Video to Music Moment Retrieval

Zijie Xin[1*] Minquan Wang[2] Ye Ma[2] Bo Wang[2] Quan Chen[2] Peng Jiang[2]
Xirong Li[1†]
[1]MoE Key Lab of DEKE, Renmin University of China
[2]Kuaishou Technology
https://rucmm.github.io/VMMR

Abstract

Adding proper background music helps complete a short video to be shared. Towards automating the task, previous research focuses on video-to-music retrieval (VMR), aiming to find amidst a collection of music the one best matching the content of a given video. Since music tracks are typically much longer than short videos, meaning the returned music has to be cut to a shorter moment, there is a clear gap between the practical need and VMR. In order to bridge the gap, we propose in this paper video to music moment retrieval (VMMR) as a new task. To tackle the new task, we build a comprehensive dataset Ad-Moment which contains 50K short videos annotated with music moments and develop a two-stage approach. In particular, given a test video, the most similar music is retrieved from a given collection. Then, a Transformer based music moment localization is performed. We term this approach Retrieval and Localization (ReaL). Extensive experiments on real-world datasets verify the effectiveness of the proposed method for VMMR.

1. Introduction

With the widespread adoption of smart mobile devices and the increasing emphasis that ordinary people place on documenting their daily lives, the development of short videos has been progressing rapidly. At the same time, people are not just limited to shooting short videos; they are more enthusiastic about making their videos look more complete by adding background music, sound effects, stickers, and other elements to gain more views and comments.

In this paper, we tackle a new challenge of video-to-music moment retrieval (VMMR) which can be formulated as selecting the most relevant music moment for videos from a given music corpus. This is a completely new field.

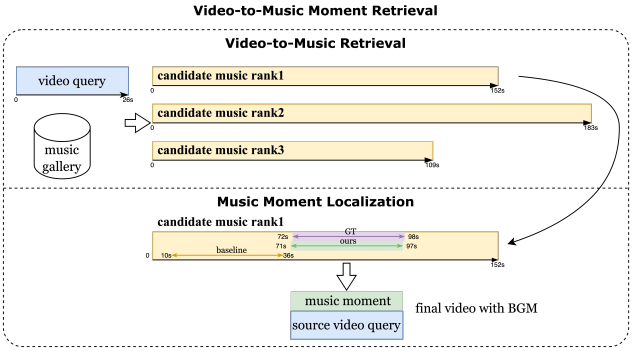


Figure 1. **Proposed Video-to-Music Moment Retrieval (VMMR) task *versus* the conventional video-to-music retrieval (VMR) task**.

Although there has been a significant amount of work focusing on Music Retrieval (MR) [21, 33], they all consider the music to be a single entity and overlook the differences between various moments within one piece of music. Although these musical moments share a similar musical style, they differ in terms of emotional expression and rhythm. For example, Taylor Swift's famous song "Love Story" employs a country music style in the intro, with a slow tempo, but switches to a rock style in the chorus, making it suitable as background music for videos with rapid rhythm changes. Overall, this song is more appropriate for lyrical videos and not suitable for tense or horror-themed videos.

However, the existing datasets related to MR tasks don't support us in training for moment retrieval. For example, the commonly used HIMV dataset [10] does not have annotations at the moment level. Annotating music moments is a very challenging task because the start and end times of moments might be precise to decimals, such as 3.1 seconds. Moreover, a piece of music may contain multiple moments with identical rhythms but different lyrics. This is a widespread phenomenon in music composition. Therefore, if we directly extract a segment from a video, for instance, from 10.1s to 30.5s, the corresponding music mo-

*This work is done when Zijie Xin is an intern at Kuaishou Technology.
†Corresponding authors: Xirong Li (xirong@ruc.edu.cn)

ment might not be the only correct annotation, because the segment from 100.1s to 130.5s could have the same rhythm. To address this issue, we designed a data collection pipeline. By simply inputting the video and the corresponding complete music track, it uses lyrics elimination and rhythm matching to output all possible music moments.

Although VMMR is a new task, Video Corpus Moment Retrieval(VCMR) has already garnered significant attention from researchers. These methods have provided us with great inspiration and insight. Escorcia et al. [5] firstly focus on VCMR and devise a ranking-based clip-query alignment model. Existing methods for VCMR fall into two categories based on how they address the learning of retrieval and localization. We follow the two-stage solution and propose a framework named ReaL for VMMR. The core target of the first stage is learning two independent encoders for query video and gallery music by contrastive learning. $\mathbf{top}k$ music can be selected from the whole corpus which is related to query video in latent space. In the second stage, we apply a cross-model attention module followed by a DETR-based structure to predict music moments for every music output from the first stage. At last, some of these predicted moments are elected as the most related moment to query video.

The main contributions of this paper are three-fold:

- To the best of our knowledge, we are the first to propose the Video-to-Music Moment Retrieval task to achieve music retrieval and moment localization from a music corpus to serve as a given video's background music.
- We build a benchmark dataset Ad–Moment by a weakly supervised timestamp collection pipeline, to facilitate the investigation of the VMMR task.
- We propose a two-stage framework ReaL and adapt two straightforward baseline methods to tackle the task, followed by a comprehensive analysis.

2. Related Work

Video-to-Music Retrieval (VMR) aims to identify the most appropriate BGM for a given video content and style from the music library. Early works on the task achieve cross-modal alignment by utilizing metadata, which offers key information about the content of videos or music quickly and simply, such as color histogram of album covers [1] or emotion tags [15, 32]. Nevertheless, metadata is not always accessible or comprehensive, especially for large data.

Recently, content-based models have taken the lead in video-to-music retrieval benchmarks. VM-NET model [10] and Seg-VM-Net [28] employed the two-branch deep network to associate videos and music by considering both inter- and intra-modal relationships. CMMVR [27] enhanced this approach by incorporating learned audio features rather than relying on handcrafted ones. Unfortunately, they primarily focus on matching with video content, making it challenging to model relationships among different snippets and capture temporal information in music, thereby limiting the ability to identify relevant music segments and align them appropriately with the video.

MVPt [33] and ViML [22] employ Transformers with self-supervised contrastive learning to improve the long-range temporal context modeling. CMVAE [35] introduces a cross-generation strategy designed to better align latent embeddings of videos and music. Some methods focus on learning distinctive properties of music such as beat [4], emotion [9] and rhythm [21]. However, these methods neglect the precise requirement in duration between visual and musical modalities. Consequently, these methods can only trim the music retrieved through fixed editing rules. To address this practical issue, we propose VMMR to integrate retrieval and localization. We also constructed a dataset and developed a two-stage framework ReaL to find the most suitable music segments as BGM.

