Partially Relevant Video Retrieval

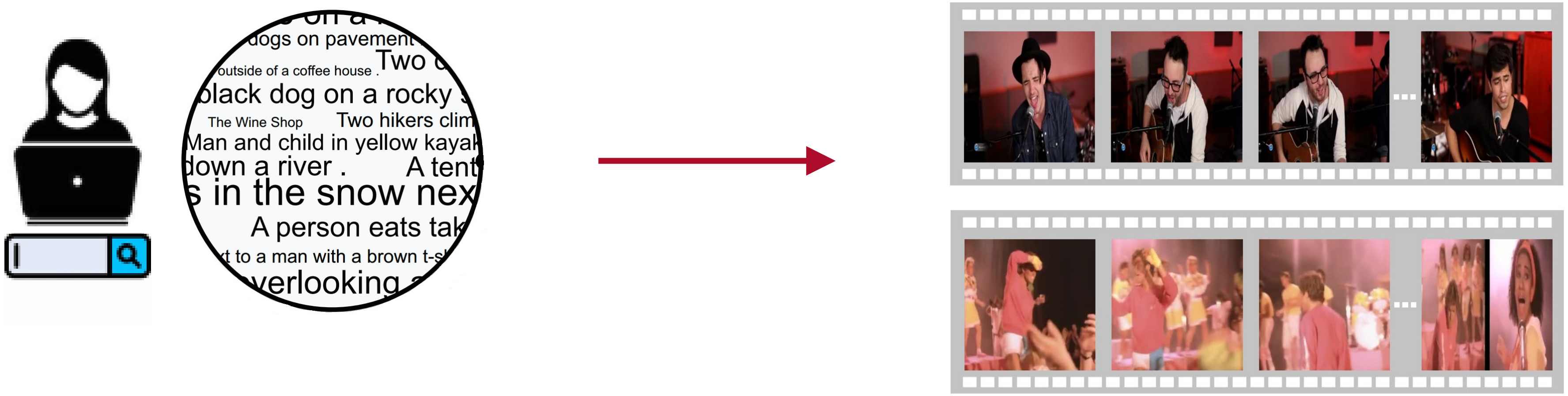
Jianfeng Dong[1], **Xianke Chen**[1], Minsong Zhang[1],
Xun Yang [2], Shujie Chen[1], Xirong Li[3], Xun Wang[1]

[1] Zhejiang Gongshang University
[2] University of Science and Technology of China
[3] Renmin University of China

Text-to-Video Retrieval（T2VR）

- Give a textual query, T2VR asks to retrieve videos that are semantically relevant to the given query from a gallery of videos.

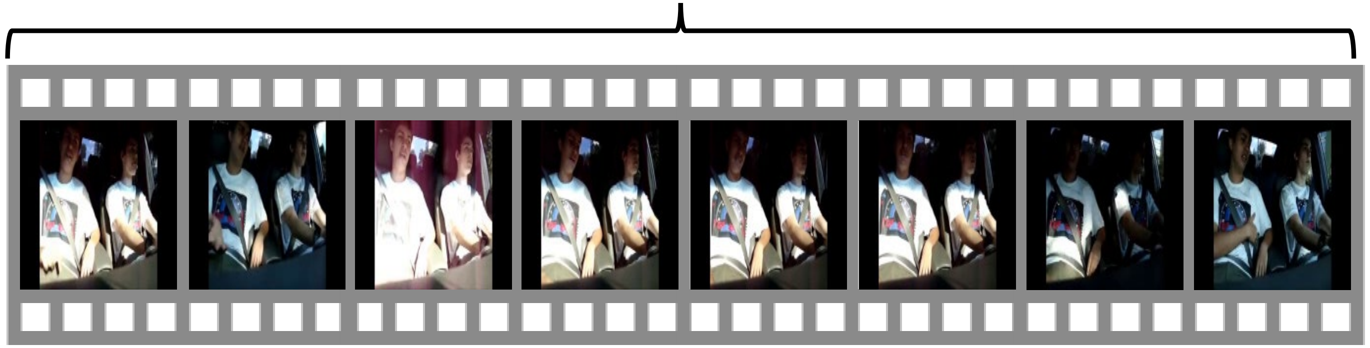
Textual Query **Videos**



Weakness of conventional T2VR methods

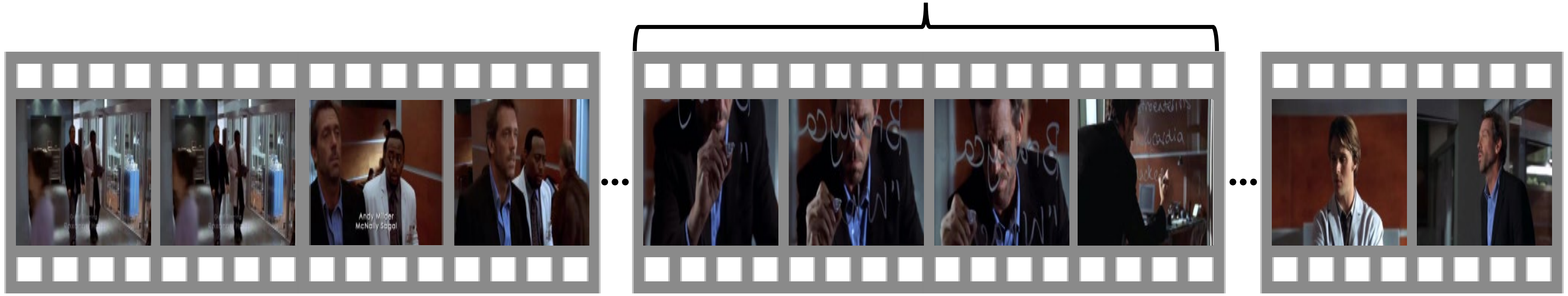
- Video-text pairs in training datasets are **fully relevant**:

Query: Two man talk to each other and drive the car.



