Music Grounding by Short Video

Zijie Xin[1] Minquan Wang[2] Jingyu Liu[1] Quan Chen[2] Ye Ma[2] Peng Jiang[2]
Xirong Li[1]

[1] Renmin University of China [2] Kuaishou Technology

https://github.com/xxayt/MGSV

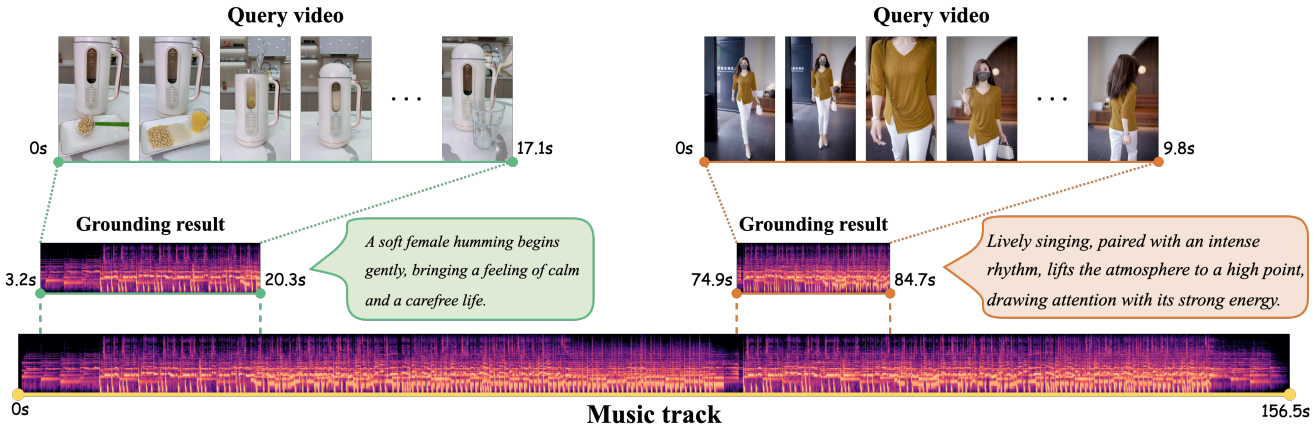


Figure 1. **Music grounding by short video** (MGSV), aiming to localize within a music-track collection a *music moment* that best serves as background music for the query video. Text in callout boxes to the right of each music moment is for illustrative purposes only.

Abstract

*Adding proper background music helps complete a short video to be shared. Previous work tackles the task by video-to-music retrieval (V2MR), aiming to find the most suitable music track from a collection to match the content of a given query video. In practice, however, music tracks are typically much longer than the query video, necessitating (*manual*) trimming of the retrieved music to a shorter segment that matches the video duration. In order to bridge the gap between the practical need for* music moment localization *and V2MR, we propose a new task termed Music Grounding by Short Video (MGSV). To tackle the new task, we introduce a new benchmark, MGSV-EC, which comprises a diverse set of 53k short videos associated with 35k different music moments from 4k unique music tracks. Furthermore, we develop a new baseline method, MaDe, which performs both video-to-music* ***ma****tching and music moment* ***de****tection within a unified end-to-end deep network. Extensive experiments on MGSV-EC not only highlight the challenging nature of MGSV but also set MaDe as a strong baseline.*

1. Introduction

Music can name the unnameable and communicate the unknowable[1]. Adding appropriate background music (BGM) helps complete short videos, a dominant form of information dissemination online. To that end, current research focuses on video-to-music retrieval (V2MR), finding amidst a collection of music tracks the one best matching the content of a given query video [4, 17, 21, 27]. In practice, music tracks (songs or instrumental music) are typically much longer than query videos (which usually last around 30 seconds). As a result, the retrieved music needs to be *manually* cut to a shorter segment to match the video duration. In order to bridge the gap between the practical need and V2MR, we introduce in this paper a novel task called **M**usic **G**rounding by **S**hort **V**ideo (MGSV), aiming to localize within the music-track collection a *music moment* that

Work performed as an intern at Kuaishou. (xinzijie@ruc.edu.cn)
Corresponding author. (xirong@ruc.edu.cn)

[1] Leonard Bernstein

best serves as BGM for the query video, see Fig. 1.

To address the new task, one might consider applying V2MR at a fine-grained segment level, with music tracks pre-trimmed into shorter segments. This approach is problematic due to the varying durations of videos. While MGSV is new, we observe in a broader context its conceptual resemblance to Video Grounding [14, 16, 30], which aims to localize a moment in a specific video w.r.t. a given textual query. In particular, the query and target modalities in MGSV (Video Grounding) are video (text) and music (video), respectively. Thus, a Video Grounding method could, in principle, be repurposed for MGSV in a *single-music* mode, assuming that the relevant music track for a query video is provided. When dealing with a collection of music tracks, V2MR is required to identify the relevant track. While combining existing Video Grounding and V2MR methods provides a good starting point, these methods were not originally designed for MGSV. By incorporating task-specific optimizations, we develop a stronger baseline that performs video-to-music **Ma**tching and music moment **De**tection (**MaDe**) in a single network.

While few public datasets (mainly HIMV [11]) exist for V2MR, each video in these datasets is paired with an equal-length music segment, yet the source tracks from which these segments were derived are absent. Therefore, they cannot be repurposed to MGSV. We construct a new dataset, **MGSV-EC**, sourced from an **e-c**ommerce video creation platform. The availability of user-generated BGM editing logs on this platform allows us to reliably trace back to the source tracks. Noting that music selection for short videos involves inherent subjectivity, our work explores if an automated system can recommend music comparable to that made by skilled practitioners. As such, this work is a starting point for answering the higher-level challenge of diversity. In sum, our main contributions are as follows:

- **Dataset**. We introduce MGSV-EC, a real-world dataset comprising 53k videos associated with 35k different music moments from 4k unique music tracks, see Fig. 2. This dataset supports the evaluation of methods for MGSV in two distinct modes: *single-music* grounding (SmG) and *music-set* grounding (MsG).
- **Baseline**. We develop MaDe, a unified end-to-end deep network that performs both video-to-music matching and music moment detection. MaDe will be open-source.
- **Evaluation**. We adapt and evaluate a diverse set of models, including six Video Grounding models (Moment-DETR [16], QD-DETR [23], TR-DETR [26], EaTR [14], and UVCOM [30]), one V2MR model (MVPt [27]), and one Video-corpus Moment Retrieval model (CONQUER [12]). The evaluation highlights the challenging nature of MGSV and establishes MaDe as a stronger baseline.