Single Video Moment Retrieval (SVMR) is a language-based task whose goal is to localize a specific moment within an untrimmed video that corresponds to a given text description. The current mainstream approaches to SVMR can be categorized into proposal-based methods [25, 39, 41] and proposal-free methods [3, 36, 40]. However, most of them rely on preprocessing steps (e.g., proposal generation) or postprocessing steps (e.g., non-maximum suppression) that are hand-crafted, making them unsuitable for end-to-end training. Inspired by works in object detection [2, 12] and video action detection [24, 38], some DETR-based methods [14, 17, 23] treat moment retrieval as a direct set prediction problem. In these methods, Moment-DETR [14] takes video and user query representations as inputs and directly produces moment coordinates in an end-to-end manner, thereby eliminating the need for any manually-designed pre- or post-processing steps. Unlike SVMR, our VMMR needs the localization model to select the most suitable segment from a complete music track based on the given video. More importantly, the ideal music segment should exactly match the videos duration. To achieve this, we incorporated a mask into the context module, indirectly ensuring the accuracy of duration prediction.

Video Corpus Moment Retrieval (VCMR) is to identify relevant video moments corresponding to a given query language description from a large collection of untrimmed videos, which are often paired with timestamped subtitles. At present, there is limited research addressing the VCMR task. Existing approaches can be divided into two categories: one-stage methods [13, 16] and two-stage methods [11, 37]. One-stage methods train both video retrieval and moment localization heads within a single model. In contrast, two-stage methods first train a video retrieval head with one model, followed by training a moment localization head using a separate model. Our VMMR task is similar to

Data split	#Musics	*Duration* (s)	#Videos	*Duration* (s)	#Moments
Total	4,050	138.9±69.6	53,194	23.9±10.7	35,393
Training	3,496	138.3±69.4	49,194	24.0±10.7	31,660
Validation	2,000	139.6±70.0	2,000	22.8±10.8	2,000
Test	2,000	139.9±70.1	2,000	22.6±10.7	2,000

Table 1. **Overview of the Ad–Moment dataset**.

VCMR but more challenging due to the difficulty in distinguishing music data rather than video frames. The shorter music clips also carry less semantic information, adding to the task's complexity. To address this, we propose a two-stage ReaL framework that separately tackles the retrieval and localization subtasks, achieving strong performance.

3. VMMR: Task and Dataset

3.1. Task Definition

We define VMMR as follows. Consider that we have a set of untrimmed complete music $\mathcal{M}$ and a set of videos $\mathcal{V}$. Given a video query $v \in \mathcal{V}$, the goal is to *retrieval* a music $m^* \in \mathcal{M}$ that conveys a suitable emotional tone to the content in the video query and *localize* the most appropriate music moment w^* in the music m^* by providing the start and end time point τ^s and τ^e.

To better address the task, we designed a two-stage framework. In the first stage, *Music Retrieval*, we select the $\mathbf{top}k$ music tracks from the corpus $\mathcal{M}$ that exhibit semantic similarities with v. Subsequently, in the second stage, *Moment Localization*, we pinpoint the most relevant moments from each of the $\mathbf{top}k$ music tracks.

3.2. Dataset Curation

Due to the massive demand for advertising short videos and the cost constraints from businesses, there's an urgent need for automated and precise music recommendations in ad video production. To train such a model in a supervised manner, it is necessary to have a dataset of corresponding quadruple (*short video, music, moment start timestamp, moment end timestamp*) data, abbreviated as (v, m, τ^s, τ^e). While large-scale datasets of videos with paired music are available, it is not easy to find datasets which also contain high-quality annotated moments with start and end time points for the corresponding music tracks. As a result, to tackle this new task, we devise a moment construction pipeline that generates moment timestamps based on video audio and corresponding available music in a weakly supervised mode. And we automatically construct an advertising domain dataset Ad–Moment. In the following part, we discuss the pipeline through steps of *data collection*, *data cleaning*, and *moment localization*.

Raw Data Gathering. To obtain a representative and diverse subset of short videos for advertising, we have authorized downloading videos from short video platforms as

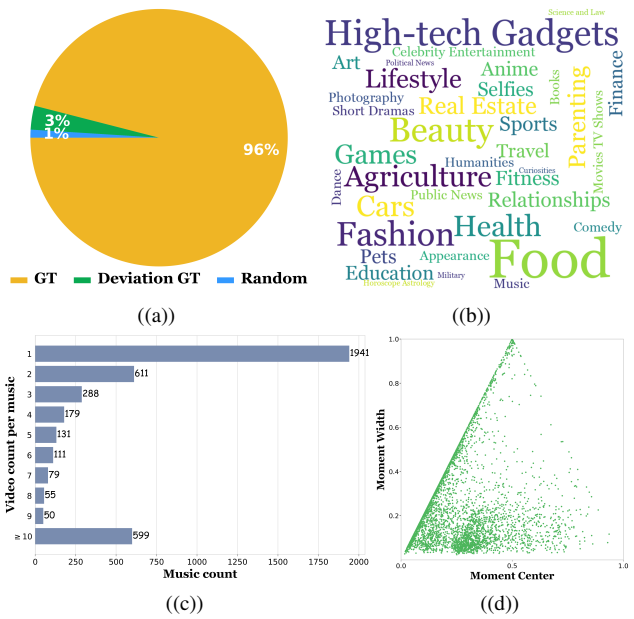


Figure 2. **Data analysis and statistics on Ad–Moment.**

candidates. For this study, we gathered nearly 1 million diverse short videos uploaded between Jan 2023 and Sep 2023. Besides the MP4 video and corresponding MP3 background music files, we also downloaded varied metadata, including titles, tags, and protobuf files containing playback information, if available. Specific annotated data instances are illustrated in the supplementary material.

Automated Data Cleaning. As the raw data is quite diverse with varied annotation quality, automated data cleaning is necessary to remove videos that cannot accurately identify BGM music or have meaningless visual content and low view counts. After removing low-quality data, we filtered videos related to the advertising themes based on video and creator tags. At this point, we obtain approximately 80k videos with duration ranging from 2.1 seconds to 753.6 seconds, with a mean value of 29.1 seconds and a median of 23.8. Due to the nature of VMMR, which involves recommending appropriate music moments for short videos, we constrain the background music to consist of high-quality segments from complete musical compositions. Through the manual review of the raw data, we empirically select videos that are associated with only one background music track. Besides, the video duration is restricted from 5 to 50 seconds, while the duration of the complete music is constrained from 30 to 240 seconds. In the end, we select 53,194 short videos, each corresponding to 4,050 complete music tracks.

