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

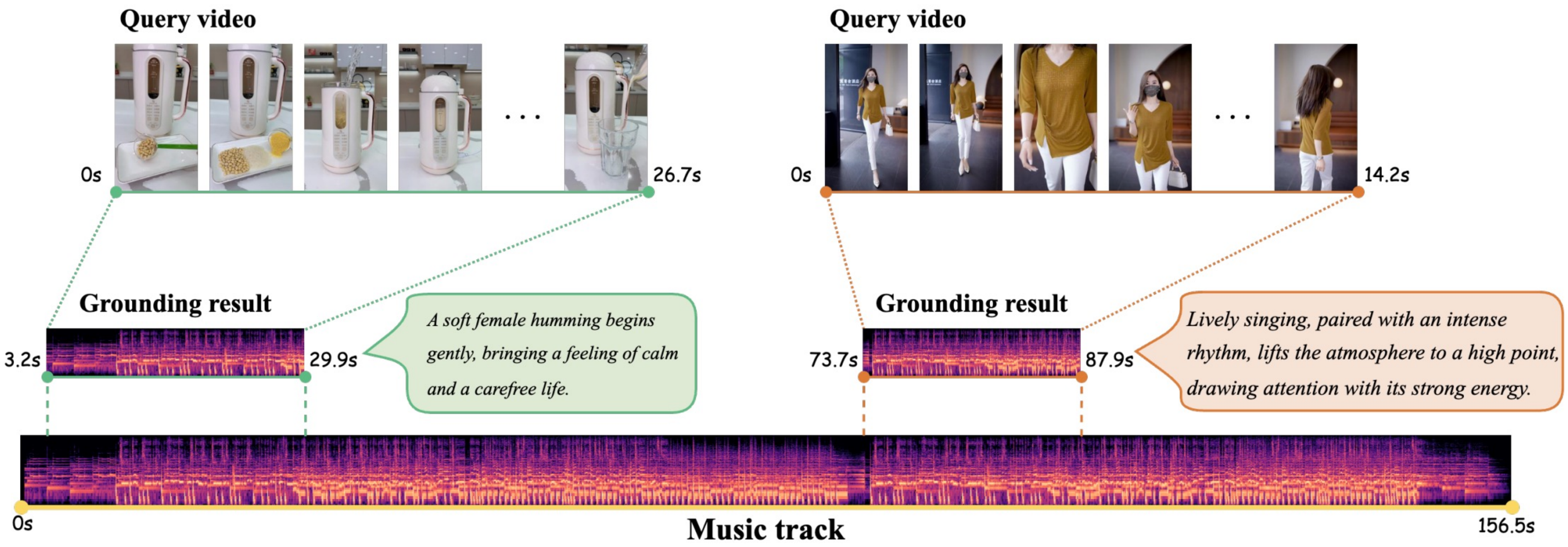
xinzijie@ruc.edu.cn

Background

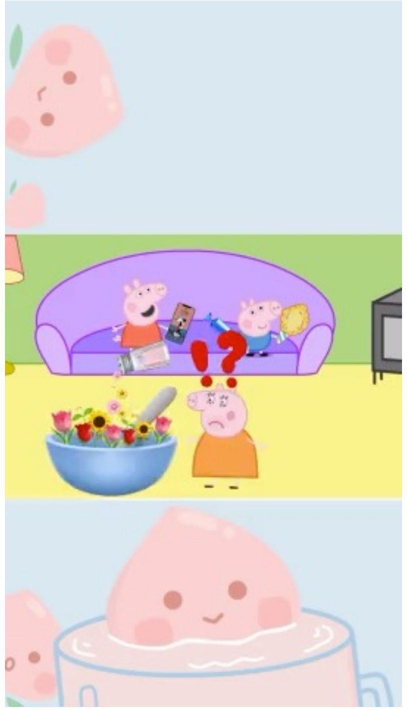
- **Short videos + Music = More engaging**: Adding background music "helps complete a short video" – think of TikToks or YouTube Shorts with catchy songs.
- **Current practice is manual**: Creators pick a song and manually trim it to fit the video. This can be tedious and requires timing the music just right.
- **Previous research (V2MR)**: Some systems can suggest a whole music track for a video (Video-to-Music Retrieval), but they don't tell you which part of the song to use.

Motivation

- **The Gap**: Music tracks are usually much longer than short videos, so just getting a song isn't enough – you need the right moment. How can we automate finding that perfect moment?