The rest of the paper is organized as follows. We discuss related work in Sec. 2. The new dataset is introduced in Sec. 3, followed by our solution in Sec. 4. Experiments are presented in Sec. 5. We conclude the paper in Sec. 6.

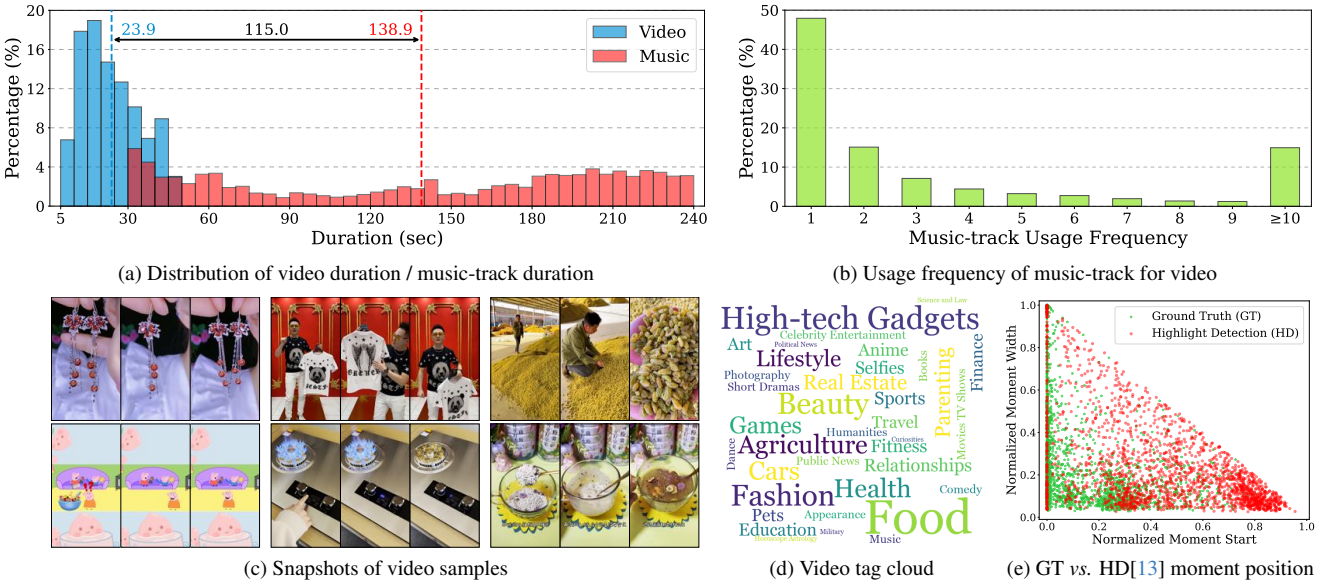
2. Related Work

In developing our solution for MGSV, we draw inspiration from recent studies on Transformers-based Video Grounding and Video-to-Music Retrieval (V2MR).

Video Grounding, *aka* video moment retrieval by natural language, involves localizing a moment within an untrimmed video that corresponds to a given natural language text. While earlier methods for this task typically rely on pre-generated segments as candidate moments or moment proposals [22, 32, 34], more recent models, equipped with DETR-like decoding modules, directly regress temporal boundaries in a proposal-free manner [14, 16, 23, 26, 30]. Key elements of these models are summarized in Tab. 1. For instance, consider Moment-DETR [16], the first work that adapted DETR [1] for the task. Given a sequence of visual tokens extracted from video frames and textual tokens from the input text, Moment-DETR first enhances the unimodal features and aligns their dimensions using fully connected (FC) layers. The enhanced tokens are then concatenated (ⓒ) along their temporal axis and passed through two self-attention (SA) blocks for multimodal feature fusion. Subsequently, a cross-attention (CA) based decoder is employed, utilizing an array of 10 trainable query tokens as Q, with the previously fused tokens serving as K and V for moment location prediction. Note that each query token for a specific CA is obtained as the sum of a *positional* token, which is shared across all CAs, and a *content* token from the preceding CA. For its first CA, Moment-DETR initializes the content token with a ZERO vector. As shown in Tab. 1, subsequent works such as QD-DETR [23], TR-DETR [26], and UVCOM [30] primarily focus on innovations in mul-

Model	**Unimodal Feat. Enhancement**	**Multimodal Feat. Fusion**	**Init. Content Token ϕ_0**	**Decoder**
Moment-DETR[16]	$FC_{\times 2}$	$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
QD-DETR[23]	$FC_{\times 2}$	$CA_{\times 2}$+$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
QD-DETR+	$FC_{\times 2}$	$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
TR-DETR[26]	$FC_{\times 2}$	VFR+$CA_{\times 2}$+$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
TR-DETR+	$FC_{\times 2}$	VFR+$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
EaTR[14]	$FC_{\times 2}$	$SA_{\times 3}$	Target-modality features	GF + $SA\text{-}CA_{\times 2}$
UVCOM[30]	$FC_{\times 2}$	Dual-$CA_{\times 3}$ + DBIA+LRP+$SA_{\times 3}$	Target-modality features	$SA\text{-}CA_{\times 3}$
UVCOM+	$FC_{\times 2}$	DBIA+LRP+$SA_{\times 3}$	Target-modality features	$SA\text{-}CA_{\times 3}$
MaDe (this paper)	FC+SA	$SA_{\times 2}$	Query-modality feature	$CA_{\times 6}$

Table 1. **Key elements in current DETR-based models for video grounding**. When adapting these models to the new MGSV task, we find that removing cross-attention (CA) blocks from their multimodal feature fusion module (where applicable) results in stronger baselines, which we denote with a "+" suffix.