Ground truth Moment Localization. After filtering out high-quality advertising short videos with paired music, the next goal is to construct fine-grained annotations for music moment time points. However, this is quite challenging because it is difficult for humans to distinguish differences

in music at the millisecond level. To that end, we develop a weakly supervised multi-modal collection pipeline to obtain highly confident timestamp data. It consists of two core steps: **Background Sound Extraction** and **Moment Localization**. In the pipeline, it is essential to annotate the music used in a given video. Then, the pipeline will provide precise temporal location information.

Background Sound Extraction. At first, we extract the raw audio from the video and the corresponding complete raw music, which contains the musical rhythms, lyrics, and dialogue. Demucs[31] is a renowned toolbox for Music Source Separation tasks. It enables the separation of the audio into the vocal track and non-vocal track components. And it splits the raw music into vocal and instrumental tracks. In our scenario, we only use the non-vocal component to avoid potential shortcuts presented in the vocal track from the video, which is not permissible in our cross-modal learning task.

Moment Localization. In the next step, we use the non-vocal track, cleared of vocal noise interference, and the instrumental track from the music for timestamp localization. Specifically, we extract the waveform data from the two audio using the torchaudio package and create a sliding window on the instrumental waveform data with a window length matching that of the non-vocal track and a step size of 0.1 seconds. Then, we calculate the cosine similarity between the waveform data features to obtain a similarity ranking list. Finally, we identify the window with the highest similarity as the ground-truth moment corresponding to the raw audio.

Analysis and statistics. To further evaluate the quality of timestamp annotations, we conducted a user study. Specifically, we randomly select GT moments, moments with a 3-second deviation with GT, and random moments from the corresponding complete music videos. These three categories are then matched with the video audio to determine the most accurate corresponding moment. Six individuals were assigned to evaluate the same 200 videos and calculate the average of their results. As shown in Fig. 2(a), over 95% of the matches aligned with the ground truth (GT). The word cloud of specific video tags in Fig. 2(b) shows the video content diversity.

Tab. 1 displays the overall data statistics. We separately allocated 2,000 videos with different music as the validation set and test set, and the remaining data as the training set. The average duration and standard deviation of all the music are 138.9 seconds and 69.6 seconds, respectively. The videos have an average duration of 23.9 seconds with a standard deviation of 10.7 seconds. As shown in Fig. 2(c), the video count per music exhibits a clear long-tail distribution, with the most frequently used music appearing in 1,941 videos. Fig. 2(d) shows that some moments are selected from music beginning. Furthermore, the length of most moments is relatively small compared to the entire music track, indicating a tendency for videos to use segments of the music rather than the full track.

4. Proposed Method for VMMR

We propose ReaL, a two-stage method that facilitates global video-music information learning for video-to-music retrieval and ensures cross-modal alignment and temporal localization within the retrieved music.

4.1. Stage I: Video-to-Music Retrieval

Feature Representation. The initial features of video and music are extracted from pre-trained networks. We will introduce the feature extraction and contextual modeling for video and music separately.

Music Feature. Given a music track, we first pad it to match the longest duration of music in the dataset with zero, which is 240 seconds, and load it by torchaudio package in the sample rate of $16k$. Next, we divide it into S overlapping audio segment $m_k = \{m_k^i\}_{i=1}^S$, where S represents the number of segments obtained from the padding music. Each segment is 10 seconds long, with a window stride of 5 seconds. We encode each music segment m_k^i into a feature vector $x_{m_k}^i$ using a pre-trained AST model [7], applied to the audio spectrograms. The music features are extracted offline without fine-tuning the AST model. These features can be formulated as $\mathbf{x}_{m_k} = \{\mathbf{x}_{m_k}^i\}_{i=1}^S$, where $\mathbf{x}_{m_k}^i \in d^{768}$.

Video Feature. Given an input video query sequence, we extract all frames at a rate of 1 fps, denoted as $v = \{v^i\}_{i=1}^W$, where W is the total number of frames in the video. We then use a pre-trained CLIP model [29] to extract visual feature $\mathbf{x}_v^i$ for each video frame v^i. The pre-trained CLIP model is not fine-tuned. The video feature are represented as $\mathbf{x}_v = \{\mathbf{x}_v^i\}_{i=1}^R$, where $\mathbf{x}_v^i \in d^{512}$ and $R \geq W$ is the maximum number of frames in the videos. Any missing frame features will be padded with zeros for videos with shorter durations.

Context Modeling. Due to the difference in dimensions of outputs from the CLIP and AST models, we compress the input features from both modalities to a unified dimension of $d = 256$ using a linear projection layer, which not only reduces the number of model parameters but also performs dimensionality reduction for features. We also select $d = 256$ as the output dimension for all subsequent encoding modules. We employ the Transformer architecture [34] for our music and video encoders. Transformers are crucial for improving model performance by encoding context sequences and modeling temporal relationships from video and music. Unlike MVPt, we use a *masked* single-block Transformer followed by mean pooling to aggregate frame or music segment features instead of relying on the [CLS] token. An important point to note is that the Transformer