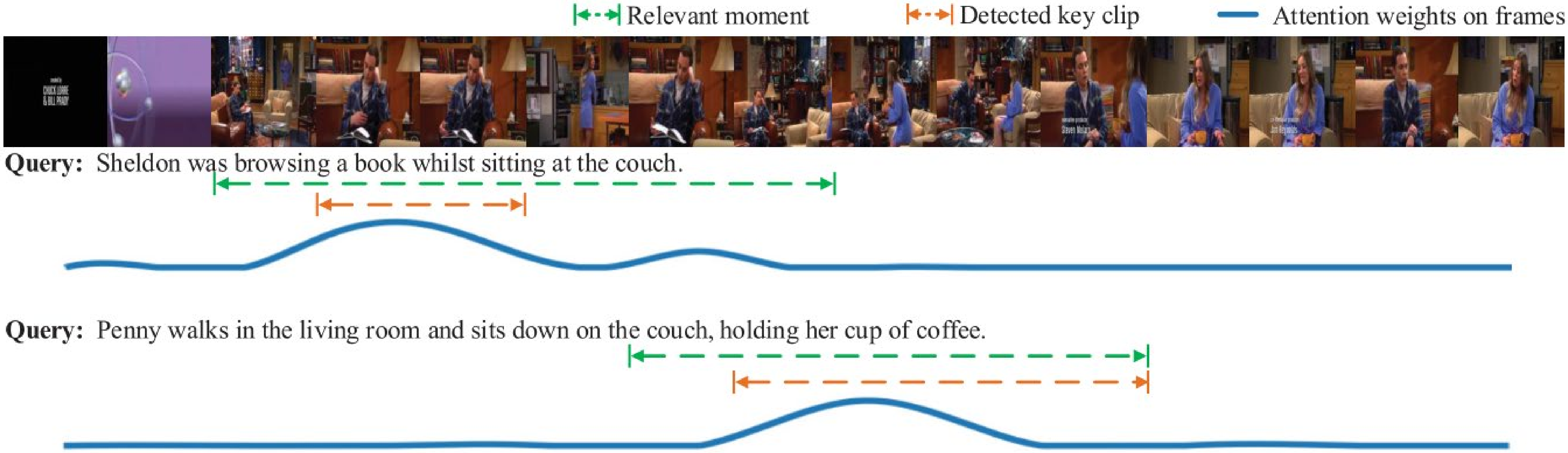
- Video-text pairs in real-world applications are mostly **partially relevant**:

Query: House writes on a glass surface with a dry erase marker.



Partially Relevant Video Retrieval（PRVR）

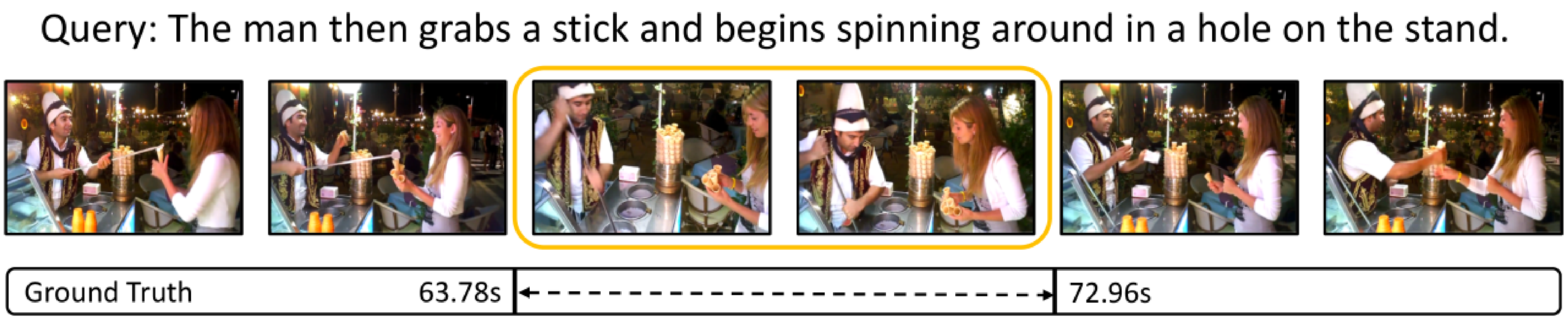
- Give a textual query, PRVR aims to retrieval a **video** which contains a (short) moment relevant w.r.t the query from **a large collection of untrimmed videos**.



How is PRVR different?

- Single Video Moment Retrieval （SVMR）

The SVMR task is to retrieve **moments** semantically relevant to the given query from **a given single untrimmed video**.