BGM Showcase Generated by our Model

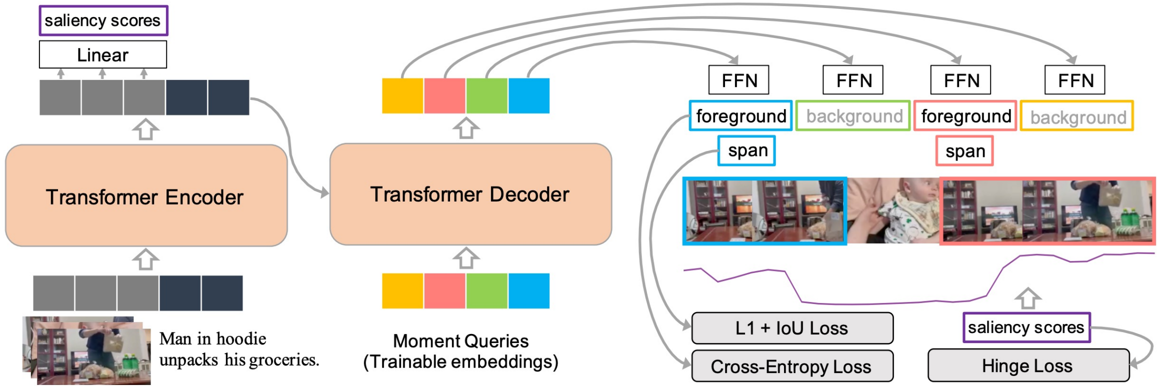


More showcases can be seen in our project page: ***https://rucmm.github.io/MGSV***

Related Work

RENMIN UNIVERSITY OF CHINA

➢ Transformers-based Video Temporal Grounding



Moment-DETR

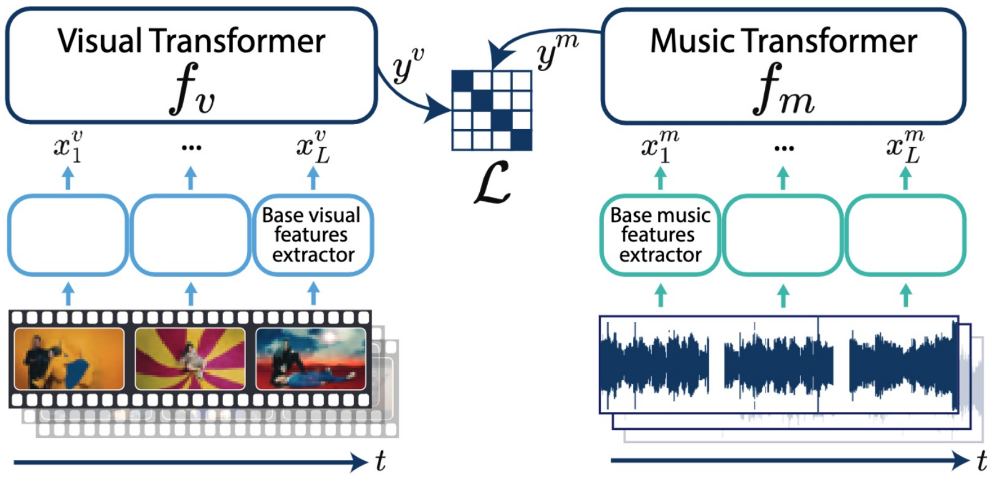
Model	Unimodal Feat. Enhancement	Multimodal Feat. Fusion	Init. Content Token ϕ_0	Decoder
Moment-DETR[16]	$FC_{\times 2}$	$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
QD-DETR[23]	$FC_{\times 2}$	$CA_{\times 2}+SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
QD-DETR+	$FC_{\times 2}$	$SA_{\times 2}$	0	$SA\text{-}CA_{\times 2}$
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]	$FC_{\times 2}$	$Dual\text{-}CA_{\times 3} + DBIA+LRP+SA_{\times 3}$	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	$CA_{\times 6}$

Key elements in current DETR-based models for video grounding

[1] Lei, et al. Detecting moments and highlights in videos via natural language queries. In NeurIPS, 2021.