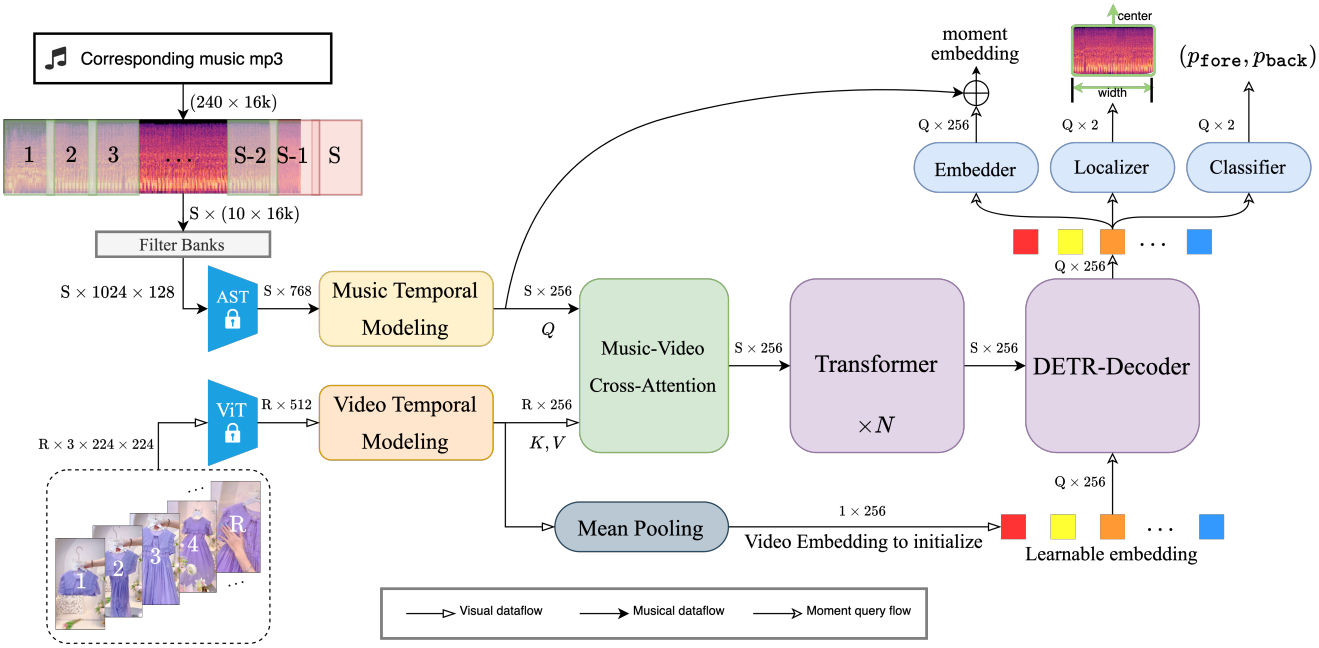


Figure 3. **Illustration of the proposed video-music moment localization model *Music–DETR***, which is composed of music/video temporal modeling, cross-modal fusion encoder, and DETR-based decoder. The decoder, following the DETR [2], performs the moment localization task. We use video embeddings to initialize the moment queries, enabling the prediction of the span range, moment classification, and moment embedding. Additionally, we optimize the alignment between the video and the moment embeddings with audio auxiliary to further constrain the training process and improve performance.

encoder shares parameters between the video and music sides and applies masking operations within each modeling module. The specific architecture is depicted in the supplementary material.

More formally, the key data flow of the visual and musical sides is expressed as follows:

$$\begin{cases} \{m^1, \dots, m^S\} & \leftarrow \text{music-to-segments}(m), \\ \{\mathbf{x}_m^1, \dots, \mathbf{x}_m^S\} & \leftarrow \text{AST}(\{m^1, \dots, m^S\}), \\ \{\hat{\mathbf{x}}_m^1, \dots, \hat{\mathbf{x}}_m^S\} & \leftarrow \text{Linear}(\{\mathbf{x}_m^1, \dots, \mathbf{x}_m^S\}), \\ \{\mathbf{y}_m^1, \dots, \mathbf{y}_m^S\} & \leftarrow \text{Transformer} \times 1(\{\hat{\mathbf{x}}_m^1, \dots, \hat{\mathbf{x}}_m^S\}), \\ \hat{\mathbf{y}}_m & \leftarrow \texttt{mean-pooling}(\{\mathbf{y}_m^1, \dots, \mathbf{y}_m^S\}). \end{cases} \tag{1}$$

$$\begin{cases} \{v^1, \dots, v^R\} & \leftarrow \text{video-to-frames}(v), \\ \{\mathbf{x}_v^1, \dots, \mathbf{x}_v^R\} & \leftarrow \text{ViT}(\{v^1, \dots, v^R\}), \\ \{\hat{\mathbf{x}}_v^1, \dots, \hat{\mathbf{x}}_v^R\} & \leftarrow \text{Linear}(\{\mathbf{x}_v^1, \dots, \mathbf{x}_v^R\}), \\ \{\mathbf{y}_v^1, \dots, \mathbf{y}_v^T\} & \leftarrow \text{Transformer} \times 1(\{\hat{\mathbf{x}}_v^1, \dots, \hat{\mathbf{x}}_v^R\}), \\ \hat{\mathbf{y}}_v & \leftarrow \texttt{mean-pooling}(\{\mathbf{y}_v^1, \dots, \mathbf{y}_v^R\}). \end{cases} \tag{2}$$

Cross-Modal Contrastive Learning. Given a batch of B (music, video) pairs, the model needs to generate and optimize $B \times B$ similarities. We apply InfoNCE [26], a symmetric cross-entropy loss, over these similarity scores to optimize the model's parameters,

$$\mathcal{L}_{m2v} = -\frac{1}{B}\sum_{k}^{B} \log \frac{\exp(s(\hat{\mathbf{y}}_{m_k}, \hat{\mathbf{y}}_v^+)/\tau)}{\sum_{j=1}^{B} \exp(s(\hat{\mathbf{y}}_{m_k}, \hat{\mathbf{y}}_{v_j})/\tau)}, \tag{3}$$

$$\mathcal{L}_{v2m} = -\frac{1}{B}\sum_{k}^{B} \log \frac{\exp(s(\hat{\mathbf{y}}_m^+, \hat{\mathbf{y}}_{v_k})/\tau)}{\sum_{j=1}^{B} \exp(s(\hat{\mathbf{y}}_{m_j}, \hat{\mathbf{y}}_{v_k})/\tau)}, \tag{4}$$

$$\mathcal{L} = \mathcal{L}_{m2v} + \mathcal{L}_{v2m}. \tag{5}$$

where $s(\hat{\mathbf{y}}_m, \hat{\mathbf{y}}_v)$ is the cosine similarity function, and τ is a learnable temperature parameter.

The loss $\mathcal{L}$ is the sum of video-to-text loss $\mathcal{L}_{m2v}$ and text-to-video loss $\mathcal{L}_{v2m}$. Notably, because multiple videos may share the same complete music, the positive music sample $\hat{\mathbf{y}}_m^+$ in $\mathcal{L}_{v2m}$ refers to all music instances corresponding to v_k in the batch instead of just $\hat{\mathbf{y}}_{m_k}$.

4.2. Stage II: Music Moment Localization

After the retrieval stage, we obtain k candidate musics that are most similar to the given video. The localization stage aims to identify the most relevant moments from each music.

Simply put, we can consider two straightforward approaches to tackle this task. We first use a sliding window to generate multiple window proposals, and then per-

form similarity retrieval by the model from the first stage. However, this approach suffers from significant efficiency issues, and the retrieval model struggles to distinguish different segments within the same music. Then, we guess that the detected highlights of the music, without considering the video content, might be sufficient for the VMMR task. Nevertheless, experiments in Tab. 4 revealed that not all ground-truth moments used the highlight music segments.