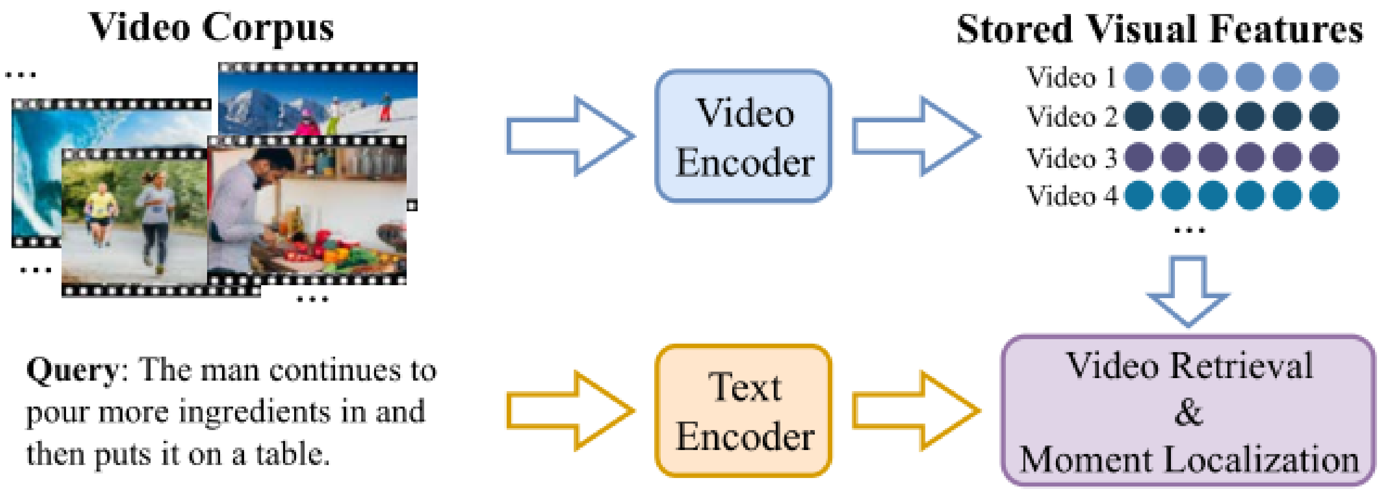


Zhang *et al.* Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos. ACM MM 2020.

How is PRVR different?

- Video Corpus Moment Retrieval （VCMR）

The VCMR task is to retrieve **moments** semantically relevant to the given query from **a large collection of untrimmed videos**.



Zhang *et al.* Video Corpus Moment Retrieval with Contrastive Learning. SIGIR 2021.

Related work

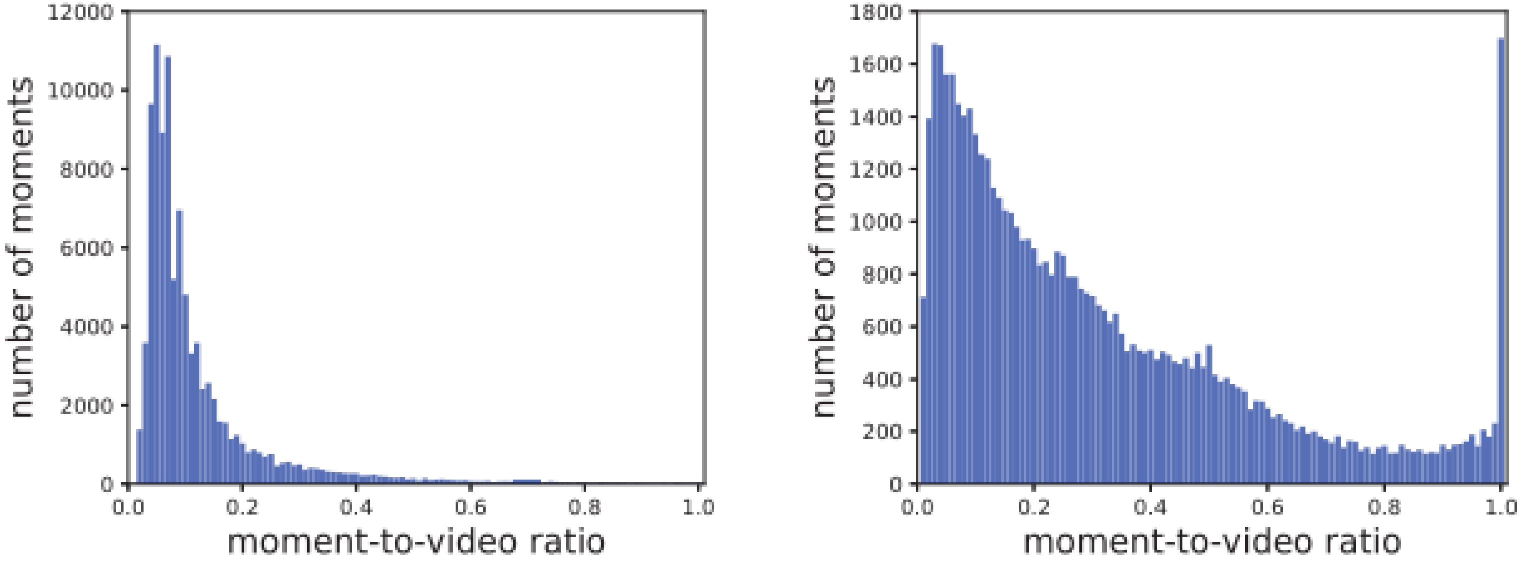
- We summarize the differences of the above-mentioned related tasks and PRVR task in two aspects.

	Labels needed in Training			Task in inference	
	Video	Clip	Moment	Retrieve target video in a collection of videos	Locate moment in a given single video
T2VR		√		√	
SVMR	√		√		√
VCMR	√		√	√	√
PRVR	√			√	