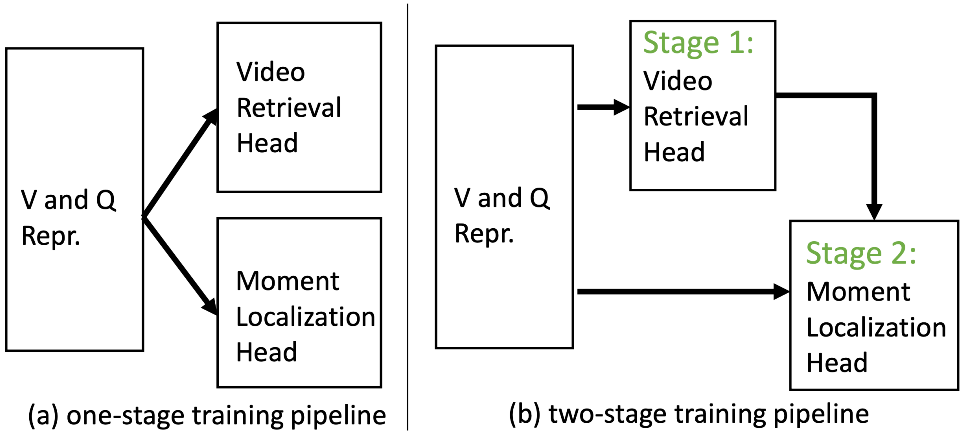
Related Work

➢ Video-to-Music Retrieval



MVPt

➢ Video Corpus Moment Retrieval

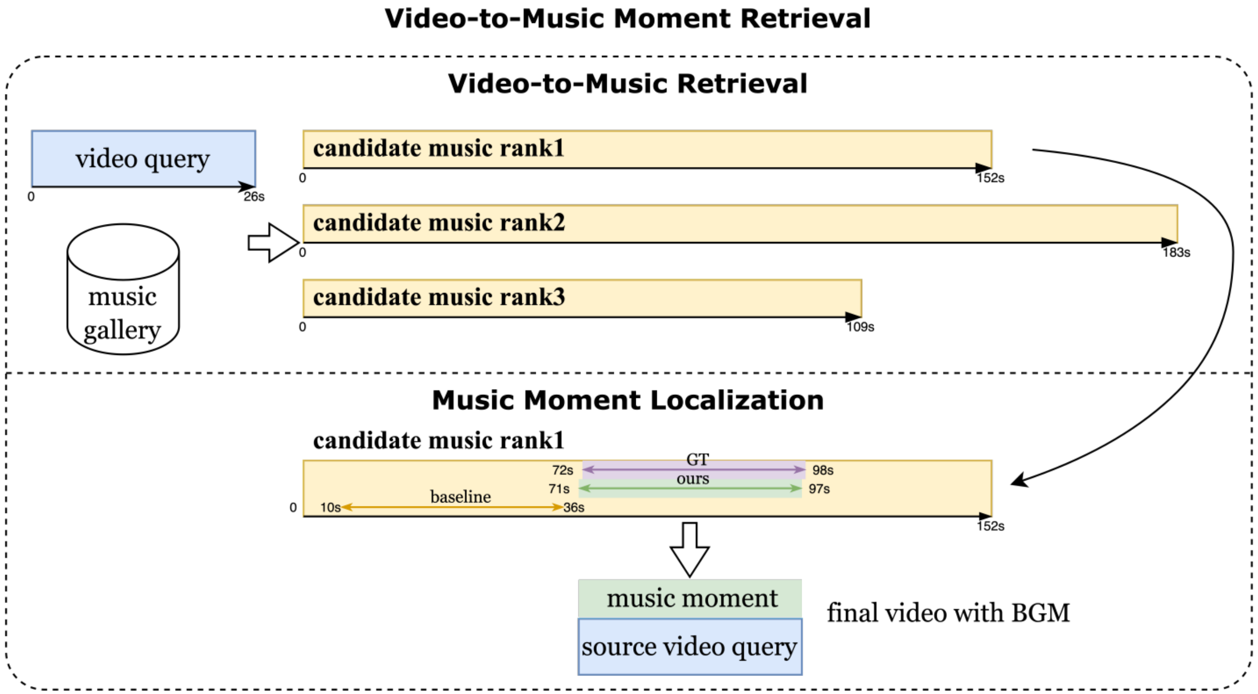


CONQUER

[1] Sur´ıs, et al. It's time for artistic correspondence in music and video. In CVPR, 2022.
[2] Hou, et al. CONQUER: Contextual query-aware ranking for video corpus moment retrieval. In ACMMM, 2021.

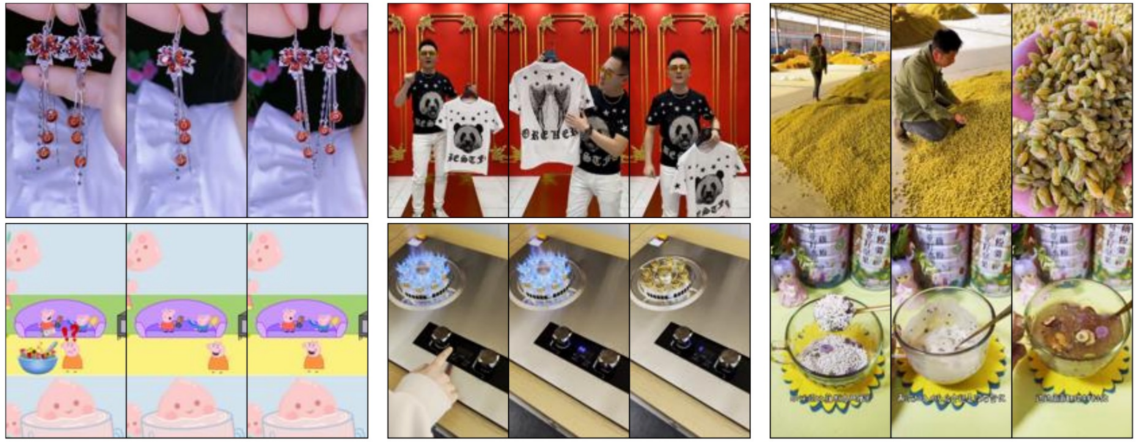
New Task - Music Grounding by Short Video