Therefore, as shown in Fig. 3, we designed Music–DETR, drawing inspiration from the DETR [2] architecture in visual object detection, which directly predicts a moment from the music for a given video in an end-to-end manner.

Cross-Modal Fusion. We use the same inputs and context modeling modules as the retrieval model, except for mean pooling, to separately obtain music segment representations $\mathbf{y}_m = \{\mathbf{y}_m^i\}_{i=1}^S$ and frame representations $\mathbf{y}_v = \{\mathbf{y}_v^i\}_{i=1}^R$. To extract video-related semantic information from the music features, we use a cross-attention Transformer [20] module to construct joint representation $\mathbf{z}_c$ for subsequent moment localization. Specifically, the keys and values from the video modality are fed into the music modality's multi-headed attention block. To ensure that the output cross-modal features are suitable for subsequent music localization, *i.e.* video-guided music features, we select the music modality as the Query.

$$\mathbf{z}_c \leftarrow \text{Cross-Attention}(\text{Q: } \mathbf{y}_m, \text{K: } \mathbf{y}_v, \text{V: } \mathbf{y}_v) \quad (6)$$

After cross-modal feature fusion, we further encode the features to $\hat{\mathbf{z}}_c$ using a transformer.

DETR-Decoder. Here, we follow the DETR method [2] in object detection to perform end-to-end temporal moment localization. The moment queries $\mathbf{q}_m$ in the decoder are composed of decoder embeddings and randomly initialized learnable embeddings. In our work, the choice of decoder embeddings in moment queries is particularly important. Compared to zero initialization [14, 23] in video moment retrieval, we believe that in multimodal tasks, the video query provides clear guidance for locating the target. Specifically, by mean-pooling the frame embeddings $\mathbf{y}_v$ to video embedding $\hat{\mathbf{y}}_v$, we then repeat it Q times to serve as the moment queries for Q predictions. This acts as an effective prior, helping the model to better focus on relevant temporal regions during the localization process. This decoder, as a cross-attention transformer, is formulated as follows,

$$\mathbf{E}_{\text{dec}} \leftarrow \text{Cross-Attention}(\text{Q: } \mathbf{q}_m, \text{K: } \hat{\mathbf{z}}_c, \text{V: } \hat{\mathbf{z}}_c) \quad (7)$$

Prediction Heads. Based on the decoder output $\mathbf{E}_{\text{dec}}$, we apply a 3-layer feed-forward network with ReLU [6] activation f_{loc} to predict the normalized moment center coordinate and width. During training, the best matching moment is selected from the Q ones with the Hungarian algorithm. We also follow the approach from DETR, where a linear layer with softmax f_{cls} is used to predict class labels. In DETR, this layer is trained by object class labels. In our context, we assign a foreground label to a predicted moment if it corresponds to ground truth, and a background otherwise. To further constrain the relationship between the decoder output and video, we use a linear layer f_{embed} to obtain a moment embedding aligned with the video representations before the cross-attention. Additionally, to facilitate the learning of the moment, we add the music embedding as a shortcut connection to it.

Loss. Our moment detection loss $\mathcal{L}_{\text{det}}$ measures the discrepancy between the ground-truth and predicted moments. Following DETR, $\mathcal{L}_{\text{det}}$ is computed as a weighted combination of an L_1 loss and a generalized IoU loss [30]. A cross-entropy loss $\mathcal{L}_{\text{cls}}$ is used to help discriminate moment (foreground) and non-moment (background) segments in a given music. Additionally, to improve the alignment between video and music moment, we use contrastive learning to compute the alignment loss $\mathcal{L}_{\text{align}}$ between the video and moment embeddings. The final loss summarizes the three losses.

5. Experiments

5.1. Experimental Setup

Datasets. Our experiments are mainly conducted on *Ad–Moment*. The dataset is divided at random into three disjoint subsets: 49k videos for training, 2k for validation, and 2k for testing, see Tab. 1.

ReaL also works for VMR. So in addition to Ad-Moment, we adopt the public HIMV dataset [10] commonly used for VMR. While HIMV have links to 200k videos, we downloaded with success only 138,265 videos as many links expired. We term this subset HIMV-138k. It is worth pointing out as no data split is publicly available, performance on HIMV is not directly comparable across papers. For a relatively fair comparison, we randomly select 2k videos for validation and another set of 2k videos for testing, with the retained 134,265 videos for training.

Evaluation criteria. For VMR, we use Recall@k, the fraction of queries that correctly retrieve the targeted music in the top k results (k=1, 5, 10, 25). For MML, we use mean Intersection over Union (IoU), given that the relevant music is already available. For VMMR, we use Recall@K, IoU=m [11], the percentage of test samples that have at least one predicted moment with IoU$>$ m in the top-k predictions (k=1,10,100 and m=0.5).

Implementation details. We perform experiments on 8 V100 GPUs. The mini-batch size is 512. An initial learning rate of 1e-4 is used to train the retrieval stage and 3e-4 for the localization stage. We train the models for 100 epochs in retrieval and 80 epochs in localization with the AdamW optimizer [19]. Following CLIP, we employ a cosine schedule

Method	Dataset	#Testset	R1	R5	R10	R25
VM-NET [10]	HIMV-200k	1k	8.2	–	23.3	35.7
MVPt [33]	YT8M-MV	2k	6.1	24.9	41.9	–
UT-CMVMR [21]	HIMV-200k	2k	10.8	28.1	36.5	51.6
ReaL	HIMV-138K	2k	11.0	26.2	35.9	51.4

Table 2. **A high-level comparison on VMR**. Note that as the dataset (and its data split) varies per method, the comparison is not head-to-head.

Input stride	R1	R5	R10	R25
2.5	2.15	5.60	**8.0**	12.3
5.0	**2.30**	**5.85**	7.8	**12.8**
7.5	2.20	5.05	7.0	11.9
10.0	2.05	4.40	6.8	9.9

Table 3. **Evaluating the influence of the input stride on ReaL for VMR**. Dataset: Ad-Moment.

[18] with a warm-up proportion of 0.05. We sample frames from all videos at 1 fps and resize all frames to 224×224 pixels.

5.2. Comparison with Other Methods

Results on VMR. Tab. 2 shows the VMR results on HIVM. Since previous methods for VMR are not open-sourced, we directly cite their published numbers. YT8M-MV, HIMV-200k and HIMV-138K are all subsets of YouTube8M under the "music video" category. Although this comparison is not entirely fair due to differences in datasets, it is evident that our simple and decoupled method achieves comparable scores even with smaller-size training data. Given that UT-CMVMR uses much larger training data and more features (optical flow and rhythm), our method is mostly comparable.