(a) Distribution of video duration / music-track duration

(b) Usage frequency of music-track for video

(c) Snapshots of video samples

(d) Video tag cloud

(e) GT *vs.* HD[13] moment position

Figure 2. **Visualization of MGSV-EC**. Note that the video tag cloud is merely to demonstrate the richness of the visual content.

timodal feature fusion to enhance the interaction between visual and textual tokens, largely through the use of cross-attention (CA) mechanisms. However, in our preliminary experiments, we find that models employing CA-based fusions are more prone to overfitting in the new task. Consequently, we remove CA from the fusion module.

There have been notable efforts to extend video grounding to a video-corpus scenario, known as Video Corpus Moment Retrieval (VCMR) [12, 31, 33]. In particular, these methods first perform text-to-video retrieval to select the top-k candidate videos from a given corpus. Subsequently, moment detection is conducted for each candidate video. It is worth noting that the above works implement moment detection in a proposal-based manner. For instance, SQuiDNet [31] treats each sampled frame as a candidate point for the start or end timestamp of the moment to be detected, thereby formulating the task as a binary classification problem. Consequently, the temporal resolution of the detection result is limited by the frame sampling frequency.

V2MR aims to retrieve music tracks relevant to a given query video through video-to-music matching. Earlier approaches utilized two-branch MLPs for cross-modal feature alignment [2, 11, 24]. In contrast, current methods are based on Transformers, employing self-attention (SA) mechanisms for long-range temporal modeling, see MVPt [27]. Building on MVPt as their backbone, ViML [21] fuses text and video input representations to query music samples, while MuseChat [4] introduces an interactive V2MR system that enhances retrieval results through conversational interactions. Despite these advancements in video-music matching, music grounding is unexplored in these efforts.

3. Dataset for MGSV: MGSV-EC

In order to develop and evaluate solutions for the new MGSV task, we build MGSV-EC, a dataset consisting of 53k professionally made **E-c**ommerce videos, with their BGMs extracted from a set of 4k unique music tracks, see Tab. 2 and Fig. 2. We describe the data curation procedure in Sec. 3.1, followed by the evaluation protocol in Sec. 3.2.

3.1. Dataset Curation

Raw Data Gathering. With permission, we gathered videos posted on an E-commerce video creation platform in the period from Jan. to Sep. 2023, resulting in an initial set of 100k videos. The videos have a frame rate of 34 and a raw resolution of 1080×1920. Each video is associated with a BGM editing log, indicating which music track was used and which part of the music was adopted as the BGM. A set of 9k music tracks was used in total. As the raw video set is quite diverse with varied data quality, automated data cleaning is performed as follows.

Data Cleaning. Given the positive correlation between

Split	Music tracks	*Duration*(s)	Query videos	*Duration*(s)	Moments
Total	4,050	138.9±69.6	53,194	23.9±10.7	35,393
Train	3,496	138.3±69.4	49,194	24.0±10.7	31,660
Val.	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 2. **Overview of the MGSV-EC dataset**. While the query videos have no overlap between train / val. / test, music tracks are partially shared across the data split. As the music tracks are meant for re-using, such a setup is practical.

video content quality and view counts / follows, we empirically exclude videos that have no more than 200 view counts or 100 follows. Videos too short (<5 seconds) or too long (>50 seconds) are also removed. With the above cleaning procedure, we preserve 53,194 short videos accompanied with 35,393 different music moments from 4,050 unique music tracks. Among the tracks, songs and instrumental music are approximately in a 6:4 ratio. As Fig. 2a shows, the video duration is 23.9 seconds on average, whilst that of the music tracks is 138.9 seconds, indicating a noticeable duration gap between the query videos and the music tracks to be retrieved. Fig. 2e shows a large disparity between the ground-truth moment positions and those detected by a music highlight extractor[2] [13]. These results clearly show the challenging nature of the new MGSV task.

Data Split. For reproducible research, we recommend a data split as follows. A set of 2k videos is randomly sampled to form a held-out test set, while another random set of 2k videos is used as a validation set (for hyper-parameter tuning, model selection, *etc*.), with the remaining data used for training, see Tab. 2.

3.2. Evaluation Protocol

Our benchmark supports two evaluation modes, *i.e*. *single-music*, wherein the music track relevant w.r.t. to a given query video is known in advance, and *music-set*, wherein the relevant music track has to be retrieved from a given music track set. Mode-specific tasks and evaluation criteria are described as follows, see Sec. 3.2.

Mode	(Sub-)Tasks	Metrics
Single-music	Grounding (SmG)	mIoU
Music-set	Video-to-Music Retrieval (V2MR)	Rk
	Grounding (MsG)	MoRk

Table 3. **Evaluation modes, (sub-)tasks and metrics**.

***Single-music* Mode**. In this mode, a user already knows in their mind which music track to use. A model for MGSV shall automatically detect and cut from the track an appropriate moment based on the query video. In order to evaluate the accuracy of single-music grounding (**SmG**), per query video we compute temporal Intersection over Union (**IoU**) between the predicted moment and the corresponding ground truth. Higher IoU is better. A mean IoU (**mIoU**) is obtained by averaging IoU scores over all query videos.

***Music-set* Mode**. In contrast to the single-music mode, the music-set mode is much more challenging as which music track is relevant w.r.t. to the query video is unknown. Hence, the effectiveness of a model is jointly determined by its performance in two sub-tasks, *i.e*. video-to-music retrieval (**V2MR**) for finding the relevant music track and music-set grounding (**MsG**) to localize the relevant moment. For V2MR, we adopt the popular Recall@k (**R**k), *i.e*. the percentage of query videos that have their corresponding music tracks ranked in the top-k retrieval results (k=1, 5, 10). To evaluate MsG, we borrow IoU-conditioned Rk from [12], calculating Moment Recall (**MoR**k) as the percentage of query videos that have corresponding music moments (with IoU>0.7) recalled in the top-k detected moments (k=1,10,100). For a fair comparison across models, we recommend generating the top-k moments using a unified post-processing strategy: picking up one moment for each of the top-k ranked music tracks.

