont Quitry, and Dominik Roblek. 2020. Pre-Training Audio Representations With Self-Supervision. *IEEE Signal Process. Lett.* 27 (2020), 600–604.
- [52] Roberto Turrin, Massimo Quadrona, Andrea Condorelli, Roberto Pagano, and Paolo Cremonesi. 2015. 30Music Listening and Playlists Dataset. In *RecSys Posters (CEUR Workshop Proceedings, Vol. 1441)*. CEUR-WS.org.
- [53] Aaron van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf

- [54] Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, and Hugo Larochelle. 2017. A Meta-Learning Perspective on Cold-Start Recommendations for Items. In *NIPS*. 6904–6914.
- [55] Maksims Volkovs, Himanshu Rai, Zhaoyue Cheng, Ga Wu, Yichao Lu, and Scott Sanner. 2018. Two-stage Model for Automatic Playlist Continuation at Scale. In *RecSys Challenge*. ACM, 9:1–9:6.
- [56] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. 2017. DropoutNet: Addressing Cold Start in Recommender Systems. In *NIPS*. 4957–4966.
- [57] Wenjie Wang, Xinyu Lin, Lihui Wang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2023. Equivariant Learning for Out-of-Distribution Cold-start Recommendation. In *ACM Multimedia*. ACM, 903–914.
- [58] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-Start Recommendation. In *ACM Multimedia*. ACM, 5382–5390.
- [59] Minz Won, Keunwoo Choi, and Xavier Serra. 2021. Semi-supervised Music Tagging Transformer. In *ISMIR*. 769–776.
- [60] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP*. IEEE, 1–5.
- [61] Zhenchao Wu and Xiao Zhou. 2023. M2EU: Meta Learning for Cold-start Recommendation via Enhancing User Preference Estimation. In *SIGIR*. ACM, 1158–1167.
- [62] Hojin Yang, Yoonki Jeong, Minjin Choi, and Jongwuk Lee. 2018. MMCF: Multimodal Collaborative Filtering for Automatic Playlist Continuation. In *RecSys Challenge*. ACM, 11:1–11:6.
- [63] Hamed Zamani, Markus Schedl, Paul Lamere, and Ching-Wei Chen. 2019. An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation. *ACM Trans. Intell. Syst. Technol.* 10, 5 (2019), 57:1–57:21.
- [64] Eva Zangerle, Martin Pichl, Wolfgang Gassler, and Günther Specht. 2014. #now-playing Music Dataset: Extracting Listening Behavior from Twitter. In *WISMM*. ACM, 21–26.
- [65] Mi Zhang, Jie Tang, Xuchen Zhang, and Xiangyang Xue. 2014. Addressing cold start in recommender systems: a semi-supervised co-training algorithm. In *SIGIR*. ACM, 73–82.
- [66] Ziwei Zhu, Shahin Sefati, Parsa Saadatpanah, and James Caverlee. 2020. Recommendation for New Users and New Items via Randomized Training and Mixture-of-Experts Transformation. In *SIGIR*. ACM, 1121–1130.

A DETAILED LOSS FORMULATIONS

(L1) WTC: Within-Track Contrastive Loss. As mentioned in 3.2, the L1 loss is formulated for cross-modal alignment within a track. As defined in Equation 1, this loss performs contrastive learning on the single-track level. Given a track, s_i , consisting of a language-audio pair, (t_{s_i}, a_{s_i}) , we formally define this loss as:

$$\ell_{s_i}^{\text{WTC}} = \text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}). \quad (13)$$

More specifically, the contrastive loss is formulated as:

$$\text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}) = \frac{1}{2} [\text{CE}(y_i^{\text{a2t}}, \hat{y}_i^{\text{a2t}}) + \text{CE}(y_i^{\text{t2a}}, \hat{y}_i^{\text{t2a}})] \quad (14)$$

where softmax-normalized audio-to-text (a2t) and text-to-audio (t2a) similarities are defined by

$$\begin{aligned} \hat{y}_i^{\text{a2t}} &= \frac{\exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_i}) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_n}) / \tau)}, \\ \hat{y}_i^{\text{t2a}} &= \frac{\exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_i}) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_n}) / \tau)}, \end{aligned} \quad (15)$$

such that $\text{sim}(\cdot)$ is the cosine similarity function, τ is a temperature hyper-parameter, $y_i^{\text{a2t}}, y_i^{\text{t2a}} \in \{0, 1\}$ are one-hot ground-truth similarity vectors indicating the correct language-audio pairing, and s_n is a negative sample, drawn from the momentum queue.