In Ad–Moment, we also evaluated the performance of the retrieval model in Tab. 3. The ablation study on the input stride size will be discussed later.

Results on MML. We used a sliding window approach with a 1-second stride and video duration as the window length. Multiple windows are selected from the full music track and ranked based on similarity by the retrieval method, ultimately choosing the highest similarity window as the predicted moment. Additionally, when selecting moments from complete music, we employed a highlight detection model [8] to find a highlight fully based on the duration of the video. For music where no highlight can be detected, we simulate the choice of a music editor by simply choosing the segment from the beginning. This also explains why many segments start from the beginning, see Fig. 2(d). Our method outperforms the baselines, see Tab. 4.

Results on VMMR. To ensure a fairer comparison, the results for the two baseline methods in the complete VMMR task are calculated based on the best retrieval performance (top line in Tab. 3). Similar to the Music Moment Localization task, ReaL method also shows higher metrics over the straightforward baseline, obtaining 105% (0.90 → 1.85) boost at R@100, IoU=0.5 in the complete VMMR task.

Method	IoU
Baseline:	
Sliding Window	0.216 (-64.1%↓)
Highlight Detection	0.353 (-41.4%↓)
This paper:	
ReaL	**0.602**
- w/o Music Shortcut Connection	0.556 (-7.6% ↓)
- w/o Moment Query Video Init	0.549 (-8.8% ↓)
- w/o $\mathcal{L}_{\text{align}}$	0.584 (-3.0% ↓)
#stride 5.0 → 2.5	0.595 (-1.2% ↓)
#stride 5.0 → 7.5	0.600 (-0.3% ↓)
#stride 5.0 → 10.0	0.586 (-2.7% ↓)

Table 4. **Music Moment Localization (MML) results**.

Method	R1	R10	R100
Baseline:			
Sliding Window	0.45	2.05	5.80
Highlight Detection	0.90	3.10	11.05
This paper:			
ReaL	**1.85**	5.70	**18.55**
- w/o Music Shortcut Connection	1.65	5.00	16.80
- w/o Moment Query Video Init	1.70	5.50	18.00
- w/o $\mathcal{L}_{\text{align}}$	1.75	5.65	17.90
#stride 5.0 → 2.5	**1.85**	**5.90**	18.00
#stride 5.0 → 7.5	1.80	5.55	17.90
#stride 5.0 → 10.0	1.70	4.85	16.50

Table 5. **VMMR results**.

5.3. Abalation Study

We provide the ablation study for the framework in terms of module details, learning objectives, and hyperparameters.

Music embedding as auxiliary. With the music embedding shortcut removed, we observe a 7.6% decrease at Mean IoU in the music moment localization subtask. This shows the necessity of music as an auxiliary component.

Video embedding to initialize. We further evaluate the impact of video embedding as initialization of decoder moment queries. By comparison, using video as the init significantly enhances the model's performance. This is because, in the case of moment retrieval as a multimodal task, the video modality provides clear guidance for locating the target, unlike in object detection. Therefore, relying solely on the random init of learnable embeddings in Fig. 3 is entirely insufficient.

Is $\mathcal{L}_{\text{align}}$ necessary? Removing the $\mathcal{L}_{\text{align}}$ results in about 3% drop at IoU in Tab. 4. The results from the localization subtask and the overall VMMR task indicate that the moment alignment is essential.

Hyperparameters. Tabs. 3 to 5 discuss the effect of the stride size when using a sliding window from the mu-

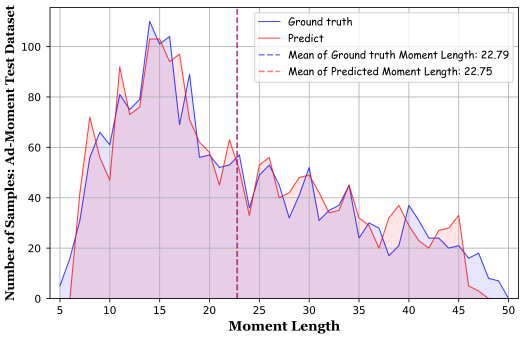


Figure 4. **Comparison of Ground truth and Predicted Moment Length.** By average, the duration difference between them is negligible.

sic track input. Performance under different sizes are presented. The VMMR performance is based on the same stride size, ensuring a consistent comparison across varying configurations. A smaller stride represents a higher sampling rate, suggesting that more dense sampling provides clear advantages for the localization task. Through temporal learning, this method demonstrates robustness to different configurations of the input music.

Further analysis for localization task. Considering real-world end-to-end applications, the duration for the predicted moment should closely align with the given video length. To this end, we further analyze the moment length by observing the predicted and ground truth moment. We visualized the number of samples grouped by the duration difference of 1-second intervals between the predicted and the ground truth moments in Fig. 4. Since the input video is constrained to the maximum duration by padding with zeros in the input set, the model indirectly learns the duration information of video during masked context modeling. By average, the mean lines of the predicted and ground truth moments are nearly identical (22.75s $\rightarrow$ 22.79s), ensuring the potential effectiveness of our approach in real-world applications.

6. Conclusions

To solve the problem that music tracks are often longer than short videos when adding background music to videos, we propose a novel multi-modal task named Video-to-Music Moment Retrieval (VMMR). It aims at retrieving music and finding the most appropriate moment from a large complete music gallery under the guidance of given video queries. To satisfy the practical demands, we collect a new dataset, named Ad–Moment, using a weakly supervised timestamp collection pipeline. We further propose a two-stage method ReaL as a strong baseline for the new task.