[2] https://github.com/remyhuang/pop-music-highlighter/

4. A Strong Baseline for MGSV: MaDe

Given a query video v and a specific music track M as multi-modal input, we aim for a music grounding network $\mathcal{G}$ that can simultaneously estimate the relevance of M w.r.t. v and detect an appropriate music moment as the video's BGM. More formally, letting p_s indicate the relevance score and p_c / p_w as the (normalized) temporal center / width of the moment, we express the network computation at a high level as:

$$(p_s, p_c, p_w) \leftarrow \mathcal{G}(v, M). \tag{1}$$

One might question the necessity of predicting p_w, as the moment length appears to be directly obtainable from the query video. We argue that enabling $\mathcal{G}$ to predict p_w enhances its awareness of the moment boundaries, thereby improving its overall grounding capability. Note that in the *single-music* mode where M is known to be relevant w.r.t. v, p_s will be ignored. As for the *music-set* mode, all the candidate music tracks will be ranked by their p_s and accordingly, the moment detected from the top-ranked music track will be chosen as the final prediction. Next, we elaborate the network architecture of $\mathcal{G}$ in Sec. 4.1, followed by the description of network training in Sec. 4.2.

4.1. The MaDe Network

The overall structure of our network is illustrated in Fig. 3. Conceptually, MaDe consists of three modules, *i.e*. 1) **multimodal embedding** (Sec. 4.1.1) that converts the given query video and music track to a sequence of visual and audio tokens, respectively, 2) **video-music matching** (Sec. 4.1.2) which estimates the relevance of the music track w.r.t. the video based on their holistic embeddings, and 3) **music moment detection** (Sec. 4.1.3) that localizes within the music track the moment best fitting the video.

4.1.1. Multimodal Embedding

Raw Feature Extraction. Following the current practice of audiovisual analysis [5, 19], we adopt pre-trained CLIP (ViT-B/32) [25] and Audio Spectrogram Transformer (AST) [7] as our (weights-frozen) video and audio encoders, respectively. In particular, the query video is de-

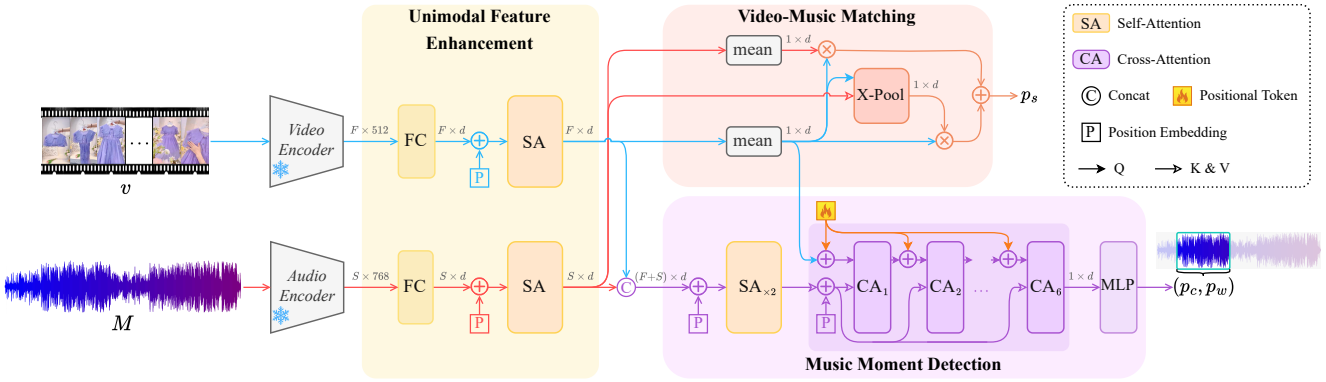


Figure 3. **Conceptual diagram of the proposed MaDe network for MGSV**. At a high level, given a query video v and a music track M as multi-modal input, MaDe yields a three-dimensional output (p_s, p_c, p_w), where p_s measures the relevance of M w.r.t. v, whilst p_c / p_w indicates the center / width of the detected music moment. The video is initially represented by a sequence of F evenly sampled frames, and the music track as a sequence of S partially overlapped Mel-spectrogram segments. Pre-trained CLIP (ViT-B/32) and Audio Spectrogram Transformer (AST) are used as our (weight-frozen) video and audio encoders, respectively. Each encoder is followed by an FC layer to compress frame / audio embeddings to a smaller size of d(=256) for cross-modal feature alignment and for parameter reduction. In the *single-music* mode where M is known to be relevant w.r.t. v, p_s will be ignored. In the *music-set* mode, we rank all candidate music tracks by their p_s and accordingly select the moment detected from the top-ranked music track as the grounding result.

coded at 1 FPS, yielding a sequence of F frames resized to 256×256. A 512-d CLS token is extracted per frame. As for the music track, we use AST's data preprocessing procedure to generate a sequence of S partially overlapped 128-bin Mel-spectrogram features. These Mel features are then fed into AST to obtain 768-d audio tokens $\{s_i\}_{i=1}^{S}$.