(L2) TTC: Track-Track Contrastive Loss. The second loss extends the formulation of the WTC (L1) loss to perform track-track

contrastive learning. For two tracks $s_i = (t_{s_i}, a_{s_i})$ and $s_j = (t_{s_j}, a_{s_j})$, which share a parent playlist p_k , this loss is formulated as

$$\ell_{s_i, s_j}^{\text{TTC}} = \frac{1}{2} [\text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_j}) + \text{Contrast}(\mathbf{t}_{s_i}, \mathbf{a}_{s_j})]. \quad (16)$$

Here, the each contrastive loss pairing is formulated as:

$$\begin{aligned} \text{Contrast}(\mathbf{a}_{s_i}, \mathbf{t}_{s_j}) &= \frac{1}{2} [\text{CE}(y_{ij}^{\text{a2t}}, \hat{y}_{ij}^{\text{a2t}}) + \text{CE}(y_{ji}^{\text{t2a}}, \hat{y}_{ji}^{\text{t2a}})], \\ \text{Contrast}(\mathbf{t}_{s_i}, \mathbf{a}_{s_j}) &= \frac{1}{2} [\text{CE}(y_{ij}^{\text{t2a}}, \hat{y}_{ij}^{\text{t2a}}) + \text{CE}(y_{ji}^{\text{a2t}}, \hat{y}_{ji}^{\text{a2t}})], \end{aligned} \quad (17)$$

where softmax-normalized audio-to-text (a2t) and text-to-audio (t2a) similarities are defined by

$$\begin{aligned} \hat{y}_{ij}^{\text{a2t}} &= \frac{\exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_j}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_n}) / \tau)}, \\ \hat{y}_{ji}^{\text{a2t}} &= \frac{\exp(\text{sim}(\mathbf{a}_{s_j}, \mathbf{t}_{s_i}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{a}_{s_j}, \mathbf{t}_{s_n}) / \tau)}, \\ \hat{y}_{ji}^{\text{t2a}} &= \frac{\exp(\text{sim}(\mathbf{t}_{s_j}, \mathbf{a}_{s_i}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{t}_{s_j}, \mathbf{a}_{s_n}) / \tau)}, \\ \hat{y}_{ij}^{\text{t2a}} &= \frac{\exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_j}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_n}) / \tau)}, \end{aligned} \quad (18)$$

such that $\text{sim}(\cdot)$ is the cosine similarity function, τ is a temperature hyper-parameter, and s_n is a negative sample, drawn from the momentum queue. And the one-hot ground-truth similarity vectors indicating the correct language-audio pairing, are calculated as:

$$\begin{aligned} y_{ij}^{\text{a2t}} &= \left\{ o \in \{0, 1\}^B : \sum_{v=1}^B o_v = O_{iv} \right\}, \\ y_{ji}^{\text{a2t}} &= \left\{ o \in \{0, 1\}^B : \sum_{v=1}^B o_v = O_{jv} \right\}, \\ y_{ij}^{\text{t2a}} &= \left\{ o \in \{0, 1\}^B : \sum_{v=1}^B o_v = O_{iv} \right\}, \\ y_{ji}^{\text{t2a}} &= \left\{ o \in \{0, 1\}^B : \sum_{v=1}^B o_v = O_{jv} \right\}, \end{aligned} \quad (19)$$

where B is the batch size and $O_{N \times N}$ is the homogeneous, track-track co-occurrence graph defined in Section 2.

(L3) TPC: Track-Playlist Contrastive Loss. The third loss extends the formulation of TTC (L2) loss to perform track-playlist contrastive learning. For a track, s_i , its parent playlist, p_k , and a set of neighbour tracks that share the parent playlist, $J = \{s_j \in p_k \setminus s_i\}$, define this loss as:

$$\ell_{p_k, s_i}^{\text{TPC}} = \frac{1}{2} [\text{Contrast}(\mathbf{a}_{p_k}, \mathbf{t}_{s_i}) + \text{Contrast}(\mathbf{t}_{p_k}, \mathbf{a}_{s_i})], \quad (20)$$