Our Method

PRVR is more practical but challenging

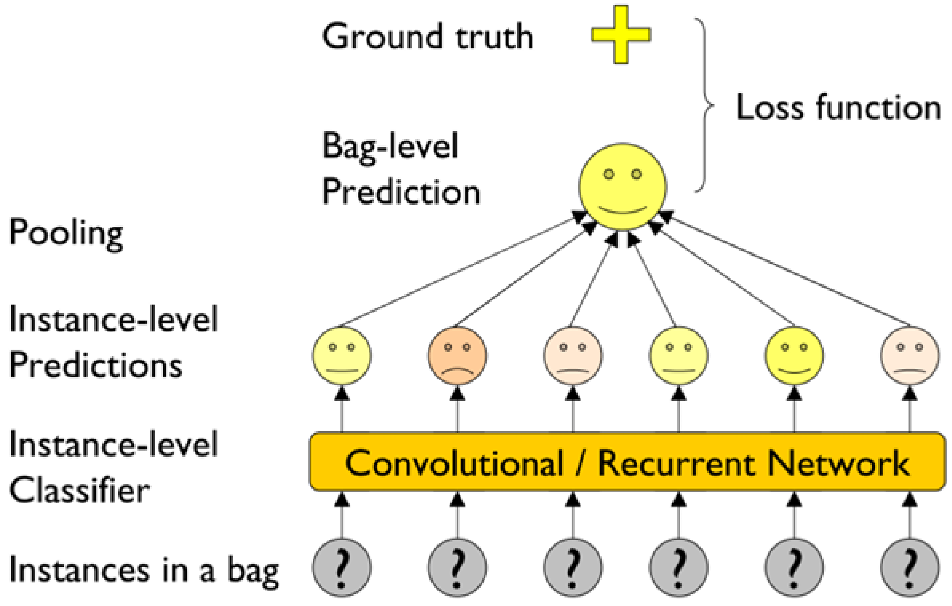
- How to make the model accurately construct the **partial relevance** between text query and its corresponding untrimmed video, and **where the relevant moment is localized and how long it lasts** are both unknown.



(a) TVR (b) ActivityNet Captions

We formulate the PRVR task as a MIL problem

- Multiple Instance Learning (MIL) is a classical framework for learning from weakly annotated data, and widely used for classification tasks.
- We formulate the PRVR task as a MIL problem. A video can simultaneously viewed as a bag of video clips and a bag of video frames.



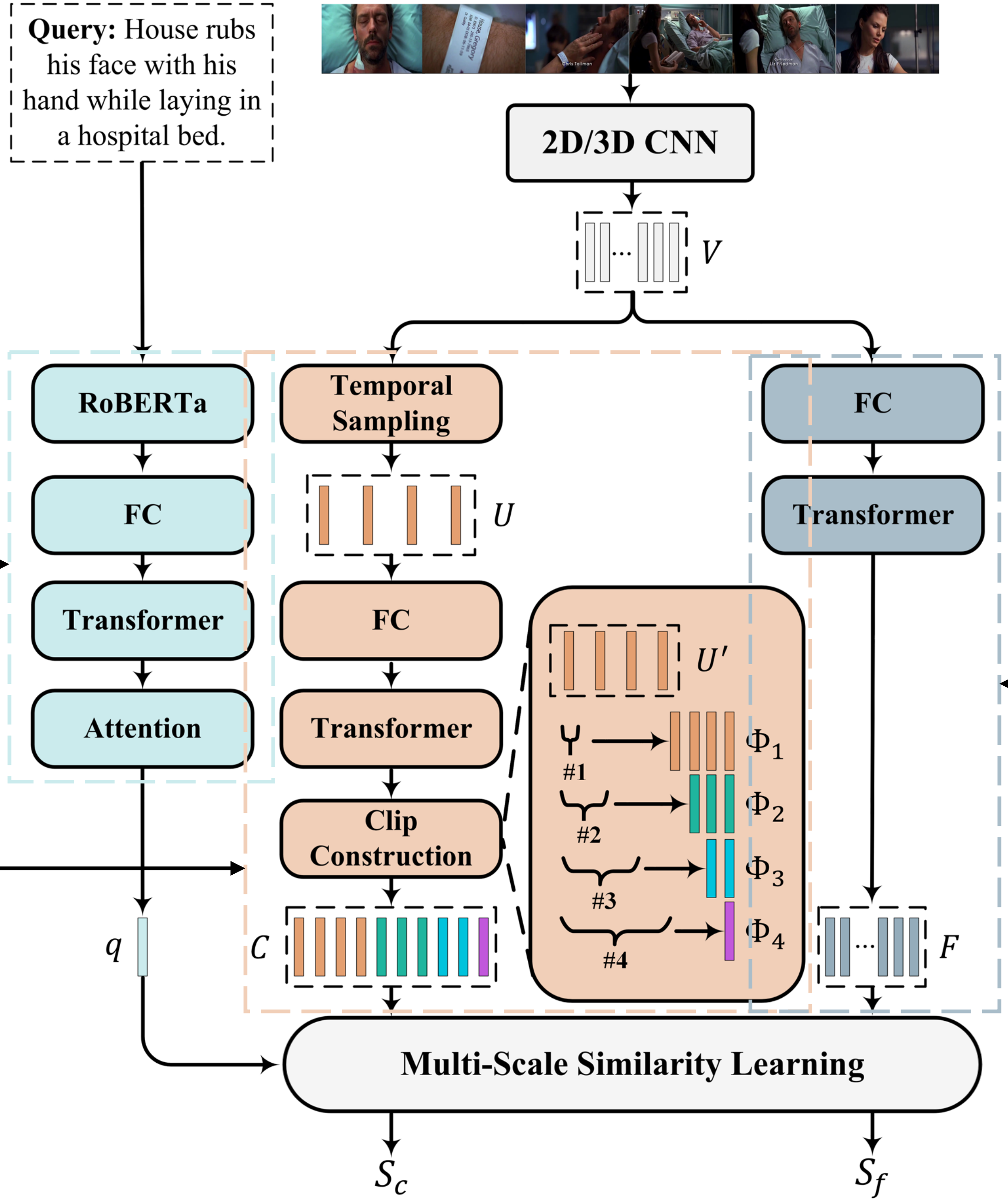
Wang *et al.* A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. ICASSP 2019.