Unimodal Feature Enhancement. For cross-modal embedding alignment and for parameter reduction, we place a fully connected (FC) layer after each encoder to compress the visual / audio tokens to a smaller size of d (256). The compressed tokens are denoted as $\{\hat{f}_i\}_{i=1}^{F}$ and $\{\hat{s}_i\}_{i=1}^{S}$, respectively. Since the visual (audio) tokens are extracted independently per frame (segment), the inter-frame (inter-segment) connection along the temporal dimension is ignored. In order to enhance the tokens by temporal modeling, we feed them to a self-attention (SA) block [29]. Given $\{\tilde{f}_i\}$ and $\{\tilde{s}_i\}$ as temporally enhanced tokens, we express the multi-modal embedding module more formally as:

$$\begin{cases} \{f_i\}_{i=1}^{F} & \leftarrow \text{ViT}(v, F), \\ \{\hat{f}_i\}_{i=1}^{F} & \leftarrow \text{FC}_{512\times d}(\{f_i\}_{i=1}^{F}), \\ \{\tilde{f}_i\}_{i=1}^{F} & \leftarrow \text{SA}(\{\hat{f}_i\}_{i=1}^{F}), \\ \{s_i\}_{i=1}^{S} & \leftarrow \text{AST}(M, S), \\ \{\hat{s}_i\}_{i=1}^{S} & \leftarrow \text{FC}_{768\times d}(\{s_i\}_{i=1}^{S}), \\ \{\tilde{s}_i\}_{i=1}^{S} & \leftarrow \text{SA}(\{\hat{s}_i\}_{i=1}^{S}). \end{cases} \tag{2}$$

4.1.2. Video-Music Matching

For video-music matching, holistic (video-level / track-level) embeddings are required. To obtain such two embeddings, denoted as $h(v)$ and $h(M)$, a simple and commonly used operation is mean pooling over $\{\tilde{f}_i\}_{i=1}^{F}$ and $\{\tilde{s}_i\}_{i=1}^{S}$, respectively. However, as shown in Fig. 2a, the large disparity between short video duration and long music-track duration suggests that the music's relevance to the video is *partial*. Hence, the mean-pooled embedding, denoted as $h_0(M)$, with the S segment embeddings treated equally, is suboptimal to reflect such partial relevance.

We improve video-text matching by adopting X-Pool [8] to obtain a *video-conditioned* track embedding $h_1(M)$. Originally developed for video-text retrieval, X-Pool aggregates an array of local features with a cross-attention (CA) mechanism. In the current context, the video embedding $h(v)$, obtained by mean pooling, is used as Q, whilst $\{\tilde{s}_i\}_{i=1}^{S}$ are used as K and V. The attention weights generated by CA are used for weighted feature aggregation. The weights are computed based on the dot product between Q and K and thus reflect segment-wise relevance w.r.t. the video. Hence, music segments closer to the video will contribute more to $h_1(M)$. The final video-music similarity p_s is calculated as the sum of the cosine similarity (cs) between $h(v)$ and $h_0(M)$ and that of $h(v)$ and $h_1(M)$. In short, we have X-Pool enhanced video-music matching as:

$$\begin{cases} h(v) & \leftarrow \text{mean-pooling}(\{\tilde{f}_i\}_{i=1}^{F}), \\ h_0(M) & \leftarrow \text{mean-pooling}(\{\tilde{s}_i\}_{i=1}^{S}), \\ h_1(M) & \leftarrow \text{X-Pool}(h(v) \text{ as } Q, \{\tilde{s}_i\}_{i=1}^{S} \text{ as } K/V), \\ p_s & \leftarrow cs(h(v), h_0(M)) + cs(h(v), h_1(M)). \end{cases} \tag{3}$$

4.1.3. Music Moment Detection

In order to predict the moment localization (p_c and p_w) based on the visual and audio tokens ($\{\tilde{f}_i\}_{i=1}^{F}$ and $\{\tilde{s}_i\}_{i=1}^{S}$), we adapt Moment-DETR [16] as follows. Similar to Moment-DETR, we first fuse the tokens via concat (ⓒ)

and two SA blocks, producing $F + S$ modality-fused tokens $\{c_i\}_{i=1}^{F+S}$. In the subsequent decoding stage, $\{c_i\}$ are used as K and V input to each CA block, see Fig. 3. Note that in Moment-DETR, each CA processes an array of 10 composite query tokens, which are the sum of 10 learnable *positional* tokens P shared among the CAs and the same amount of *content tokens* generated by the previous CA. In other words, given ϕ_k as the output of the k-th CA block, we have $P + \phi_{k-1}$ as Q. Concerning the choice of ϕ_0, *i.e.* the content token for the first CA, we propose to use the video embedding $h(v)$ instead of $\mathbf{0}$ in Moment-DETR. Such a simple tactic is important for music moment detection. Moreover, as we target at grounding the best music moment for a given video, we reduce the number of query tokens from 10 to 1. As such, the SA preceding each CA in Moment-DETR and meant for processing the query token sequence is no longer needed. Hence, different from previous DETR-based models, our decoder has 6 CAs and 0 SA, see Tab. 1. Finally, we obtain p_c and p_w by feeding the output of the last CA ϕ_6 to an MLP[3]. The detection module is expressed more formally as:

$$\begin{cases} \{c_i\}_{i=1}^{F+S} & \leftarrow \text{SA}_{\times 2}(\{\tilde{f}_i\}_{i=1}^{F} \copyright \{\tilde{s}_j\}_{j=1}^{S}), \\ \phi_0 & \leftarrow h(v), \\ \phi_k & \leftarrow \text{CA}_k(P + \phi_{k-1} \text{ as } Q, \{c_i\}_{i=1}^{F+S} \text{ as } K/V), \\ (p_c, p_w) & \leftarrow \text{MLP}(\phi_6). \end{cases} \tag{4}$$

4.2. Network Training

Loss for Video-Music Matching. We adopt a symmetric (video-to-music and music-to-video) InfoNCE loss, commonly used for training cross-modal matching networks [3, 10, 25, 28]. Recall that our video-music similarity combines the mean-pooled embedding based similarity and the X-Pooled embedding based similarity, see Eq. (3). To compute such a multi-similarity based loss, we see two options: a single loss calculated using the combined similarity or a joint loss with its components calculated separately using the individual similarities. Previous work on video-text retrieval [18] shows that the latter is better. We follow their recommendation, calculating the InfoNCE loss for each similarity and using their equal combination as the matching loss.