References

[1] Eric Brochu, Nando de Freitas, and Kejie Bao. The sound of an album cover: A probabilistic approach to multimedia. In *International Conference on Artificial Intelligence and Statistics*, 2003. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 5, 6

[3] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10551–10558, 2020. 2

[4] Yuki Era, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Video-music retrieval with fine-grained cross-modal alignment. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2005–2009. IEEE, 2023. 2

[5] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763v1*, 2019. 2

[6] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 6

[7] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 4

[8] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1783–1794, 2006. 7

[9] Xin Gu, Yinghua Shen, and Chaohui Lv. A dual-path cross-modal network for video-music retrieval. *Sensors*, 23(2): 805, 2023. 2

[10] Sungeun Hong, Woobin Im, and Hyun S Yang. Cbvmr: content-based video-music retrieval using soft intra-modal structure constraint. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pages 353–361, 2018. 1, 2, 6, 7

[11] Zhijian Hou, Chong-Wah Ngo, and Wing Kwong Chan. Conquer: Contextual query-aware ranking for video corpus moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3900–3908, 2021. 2, 6

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 2

[13] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference,*

Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, pages 447–463. Springer, 2020. 2

[14] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2, 6

[15] Bochen Li and Aparna Kumar. Query by video: Cross-modal music retrieval. In *ISMIR*, pages 604–611, 2019. 2

[16] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[17] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 2

[18] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 6

[21] Tianjun Mao, Shansong Liu, Yunxuan Zhang, Dian Li, and Ying Shan. 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)*, pages 7890–7894. IEEE, 2024. 1, 2, 7

[22] Daniel McKee, Justin Salamon, Josef Sivic, and Bryan Russell. Language-guided music recommendation for video via prompt analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14784–14793, 2023. 2

[23] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 2, 6

[24] Megha Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *arXiv preprint arXiv:2101.08540*, 2021. 2

[25] Ke Ning, Lingxi Xie, Jianzhuang Liu, Fei Wu, and Qi Tian. Interaction-integrated network for natural language moment localization. *IEEE Transactions on Image Processing*, 30: 2538–2548, 2021. 2

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[27] Laure Prétet, Gael Richard, and Geoffroy Peeters. Cross-modal music-video recommendation: A study of design choices. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021. 2

[28] Laure Prétet, Gaël Richard, Clément Souchier, and Geoffroy Peeters. Video-to-music recommendation using temporal alignment of segments. *IEEE Transactions on Multimedia*, 25:2898–2911, 2022. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[31] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023. 4

[32] Ki-Ho Shin and In-Kwon Lee. Music synchronization with video using emotion similarity. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 47–50, 2017. 2

[33] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10564–10574, 2022. 1, 2, 7

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[35] Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. Cross-modal variational auto-encoder for content-based micro-video background music recommendation. *IEEE Transactions on Multimedia*, 25:515–528, 2021. 2

[36] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9159–9166, 2019. 2

[37] Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv preprint arXiv:2011.09046*, 2020. 2

[38] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 2

[39] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 2

[40] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering

framework. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4252–4266, 2021. 2

[41] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020. 2

A. Appendix

In this material, we provide additional figures and explanations omitted from the main paper due to space constraints.

Architecture of Retrieval Model. The overall architecture of the retrieval model is depicted in Fig. 5, which is described in the Method Section. Given a video, we can obtain the $\mathbf{top}k$ candidate music by computing similarity and sorting. The video/music temporal modeling consists of a linear layer for compressed dimensions and a single block Transformer mentioned in the main paper.

Automatic Dataset Construction. To give a more detailed explanation of how the annotated data is processed, we show a diagram in Fig. 6. After performing automated data cleaning to remove low-quality content, we conduct moment localization on the filtered raw advertising short videos. Specifically, given a short video and its corresponding raw background music track, we employ *Background Sound Extraction* and *Moment Localization* mentioned in the main paper to identify the correct start and end time points in a weakly supervised manner. This leads to Ad–Moment, a timestamp-annotated set of 53k video-music moment pairs for MML task.

More Visualization Results. To better demonstrate the practicability of our method in real-world application scenarios of short advertisement videos, we provide some annotated samples and visualization results of the Ad–Moment dataset in Fig. 7. Through the video title and tags, one can gain a clearer understanding of the content and themes expressed in the video. Additionally, the music title and description can provide insights into the type of music corresponding to the video. We presented three types of moments, where both the predicted moment and the highlight moment are selected based on the corresponding music in the music moment localization (MML) task. It can be observed that the predicted moment closely aligns with the ground truth, where the duration of the ground truth one matches the video length exactly. To indicate the difference between music and moments, we add corresponding audio descriptions.