Framework

Sentence Representation

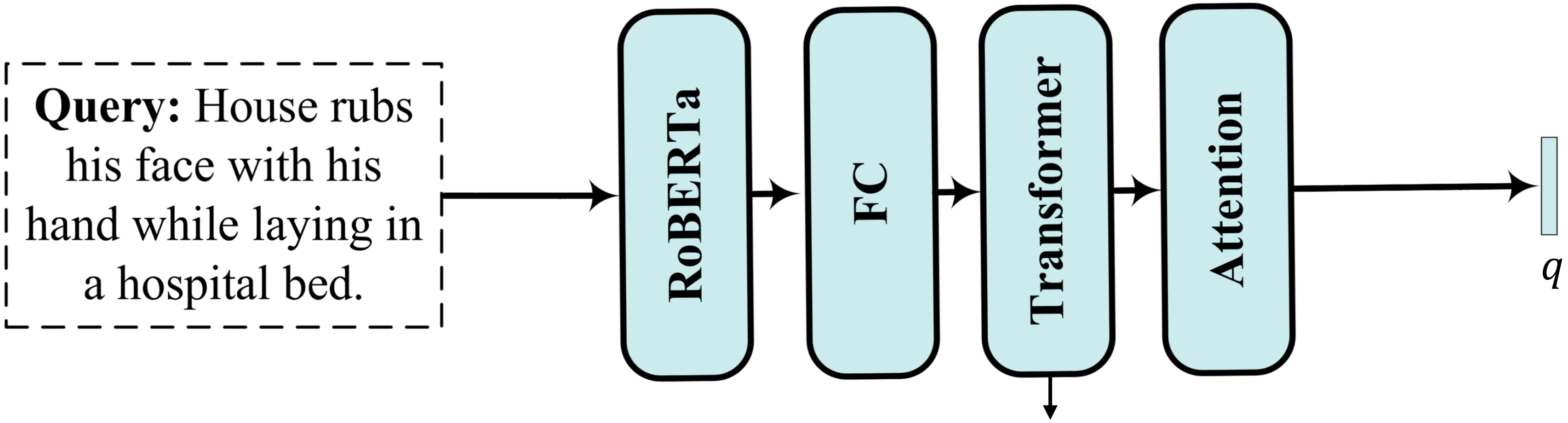
Clip-scale Video Representation

Frame-scale Video Representation



Sentence Representation

- We adopt the method by [Lei et al. ECCV 2020] to encode text query, considering its good performance on VCMR.

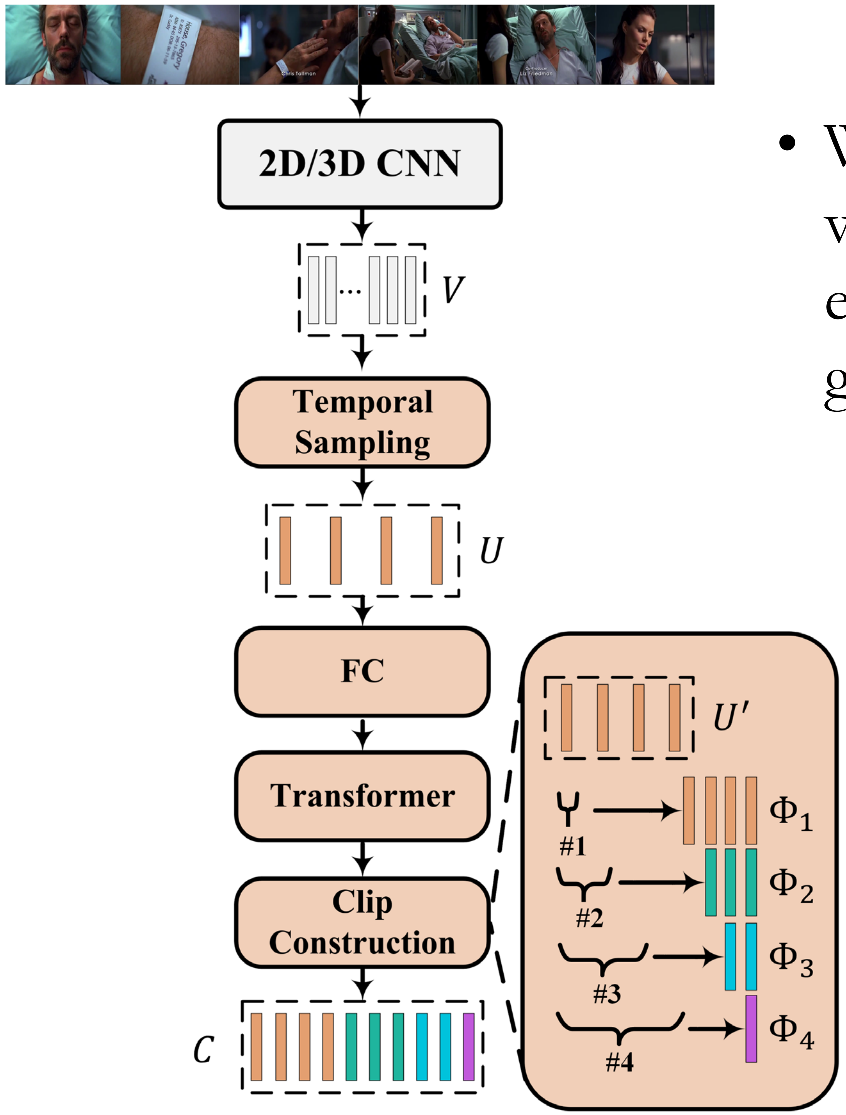


$$Q = \{q_i\}_{i=1}^{n_q} \in \mathbb{R}^{d \times n_q}$$

Attention $\left\{ \; q = \sum_{i=1}^{n_q} \alpha_i^q \times q_i \,, \alpha^q = Softmax(w^T Q) \right.$

Lei et al. TVR: A large-scale dataset for video-subtitle moment retrieval. In ECCV 2020.