- ✓ **New Task (MGSV):** "Music Grounding by Short Video" means given a short video, find the exact segment (start and end time) of a music track that best fits the video.
- ✓ **Difference from previous approach:** Unlike just retrieving a whole song, MGSV finds *which part* of the song to use, bridging the gap between track selection and manual editing.



Dataset Construction – **MGSV-EC**

- **New Benchmark "MGSV-EC"**: To train and evaluate this new task, we built a large dataset called MGSV-EC (E-commerce). It has ~**53,000 short videos**, each paired with a specific music segment (total **35,000 unique music clips** from **4,000 songs**).
- **Data source**: These video–music pairs come from an E-commerce video creation platform. 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.

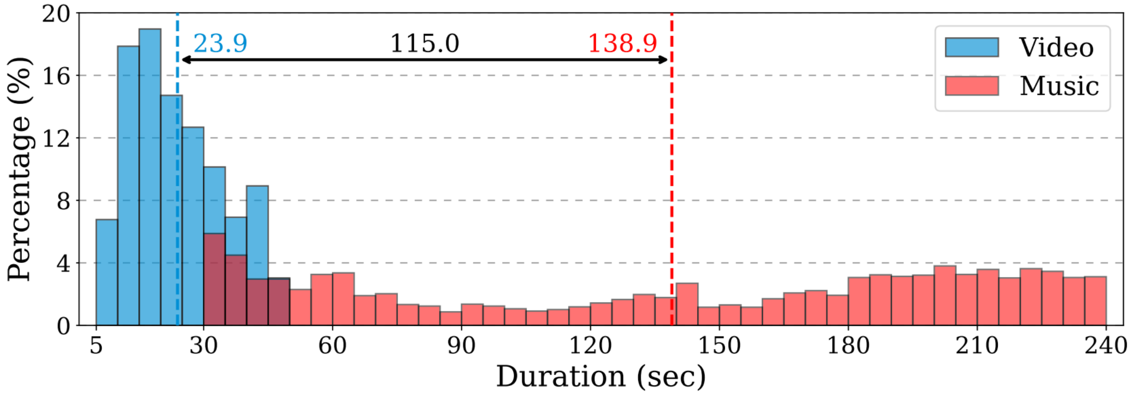


Snapshots of video samples

Video tag cloud

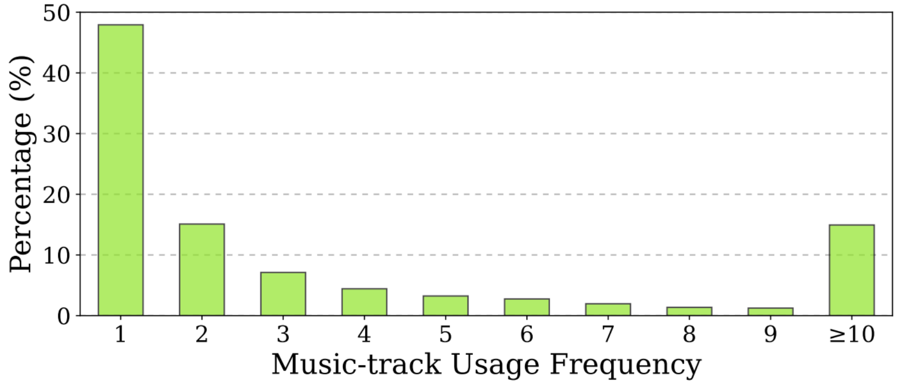
Dataset Analysis & Statistics

➢ Large & Diverse



Distribution of video duration / music-track duration

➢ Long-tail distribution



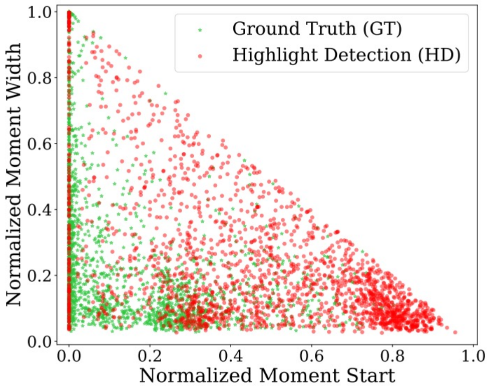
Usage frequency of music-track for video

➢ Train/Val/Test split

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

Overview of the MGSV-EC dataset

➢ Directly using a chorus or intro is not enough



GT *vs.* HD moment position

Evaluation Protocol

- **Single-music Mode**: 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.
- **Music-set Mode**: 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.