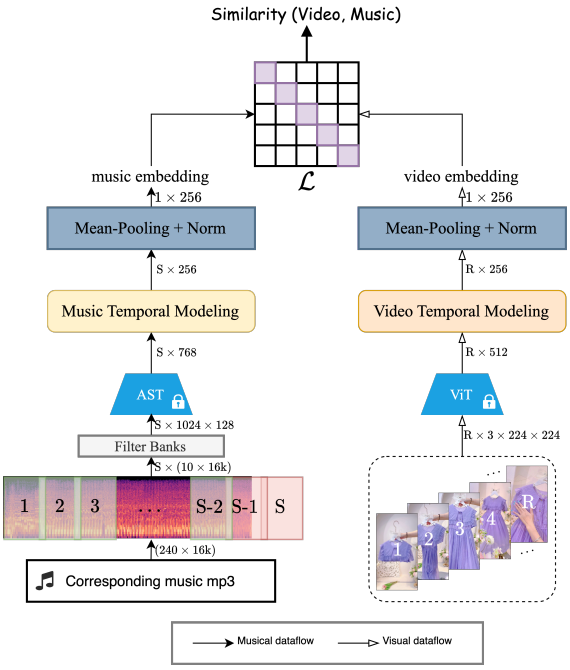


Figure 5. **Illustration of the retrieval model in stage I.**

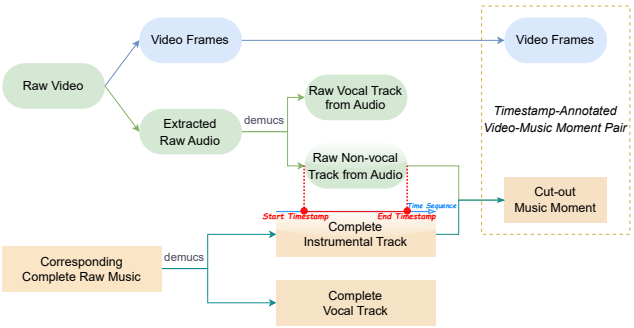
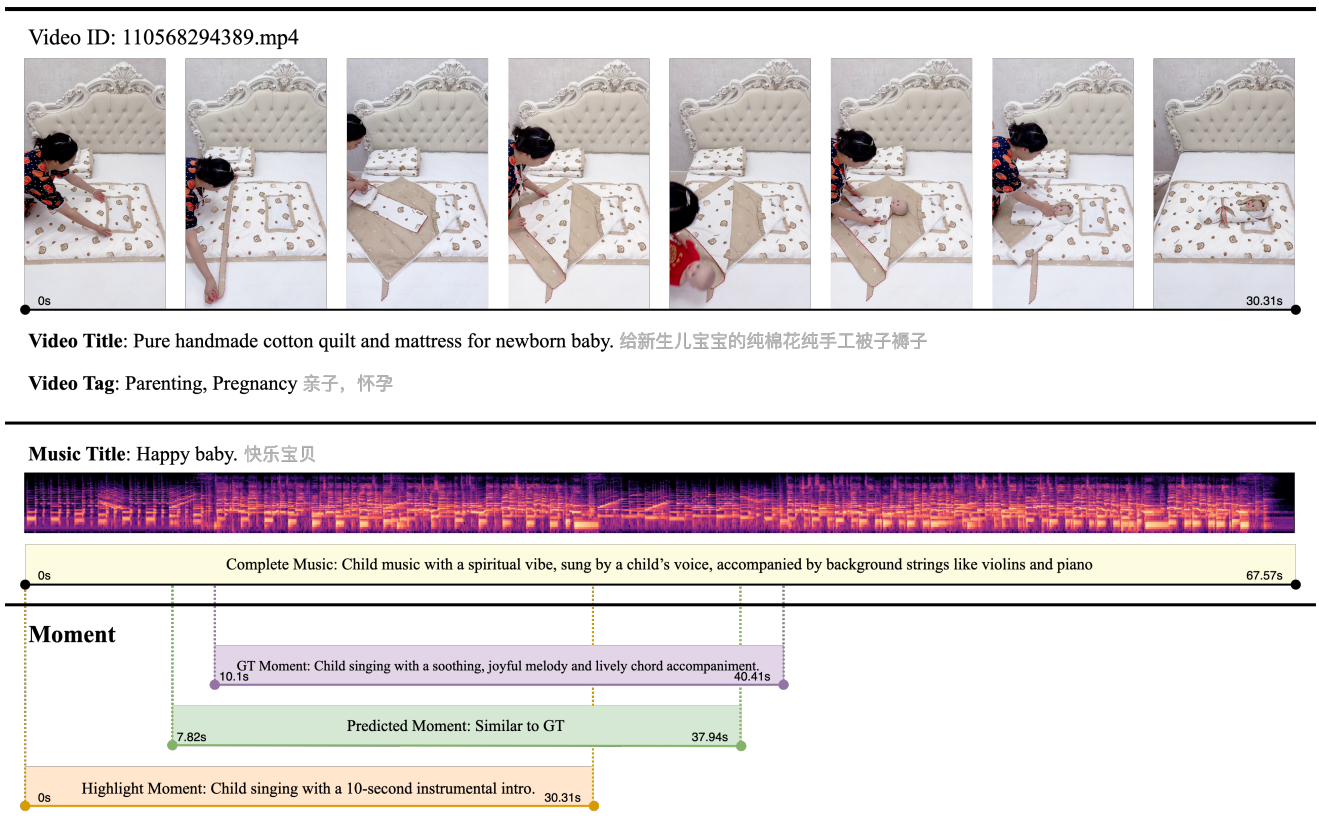
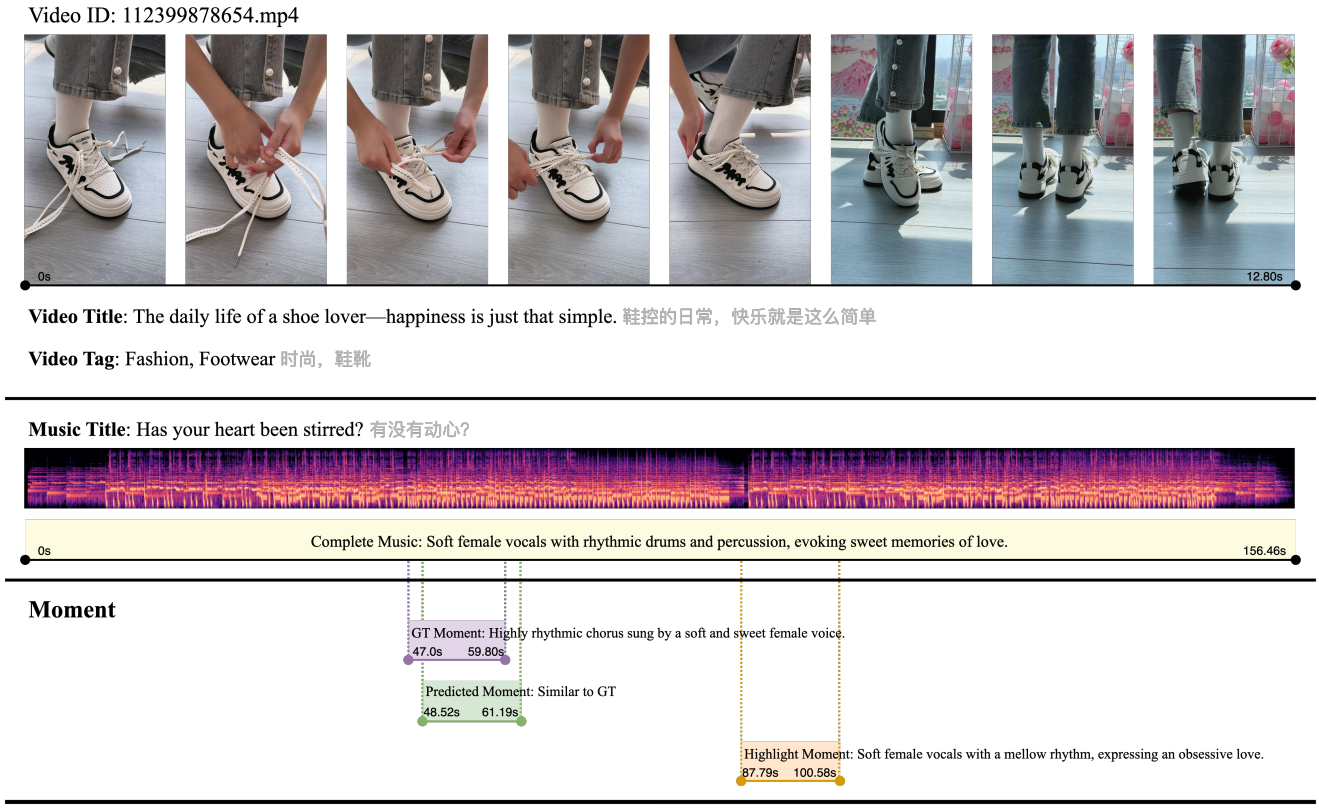


Figure 6. **Conceptual diagram of the weakly supervised multi-modal timestamp collection pipeline for the proposed Ad–Moment dataset.**



((a))



((b))

Figure 7. **Data and MML localization results visualization on Ad–Moment dataset.** Note that the authors provide the descriptions of the music tracks and moments, which the model does not use. They are only intended to help visualize the music.