Loss for Music Moment Detection. In order to measure the temporal misalignment between the predicted moment and the ground truth, we adopt the loss from Moment-DETR [16], which comprises an L1 regression loss and a generalized Intersection over Union (IoU) loss. Note that for stable and fast training, the ground-truth moment center and width are normalized by the maximum music duration.

The matching loss and the detection loss are equally combined. As shown in Fig. 3, except for the weights-frozen video and audio encoders, our network is end-to-end trained by minimizing the combined loss.

[3] $\text{FC}_{d\times d} \rightarrow \text{ReLU} \rightarrow \text{FC}_{d\times d} \rightarrow \text{ReLU} \rightarrow \text{FC}_{d\times 2} \rightarrow \text{sigmoid}$

Implementation Details. We use Transformers in their default setup [29]: sinusoidal positional encoding, 8 heads, and a dropout rate of 0.1. Trainable weights are initialized with Kaiming init [9], except for the detection module, which uses Xavier init [6] as Moment-DETR. Our optimizer is Adam [15] with an initial learning rate of 1e-4. Following CLIP, we employ a cosine schedule [20] with a warm-up proportion of 0.02. The network is trained for 100 epochs with mini-batch size of 512. Our experimental environment is PyTorch 1.13.1 and NVIDIA 3090 GPUs.

5. Evaluation

5.1. Single-music Grounding

Baselines. For a fair and reproducible comparison, we opt for SOTA video grounding models that are DETR-based and open-source. Accordingly, we collect the following five models and repurpose them for the single-music grounding (SmG) task: Moment-DETR[4] (NIPS21) [16], QD-DETR[5] (CVPR23) [23], EaTR[6] (ICCV23) [14], TR-DETR[7] (AAAI24) [26], and UVCOM[8] (CVPR24) [30]. These models are provided with the same raw features, the same training data and the same detection loss as MaDe.

As shown in Tab. 1, QD-DETR, TR-DETR and UVCOM employ CA blocks for multimodal feature fusion. We observe in our preliminary experiments that the CAs tend to cause these models to overfit to the training data. We term their CA-removed counterparts as QD-DETR+, TR-DETR+ and UVCOM+, respectively. In total, we have eight baselines for SmG. Note that they do not support video-music matching and are therefore inapplicable to MsG.

Results. The performance of the baselines is shown in Tab. 4. The superior performance of QD-DETR+ and TR-DETR+ compared to their original counterparts (mIoU 0.634 *versus* 0.423 and 0.630 *versus* 0.393) clearly reveals the negative impact of their CA-based fusion on the performance. By contrast, UVCOM (mIoU 0.652) is relatively stable and better. We attribute this result to UVCOM's dual CAs wherein the query (video) features and the target (music) features are each used as Q respectively, making the query information better preserved during feature fusion. The higher mIoU of UVCOM+ (0.661) indicates that removing the dual CAs is also beneficial. With a mIoU score of 0.722, our MaDe clearly outperforms all the baselines.

5.2. Music-set Grounding

Baselines. To tackle the MsG task, a straightforward solution is to combine UVCOM+ with an existing V2MR model. We implement MVPt (CVPR22) [27], a popular

[4] https://github.com/jayleicn/moment_detr
[5] https://github.com/wjun0830/QD-DETR
[6] https://github.com/jinhyunj/EaTR
[7] https://github.com/mingyao1120/TR-DETR
[8] https://github.com/EasonXiao-888/UVCOM

Model	#Params (M)	SmG	V2MR			MsG		
		mIoU	*R1*	*R5*	*R10*	*MoR1*	*MoR10*	*MoR100*
Video Grounding re-purposed:								
TR-DETR, AAAI'24 [26]	7.8	0.393	–	–	–	–	–	–
QD-DETR, CVPR'23 [23]	6.9	0.423	–	–	–	–	–	–
EaTR, ICCV'23 [14]	8.5	0.588	–	–	–	–	–	–
Moment-DETR, NIPS'21 [16]	4.3	0.630	–	–	–	–	–	–
TR-DETR+	6.2	0.630	–	–	–	–	–	–
QD-DETR+	5.4	0.634	–	–	–	–	–	–
UVCOM, CVPR'24 [30]	14.5	0.652	–	–	–	–	–	–
UVCOM+	12.9	0.661	–	–	–	–	–	–
Video-to-Music Retrieval:								
MVPt, CVPR'22 [27]	3.6	–	2.4	6.8	9.4	–	–	–
MVPt+	3.6	–	6.7	11.9	14.9	–	–	–
Composite solution:								
MVPt+ / UVCOM+	16.5	0.661	6.7	11.9	14.9	5.4	11.8	23.0
Video Corpus Moment Retrieval re-purposed:								
CONQUER, MM'21 [12]	39.4	0.572	5.8	11.0	13.5	4.4	9.6	18.4
MaDe (*this paper*)	10.5	**0.722**	**8.8**	**16.3**	**19.8**	**8.3**	**17.6**	**30.7**

Table 4. **Overall results**. #Params excludes the (weights-frozen) video / audio encoders.

Model	SmG	V2MR			MsG		
	mIoU	*R1*	*R5*	*R10*	*MoR1*	*MoR10*	*MoR100*
MVPt+ / UVCOM+	$0.676_{\pm 0.013}$	$7.0_{\pm 0.4}$	$12.5_{\pm 0.7}$	$15.5_{\pm 0.5}$	$5.6_{\pm 0.1}$	$12.0_{\pm 0.3}$	$24.3_{\pm 1.0}$
MaDe	$0.743_{\pm 0.017}$	$9.0_{\pm 0.5}$	$17.3_{\pm 0.7}$	$21.1_{\pm 1.0}$	$8.4_{\pm 0.3}$	$18.0_{\pm 0.5}$	$32.7_{\pm 1.4}$

Table 5. **Average performance across three distinct data splits**.