Clip-scale Video Representation

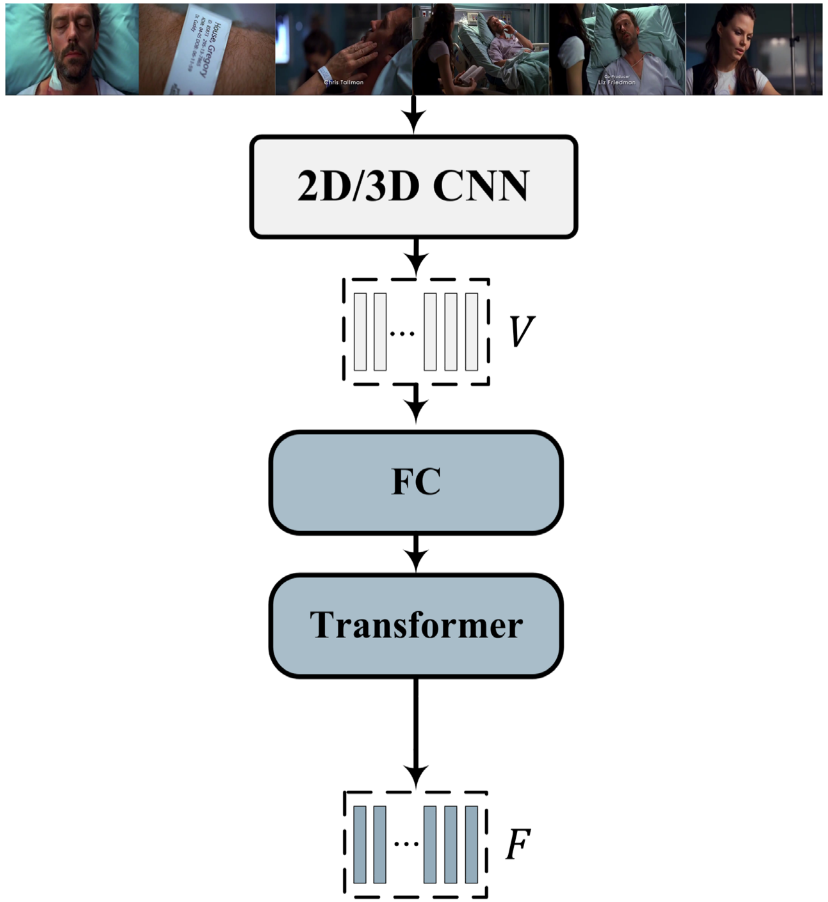


- We downsample the features into a fixed number of feature vectors and use an FC layer and a one-layer Transformer to encode it, then employ a multi-scale sliding window strategy to generate video clips vectors.

Clip construction
- Varied sliding window sizes $\{1,2,\ldots,n_u\}$
- Resultant feature sequences $\{\Phi_1,\Phi_2,\ldots,\Phi_{n_u}\}$
- video clips vectors: $C=\{\Phi_1,\Phi_2,\ldots,\Phi_{n_u}\}=\{c_1,c_2,\ldots,c_{n_c}\}$

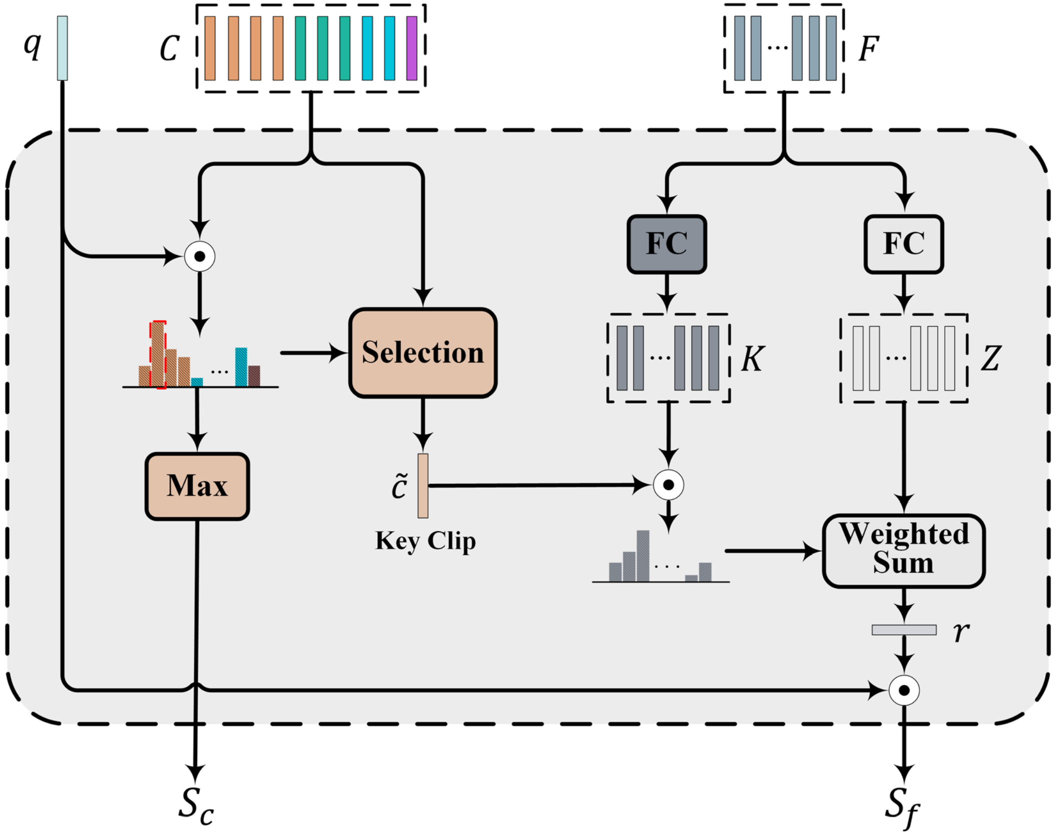
Frame-scale Video Representation

- We utilize an FC layer and a one-layer Transformer to obtain frame-scale video representation $F \in \mathbb{R}^{d \times n_v}$.