where the contrastive pairing is defined by

$$\begin{aligned} \text{Contrast}(\mathbf{a}_{p_k}, \mathbf{t}_{s_i}) &= \frac{1}{2} [\text{CE}(y_{ki}^{\text{a2t}}, \hat{y}_{ki}^{\text{a2t}}) + \text{CE}(y_{ik}^{\text{t2a}}, \hat{y}_{ik}^{\text{t2a}})], \\ \text{Contrast}(\mathbf{t}_{p_k}, \mathbf{a}_{s_i}) &= \frac{1}{2} [\text{CE}(y_{ki}^{\text{t2a}}, \hat{y}_{ki}^{\text{t2a}}) + \text{CE}(y_{ik}^{\text{a2t}}, \hat{y}_{ik}^{\text{a2t}})], \end{aligned} \quad (21)$$

where softmax-normalized audio-to-text (a2t) and text-to-audio (t2a) similarities are defined by

$$\begin{aligned} \hat{y}_{ki}^{a2t} &= \frac{\exp(\text{sim}(\mathbf{a}_{p_k}, \mathbf{t}_{s_i}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{a}_{p_k}, \mathbf{t}_{s_n}) / \tau)}, \\ \hat{y}_{ik}^{a2t} &= \frac{\exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{p_k}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{a}_{s_i}, \mathbf{t}_{s_n}) / \tau)}, \\ \hat{y}_{ik}^{t2a} &= \frac{\exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{p_k}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{t}_{s_i}, \mathbf{a}_{s_n}) / \tau)}, \\ \hat{y}_{ki}^{t2a} &= \frac{\exp(\text{sim}(\mathbf{t}_{p_k}, \mathbf{a}_{s_i}) / \tau)}{\sum_{n=1}^N \exp(\text{sim}(\mathbf{t}_{p_k}, \mathbf{a}_{s_n}) / \tau)}, \end{aligned} \quad (22)$$

where the playlist embeddings, \mathbf{a}_{p_k} , \mathbf{t}_{p_k} , are defined by the output of the *fusion* layer defined in Section 3.2 as:

$$\begin{aligned} \mathbf{a}_{p_k} &= \text{Self-Att}(\{\mathbf{a}_{s_j} : s_j \in J\}), \\ \mathbf{t}_{p_k} &= \text{Self-Att}(\{\mathbf{t}_{s_j} : s_j \in J\}), \end{aligned} \quad (23)$$

and that $\text{sim}(\cdot)$ is the cosine similarity function, τ is a temperature hyper-parameter, and s_n is a negative sample, drawn from the momentum queue. More formally, if we consider $A_J = \{\mathbf{a}_{s_j} : s_j \in J\}$ and $T_J = \{\mathbf{t}_{s_j} : s_j \in J\}$ we can define this attention layer as:

$$\begin{aligned} \text{Self-Att}(A_J) &= \text{softmax}\left(\frac{A_J W_Q^T A_J W_K^T}{\sqrt{d}}\right) A_J W_V^T \\ \text{Self-Att}(T_J) &= \text{softmax}\left(\frac{T_J W_Q^T T_J W_K^T}{\sqrt{d}}\right) T_J W_V^T \end{aligned} \quad (24)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{|J| \times d}$, $d = 256$. Meanwhile, the one-hot ground-truth similarity vectors indicating the correct language-audio pairing, are calculated as:

$$\begin{aligned} y_{ki}^{a2t} &= \left\{ o \in \{0, 1\}^B : \sum_v r_v = R_{iv} \right\}, \\ y_{ik}^{a2t} &= \left\{ o \in \{0, 1\}^B : \sum_v r_v = R_{jv} \right\}, \\ y_{ik}^{t2a} &= \left\{ o \in \{0, 1\}^B : \sum_v r_v = R_{iv} \right\}, \\ y_{ki}^{t2a} &= \left\{ o \in \{0, 1\}^B : \sum_v r_v = R_{jv} \right\}, \end{aligned} \quad (25)$$

where B is the batch size and $R_{M \times N}$ is the bipartite, playlist-track interaction graph defined in Section 2.