V2MR baseline that has served as a backbone in subsequent works [4, 21]. While the original MVPt adopts the CLS token from the last SA block as the music embedding, we find that mean pooling over the output of the last SA yields better performance. We refer to this variant as MVPt+. By first using MVPt+ to find a set of candidate music tracks and then using UVCOM+ to detect moments within the candidate tracks, we obtain a composite solution "MVPt+ / UVCOM+" for MsG. In addition, we include CONQUER[9] (MM21) [12], a moment-proposal based VCMR model, retraining it in the current setup.

Results. As Tab. 4 shows, the composition solution outperforms CONQUER in both SmG (mIoU 0.661 *versus* 0.572) and MsG (MoR1 5.4 *versus* 4.4) tasks, suggesting the advantage of the moment-proposal free solution in generating more accurate music grounding results. Compared to the above baselines, MaDe achieves consistently superior results across all metrics (R1 8.8 and Mo-R1 8.3), thereby establishing itself as a strong baseline for MGSV.

To verify the robustness of our findings across different data configurations, we conducted extra experiments using two newly generated random splits, following the same data partitioning protocol (Sec. 3.1). The superior performance of MaDe compared to the best baseline remains, see Tab. 5.

Human Evaluation. We further conduct a human evaluation on a random subset of 100 test videos to subjectively

[9] https://github.com/houzhijian/CONQUER

#	Setup	SmG	MsG		
		mIoU	*MoR1*	*MoR10*	*MoR100*
0	Full-setup	**0.722**	**8.3**	**17.6**	**30.7**
Uni-modal Feature Enhancement:					
1	*w/o* SA	0.699	6.4	14.3	27.4
2	SA $\rightarrow$ MLP	0.707	6.8	14.1	26.9
Video-Music Matching:					
3	$cs(h(v), h_0(M))$ as p_s	0.708	7.1	15.8	29.1
4	$cs(h(v), h_1(M))$ as p_s	0.707	6.3	15.1	29.0
5	single loss	0.715	7.5	16.8	29.2
6	$h_0(M) + h_1(M)$	0.716	6.9	16.0	28.3
Music Moment Detection:					
7	*w/o* $SA_{\times 2}$	0.705	8.1	16.7	29.1
8	$SA_{\times 2} \rightarrow$ CA	0.697	7.1	16.4	29.1
9	**0** as ϕ_0	0.709	7.4	16.4	29.0
10	$h_0(M)$ as ϕ_0	0.719	8.0	17.4	30.4
11	$h_1(M)$ as ϕ_0	0.718	7.5	16.9	29.5
12	#Query-tokens: $1 \rightarrow 10$	0.716	7.6	16.7	30.6
13	$(p_c, p_w) \rightarrow p_c$	0.706	7.3	16.5	29.0

Table 6. **Ablation study of MaDe**.

assess the quality of the MsG results. In particular, each video is individually paired with the music moment predicted by MaDe and its counterpart by MVPt+ / UVCOM+. A subject is asked to watch the two resultant videos (labeled as Video A and Video B without revealing the method used) and, based on their overall visual and audio perception, select one of the following four options: *A is better*, *B is better*, *neither acceptable*, or *both acceptable*. Our evaluation team comprised 15 males and 5 females, all in their 20s and 30s, with frequent exposure to online short videos and music. Fig. 4 shows that MaDe exceeds the composite solution by approximately 10% (38.1% *vs*. 28.6%).

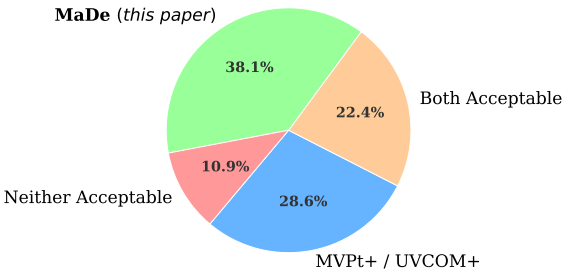


Figure 4. **Human evaluation results**.

5.3. Ablation Study of MaDe

For a better understanding of the superior performance of MaDe, an ablation study is conducted as follows, with the results summarized in Tab. 6.

SA for Feature Enhancement. Removing SA (Setup-1) or replacing it by an MLP (Setup-2) results in performance loss, with MoR1 decreased to 6.4 and 6.8, respectively.

Num. CAs	MoR1	MoR10	MoR100
1	7.6	16.7	28.3
2	7.9	17.3	28.5
3	7.4	16.0	27.2
4	8.2	16.9	30.0
5	7.3	16.2	29.2
6	**8.3**	**17.6**	**30.7**

Table 7. **Numbers of CAs for decoding**.

X-Pool for Video-Music Matching. Using only the mean-pooled embeddings (Setup-3) results in a drop in MoR1 from 8.3 to 7.1. Using exclusively the X-Pooled embedding (Setup-4) leads to a more significant decline to 6.3. This result demonstrates that the two types of embeddings are complementary to each other. Regarding their joint exploitation, minimizing the joint loss outperforms using a single loss (Setup-5), and the combination based on similarity is superior to the feature-addition counterpart (Setup-6).

SA for Multimodal Feature Fusion. Removing $\text{SA}_{\times 2}$ (Setup-7) or replacing them with a CA (the music tokens as Q)(Setup-8) consistently results in performance loss.

Choice of the Initial Content Token ϕ_0. We try three alternatives to $h(v)$, that is, $\mathbf{0}$ (Setup-9) and two target-modality features $h_0(M)$ (Setup-10) and $h_1(M)$ (Setup-11). The choice of using $h(v)$ as ϕ_0 remains the best.

Number of the Query Tokens. We experimented with 10 query tokens, the default value in DETR and Moment-DETR, but observed lower performance (Setup-12).

Predicting p_c only. We modified MaDe's detection head to predict only the moment center p_c, using the duration of the query video as the moment width (Setup-13). The resulting lower performance confirms the necessity of learning to predict temporal boundaries explicitly.

Number of CAs for Decoding. The influence of the number of CAs used for decoding is reported in Tab. 7.