Multi-scale Similarity

- We devise a Key Clip Guided Attention to select the most important clip representation and aggregate frame features.



Clip-scale similarity:

$$S_c(v,q) = max\{cos(c_1,q), cos(c_2,q), \ldots, cos(c_{n_c},q)\}$$

Aggregated frame feature $\begin{cases} K = W_k F, Z = W_v F \\ r = softmax(\tilde{c}^T K) Z^T \end{cases}$

Frame-scale similarity:

$$S_f(v,q) = cos(r,q)$$

Similarity Learning and Model Inference

- We jointly use the **triplet ranking loss** and **InfoNCE loss** to learn the clip-scale and frame-scale similarity between video and text query.

Triplet ranking loss:

$$\mathcal{L}^{trip} = \frac{1}{n} \sum_{(q,v) \in \mathcal{B}} [max(0, m + S(q^-, v) - S(q, v)) + max(0, m + S(q, v^-) - S(q, v))],$$

InfoNCE loss:

$$\mathcal{L}^{nce} = -\frac{1}{n} \sum_{(q,v) \in \mathcal{B}} \left[log\left(\frac{S(q,v)}{S(q,v) + \sum_{q_i^- \in \mathcal{N}_q} S(q_i^-, v)} \right) + log\left(\frac{S(q,v)}{S(q,v) + \sum_{v_i^- \in \mathcal{N}_v} S(q, v_i^-)} \right) \right],$$

- After the model has been trained, the similarity between a video and a sentence query is computed as the sum of their clip-level similarity and frame-level similarity.

$$S(v,s) = \alpha S_c(v,s) + (1-\alpha) S_f(v,s)$$

Experiments

Datasets and Evaluation Metrics

- We re-purpose three datasets commonly used for VCMR, i.e., TVR, Activitynet Captions, and Charades-STA, considering their natural language queries partially relevant with the corresponding videos.

- We utilize the rank-based metrics, namely $R@K$ (K=1,5,10,100) to evaluate PRVR models. $R@K$ is the fraction of queries that correctly retrieve desired items in the top K of the ranking list.

Datasets Download:

https://github.com/HuiGuanLab/ms-sl/tree/main/dataset

Experiments

- R1: How does the proposed method perform compared with baseline methods?
- R2: How the effects of the different components in our method?
- R3: How much does our model improve the performance of VCMR methods?
- R4: How the complexity of the proposed method compared with baseline methods?

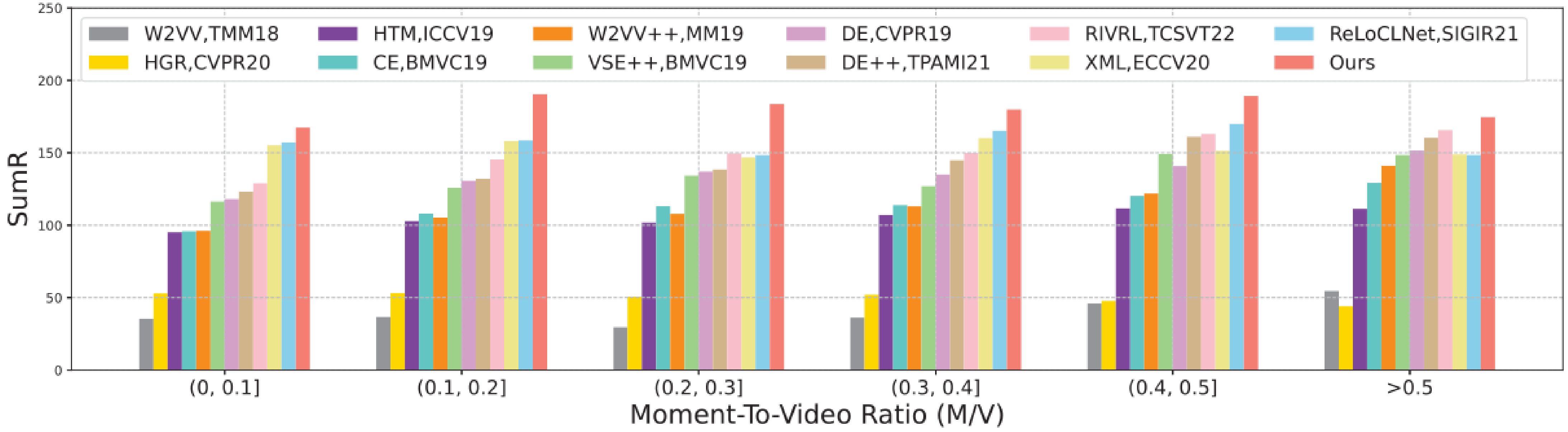
Performance comparison on TVR

Model	R@1	R@5	R@10	R100	SumR
T2VR models:					
W2VV, TMM18 [10]	2.6	5.6	7.5	20.6	36.3
HGR, CVPR20 [7]	1.7	4.9	8.3	35.2	50.1
HTM, ICCV19 [42]	3.8	12.0	19.1	63.2	98.2
CE, BMVC19 [37]	3.7	12.8	20.1	64.5	101.1
W2VV++, MM19 [31]	5.0	14.7	21.7	61.8	103.2
VSE++, BMVC19 [15]	7.5	19.9	27.7	66.0	121.1
DE, CVPR19 [11]	7.6	20.1	28.1	67.6	123.4
DE++, TPAMI21 [12]	8.8	21.9	30.2	67.4	128.3
RIVRL, TCSVT22 [13]	9.4	23.4	32.2	70.6	135.6
VCMR models w/o moment localization:					
XML, ECCV20 [29]	10.0	26.5	37.3	81.3	155.1
ReLoCLNet, SIGIR21[68]	10.7	28.1	38.1	80.3	157.1
Ours	**13.5**	**32.1**	**43.4**	**83.4**	**172.3**

- Our proposed model consistently performs the best compared with conventional T2VR models and models developed for VCMR.