B COLD-START IMPLEMENTATIONS

B.1 ItemKNN [45]

B.1.1 Formulation. Given the pre-trained content features $\mathbf{E}_{\mathcal{S}}$ for all tracks in testing set and the playlist representation $\mathbf{e}_{\bar{p}_q}$, we calculate the scores $\hat{y}_{\bar{p}_q}$ for the input partial playlist \bar{p}_q with inner product to find held-out tracks. Formally:

$$\hat{\mathcal{S}} = \arg \max_{\mathcal{S}} (\hat{y}_{\bar{p}_q}), \quad \hat{y}_{\bar{p}_q} = \mathbf{e}_{\bar{p}_q} \cdot \mathbf{E}_{\mathcal{S}}, \quad (26)$$

where $\hat{\mathcal{S}}$ are the predicted tracks. For the final recommendation, we keep the top-k highest tracks, thus selecting $|\hat{\mathcal{S}}| = K$.

B.2 DropoutNet. [56]

B.2.1 Formulation. This method uses two stages of training.

- **WMF model:** Given a playlist-track interaction matrix $\mathbf{R}_{M \times N}$, WMF learns the latent collaborative embeddings, $\mathbf{Z}_{\mathcal{S}}$ and $\mathbf{Z}_{\mathcal{P}}$.
- **DropoutNet model:** Given the learned collaborative embeddings, $\mathbf{Z}_{\mathcal{S}}$ and $\mathbf{Z}_{\mathcal{P}}$, content features $\mathbf{E}_{\mathcal{S}}$ and $\mathbf{E}_{\mathcal{P}}$, generated from $\Phi(\cdot)$ DropoutNet concatenates them and passes trains DNN models, $f_{\bar{p}}(\cdot)$ and $f_{\mathcal{S}}(\cdot)$, with mean squared error (MSE) to recover these concealed entries and perform score prediction.

During testing, where our cold-start setting precludes the possibility of calculating collaborative features, we follow the original paper's formulation to generate collaborative features by averaging them from the training set:

$$\tilde{\mathbf{z}}_p = \text{mean}(\{\mathbf{z}_p : p \in \mathcal{P}\}), \quad \tilde{\mathbf{z}}_s = \text{mean}(\{\mathbf{z}_s : s \in \mathcal{S}\}), \quad (27)$$

Thus, we input the partial playlist \bar{p}_q in Equation ?? and obtain the predicted scores $\hat{y}_{\bar{p}_q}$ for \bar{p}_q . Final predictions are generated using:

$$\hat{y}_{\bar{p}_q} = f_{\bar{p}}(\mathbf{e}_{\bar{p}} \parallel \tilde{\mathbf{z}}_p) \cdot f_{\mathcal{S}}(\mathbf{E}_{\mathcal{S}} \parallel \tilde{\mathbf{Z}}_{\mathcal{S}}), \quad (28)$$

where $\tilde{\mathbf{Z}}_{\mathcal{S}}$ is a matrix formed by row-wise stacking of $\tilde{\mathbf{z}}_s$ to match the shape requirement of concatenation with $\mathbf{E}_{\mathcal{S}}$.

B.2.2 Hyperparameter Settings. We set the learning rate to 0.005, L2 regularization to 0.1 and utilize the SGD optimizer with momentum of 0.9 for both WMF and DropoutNet training. The embedding size of WMF is 64 and the dimension of transformed features for DropoutNet is 256. Additionally, we use three-layer MLPs to implement each of the transformation DNN models, $f_{\bar{p}}$ and $f_{\mathcal{S}}$, and a ratio of 0.2 for the feature dropout.

B.3 CLCRec [58]

B.3.1 Formulation. Unlike DropoutNet, CLCRec initializes the embeddings and learns them during the training procedure. Here, its training loss is:

$$\ell_{\text{CLCRec}} = \text{Contrast}(\mathbf{z}_p, \mathbf{z}_{s_i}) + \text{Contrast}(f(\mathbf{e}_{s_i}), f(\mathbf{e}_{s_j})), \quad (29)$$

where $f(\cdot)$ is a DNN model for transforming the pre-trained content features. CLCRec randomly chooses between the content features and collaborative embeddings using a predefined probability value as it learns to maximize the correlation between learned content features and latent playlist representations. The score prediction is similar with ItemKNN, but equipped with an additional DNN model, which can be denoted as:

$$\hat{y}_{\bar{p}_q} = f(\mathbf{e}_{\bar{p}_q}) \cdot f(\mathbf{E}_{\mathcal{S}}). \quad (30)$$

B.3.2 Hyperparameter Settings. We tune the learning rate to 0.001, L2 regularization to 0.1 and use the Adam optimizer. The temperature for contrastive learning is 2.0 and we set the probability of replacement to 0.5.