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

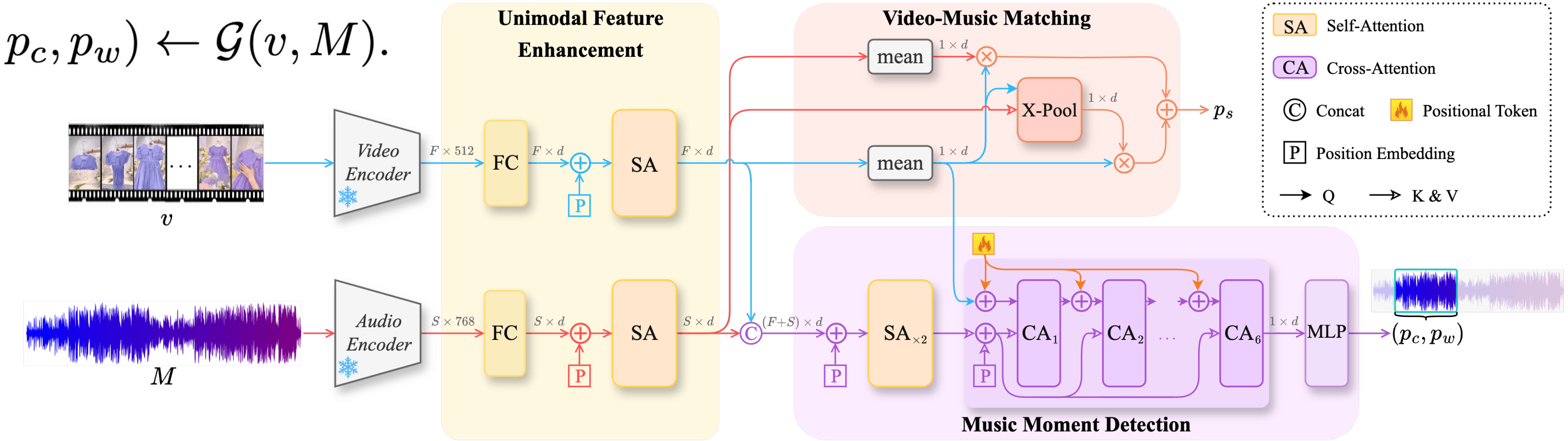
Evaluation modes, (sub-)tasks and metrics

Method Overview – Matching + Detection

RENMIN UNIVERSITY OF CHINA

- **Two sub-tasks in one**: Solving MGSV involves (1) finding the right music track and (2) finding the right segment within that track. The authors' solution combines both into a single unified approach.
- **High-level idea**: The model takes the video and a candidate song as input and outputs a predicted start time and end time that tells which part of the song matches the video.
- **Our model MaDe:** Our proposed model (MaDe, for Matching & Detection) tackles video-to-music matching and moment detection together in an end-to-end deep neural network. Instead of doing retrieval first then trimming, it learns to do both jointly.