CA for fusion. We conduct a controlled comparison between QD-DETR, UVCOM and their CA-free variants (QD-DETR+, UVCOM+) at varied training data scales. Fig. 5 reveals an inverse relationship: QD-DETR declines with more data, while QD-DETR+ improves. This result suggests that CA-based multimodal feature fusion is the source of incompatibility with the MGSV task's demands.

Efficiency Analysis. Given the raw visual / audio tokens precomputed, MaDe performs music grounding for 512 video-music pairs in 0.09 seconds on average, see Tab. 8.

6. Conclusions

In this paper, we have introduced Music Grounding by Short Video (MGSV) as a novel task and constructed MGSV-EC, a large-scale real-world dataset comprising 53k short videos associated with 35k different music moments from 4k unique tracks. Through extensive experiments on

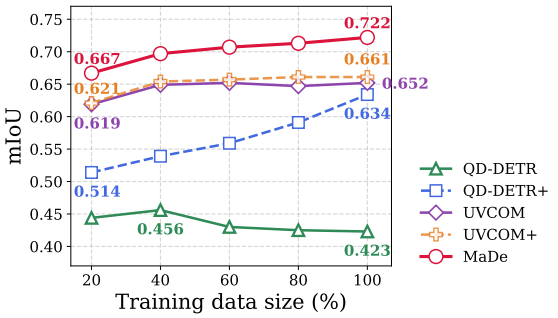


Figure 5. **SmG performance with varying training data sizes**.

Module	#Params (M)	Time cost (sec.)
Feature enhancement	2.1	0.006
Video-music matching	0.3	0.052
Music moment detection	8.1	0.033
Total	10.5	0.091

Table 8. **Inference efficiency per mini-batch**. Batch size: 512. Raw visual / audio tokens are precomputed and thus excluded.

MGSV-EC, we draw the following conclusions. Among the adapted Video Grounding models for single-music grounding, UVCOM achieves the best performance. Removing cross-attention (CA) blocks from the multimodal feature fusion module consistently improves model performance. For the Video-to-Music Retrieval model MVPt, obtaining video and music embeddings by mean pooling outperforms its default option of using the CLS token. The superior performance of MaDe on both single-music grounding and music-set grounding can be collectively attributed to the following optimizations: employing SA for unimodal feature enhancement, avoiding CA for multimodal feature fusion, jointly utilizing mean pooling and X-Pool for video-music matching, and using the video embedding as the initial content token for decoding. As such, MaDe serves as a strong baseline for MGSV.

Limitations. Our study has clearly shown that CA-based multimodal feature fusion negatively impacts music grounding performance. The surface reason is that such a fusion mechanism tends to cause overfitting. However, the question of how video-music interaction needed for Music Grounding fundamentally differs from the text-video interaction required by Video Grounding remains open. How to customize music grounding with respect to the users' personal preferences is worthy of future research.

Acknowledgments This research was supported by NSFC (No.62172420) and Kuaishou. We thank Yingtong Liu, Yuchuan Deng, Yiyi Chen, and Bo Wang for valuable discussion and feedback on this research.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[2] Houlun Chen, Xin Wang, Hong Chen, Zeyang Zhang, Wei Feng, Bin Huang, Jia Jia, and Wenwu Zhu. VERIFIED: A video corpus moment retrieval benchmark for fine-grained video understanding. In *NeurIPS*, 2024. 3

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 6

[4] Zhikang Dong, Bin Chen, Xiulong Liu, Pawe Polak, and Peng Zhang. MuseChat: A conversational music recommendation system for videos. In *CVPR*, 2024. 1, 3, 7

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, 2023. 4

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 6

[7] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech*, 2021. 4

[8] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-Pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022. 5

[9] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 6

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 6

[11] Sungeun Hong, Woobin Im, and Hyun S Yang. CBVMR: Content-based video-music retrieval using soft intra-modal structure constraint. In *ICMR*, 2018. 2, 3

[12] Zhijian Hou, Chong-Wah Ngo, and Wing Kwong Chan. CONQUER: Contextual query-aware ranking for video corpus moment retrieval. In *ACMMM*, 2021. 2, 3, 4, 7

[13] Yu-Siang Huang, Szu-Yu Chou, and Yi-Hsuan Yang. Pop music highlighter: Marking the emotion keypoints. *ISMIR*, 2018. 3, 4

[14] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, 2023. 2, 6, 7

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[16] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 2, 5, 6, 7

[17] Bochen Li and Aparna Kumar. Query by video: Cross-modal music retrieval. In *ISMIR*, 2019. 1

[18] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. SEA: Sentence encoder assembly for video retrieval by textual queries. *TMM*, 2021. 6

[19] Jingyu Liu, Minquan Wang, Ye Ma, Bo Wang, Aozhu Chen, Quan Chen, Peng Jiang, and Xirong Li. D&M: Enriching e-commerce videos with sound effects by key moment detection and SFX matching. In *AAAI*, 2025. 4

[20] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[21] Daniel McKee, Justin Salamon, Josef Sivic, and Bryan Russell. Language-guided music recommendation for video via prompt analogies. In *CVPR*, 2023. 1, 3, 7

[22] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 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 *CVPR*, 2023. 2, 6, 7

[24] Laure Prétet, Gaël Richard, Clément Souchier, and Geoffroy Peeters. Video-to-music recommendation using temporal alignment of segments. *TMM*, 2022. 3

[25] 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 *ICML*, 2021. 4, 6

[26] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. TR-DETR: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *AAAI*, 2024. 2, 6, 7

[27] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *CVPR*, 2022. 1, 2, 3, 6, 7

[28] Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, and Xirong Li. Holistic features are almost sufficient for text-to-video retrieval. In *CVPR*, 2024. 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5, 6

[30] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, 2024. 2, 6, 7

[31] Sunjae Yoon, Jiajing Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Changdong Yoo. Selective query-guided debiasing for video corpus moment retrieval. In *ECCV*, 2022. 3

[32] 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 *CVPR*, 2019. 2

[33] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *SIGIR*, 2021. 3

[34] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 2