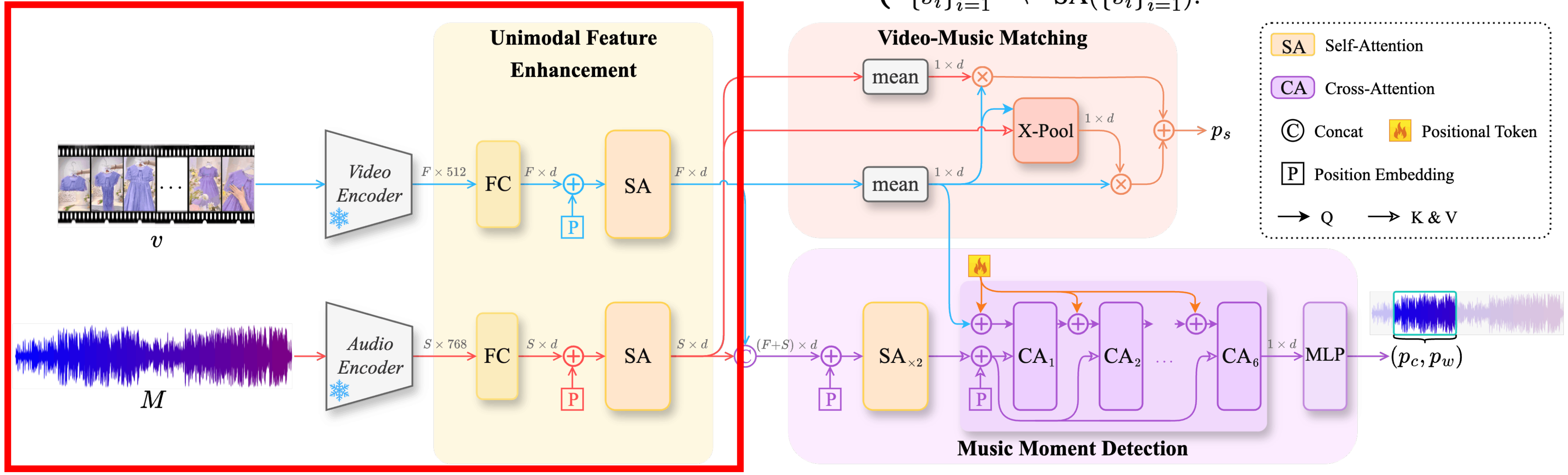
$$(p_s, p_c, p_w) \leftarrow \mathcal{G}(v, M).$$



Model Architecture Diagram

1. **Feature Extraction**: The video (as frames) and music (as spectrogram) are processed separately by pre-trained encoders, ViT for video and AST for audio. These generate high-level feature sequences over time.
2. **Temporal Modeling**: Each modality passes through its own temporal module to capture how content patterns—like rhythm in music or scene changes in video.

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

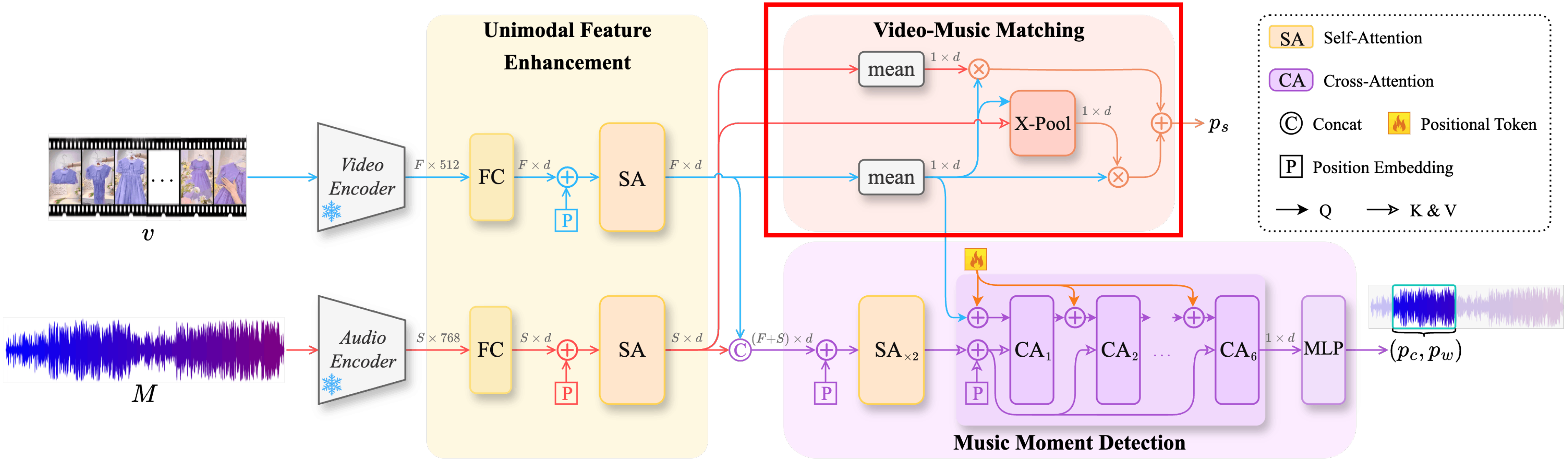


Model Architecture Diagram

RENMIN UNIVERSITY OF CHINA

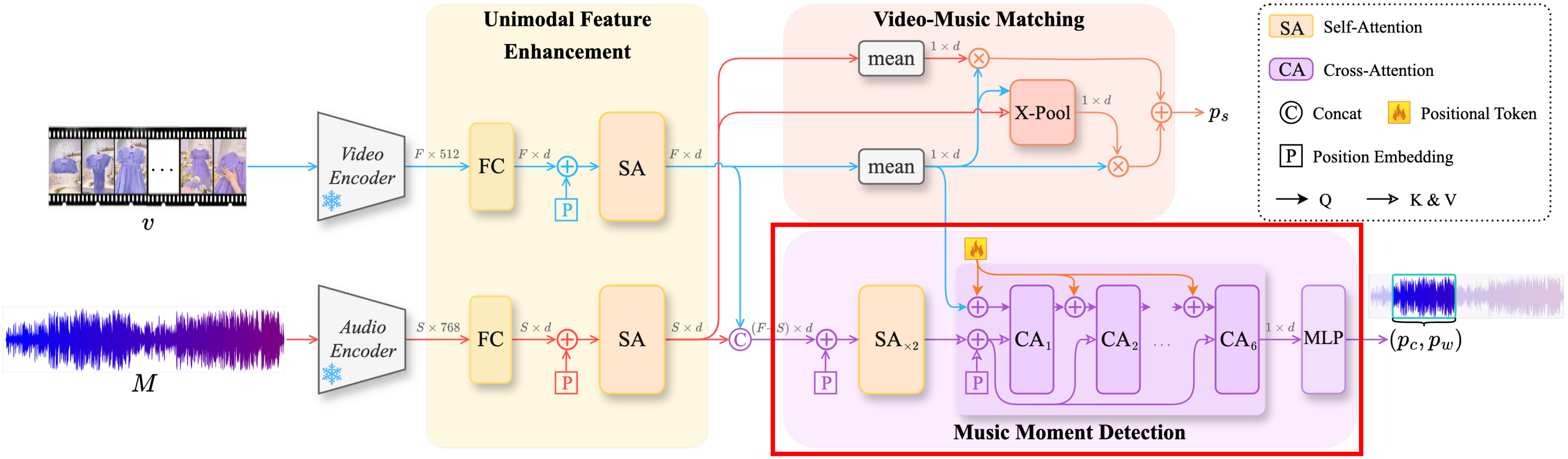
3. Cross-Modal Fusion & Matching: A cross-attention module lets the video attend to relevant parts of the music. A transformer then fuses the two modalities into a shared representation.

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



4. DETR-inspired Decoder & Output: A DETR-style decoder uses learnable queries to predict the start and end of the best-matching music segment. The model is trained end-to-end by comparing predictions to ground-truth segments.

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



Performance Evaluation

Experimental setup: The model is evaluated on the MGSV-EC dataset.

Metrics involve retrieval success at various ranks (*e.g.* the correct segment make it to the top-1, top-5, top-10, *etc.*) and localization accuracy (how well the predicted segment overlaps the ground truth).

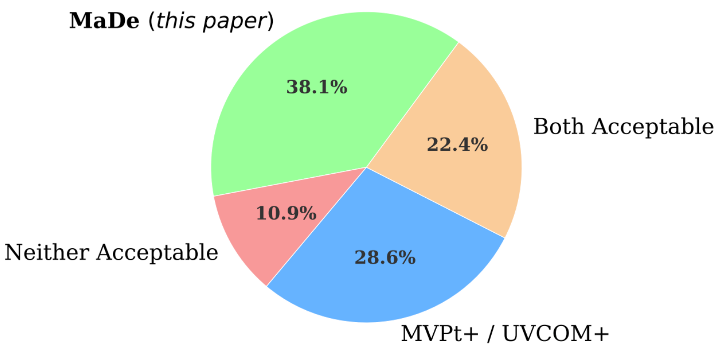
Model	#Params (M)	SmG *mIoU*	V2MR *R1*	V2MR *R5*	V2MR *R10*	MsG *MoR1*	MsG *MoR10*	MsG *MoR100*
Video Grounding re-purposed:								
TR-DETR, AAAI'24 [26]	7.8	0.393	–	–	–	–	–	–
QD-DETR, CVPR'23 [23]	6.9	0.423	–	–	–	–	–	–
EaTR, ICCV'23 [14]	8.5	0.588	–	–	–	–	–	–
Moment-DETR, NIPS'21 [16]	4.3	0.630	–	–	–	–	–	–
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**

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

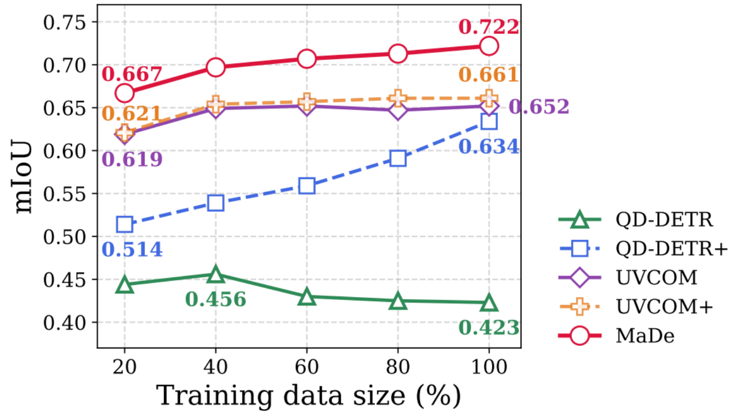
Ablation Study

#	Setup	SmG mIoU	MsG MoR1	MsG MoR10	MsG 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	$\mathbf{0}$ 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

Ablation study of MaDe



Human evaluation results

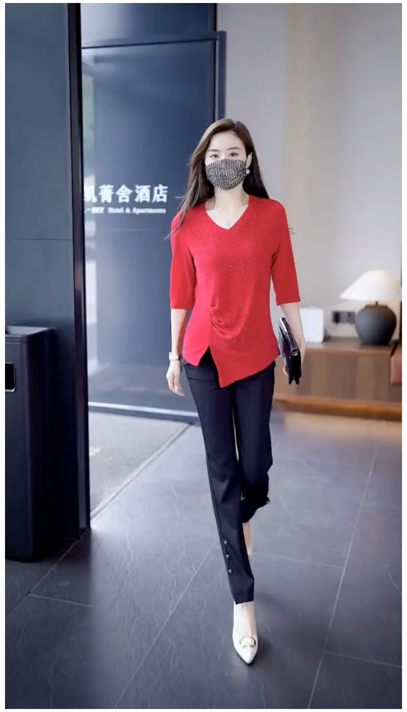


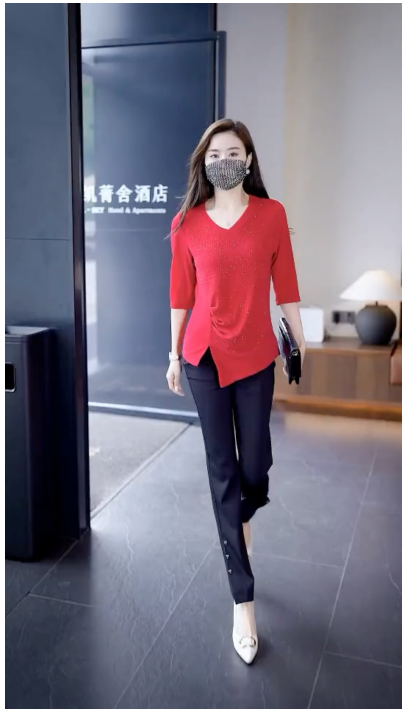
SmG performance with varying training data sizes

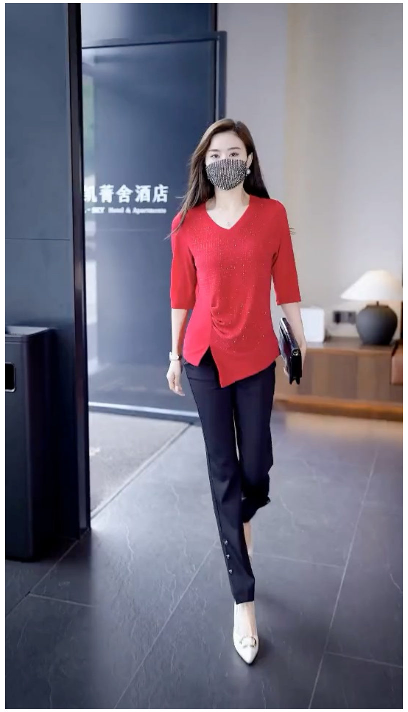
Generated BGM Comparisons

中国人民大學 RENMIN UNIVERSITY OF CHINA

Original background music | BGM generated by **MaDe** *(ours)* | Composite solution







More BGM Comparsions can be seen in our project page: ***https://rucmm.github.io/MGSV***

Conclusion

Contributions

- **New Task**: Introduced MGSV and showed the limits of traditional video-to-music retrieval.
- **Large-Scale Dataset**: Created MGSV-EC with 53k video–music pairs, using a semi-automatic annotation method.
- **Our Model**: Proposed MaDe, an end-to-end model that combines retrieval and localization, outperforming simple baselines.

Findings & Future Directions :

Despite difficulty of the task, our model learns meaningful video–audio matches. While accuracy is still modest, the model proves the task is feasible and provides a solid foundation for future improvements.

Music Grounding by Short Video

Thank you for your attention!

7 x 29
北京 海淀

arXiv:2408.16990

xxayt/MGSV

rucmm.github.io/MGSV

Welcome to star/cite if useful!

Feel free to contact us if there are any questions!