Performance comparison on TVR

- Current video retrieval baseline models better address queries of larger relevance to the corresponding video while our method is less sensitive to irrelevant content in videos.



Performance comparison on Activitynet Captions and Charades-STA

- On both two datasets, our model is still at the leading position.

Model	R@1	R@5	R@10	R100	SumR
T2VR models:					
W2VV [10]	2.2	9.5	16.6	45.5	73.8
HTM [42]	3.7	13.7	22.3	66.2	105.9
HGR [7]	4.0	15.0	24.8	63.2	107.0
RIVRL [13]	5.2	18.0	28.2	66.4	117.8
VSE++ [15]	4.9	17.7	28.2	67.1	117.9
DE++ [12]	5.3	18.4	29.2	68.0	121.0
DE [11]	5.6	18.8	29.4	67.8	121.7
W2VV++ [31]	5.4	18.7	29.7	68.8	122.6
CE [37]	5.5	19.1	29.9	71.1	125.6
VCMR models w/o moment localization:					
ReLoCLNet [68]	5.7	18.9	30.0	72.0	126.6
XML [29]	5.3	19.4	30.6	73.1	128.4
Ours	**7.1**	**22.5**	**34.7**	**75.8**	**140.1**

On Activitynet Captions

Model	R@1	R@5	R@10	R100	SumR
T2VR models:					
W2VV [10]	0.5	2.9	4.7	24.5	32.6
VSE++ [15]	0.8	3.9	7.2	31.7	43.6
W2VV++ [31]	0.9	3.5	6.6	34.3	45.3
HGR [7]	1.2	3.8	7.3	33.4	45.7
CE [37]	1.3	4.5	7.3	36.0	49.1
DE [11]	1.5	5.7	9.5	36.9	53.7
DE++ [12]	1.7	5.6	9.6	37.1	54.1
RIVRL[13]	1.6	5.6	9.4	37.7	54.3
HTM [42]	1.2	5.4	9.2	44.2	60.0
VCMR models w/o moment localization:					
ReLoCLNet [68]	1.2	5.4	10.0	45.6	62.3
XML [29]	1.6	6.0	10.1	46.9	64.6
Ours	**1.8**	**7.1**	**11.8**	**47.7**	**68.4**

On Charades-STA

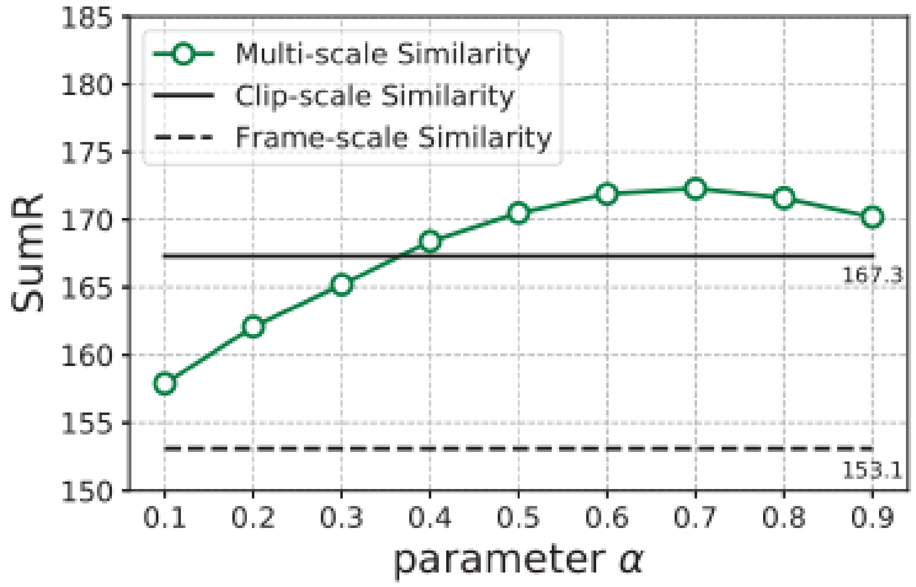
Experiments

- R1: How does the proposed method perform compared with baseline methods?
- R2: How the effects of the different components in our method?
- R3: How much does our model improve the performance of VCMR methods?
- R4: How the complexity of the proposed method compared with baseline methods?

Ablation Studies on TVR

- Removing each component from our method would result in relative performance degeneration, which shows the importance of each component.

Model	R@1	R@5	R@10	R100	SumR
Full setup	**13.5**	**32.1**	**43.4**	**83.4**	**172.4**
w/o frame-scale branch	12.3	30.5	41.5	82.3	166.6
w/o clip-scale branch	8.0	21.0	30.0	74.0	133.0
w/o key clip guide	12.2	30.6	41.0	82.4	166.3
w/o InfoNCE	11.3	29.1	40.1	81.3	161.8
w/o Triplet loss	11.2	29.2	40.4	81.9	162.6

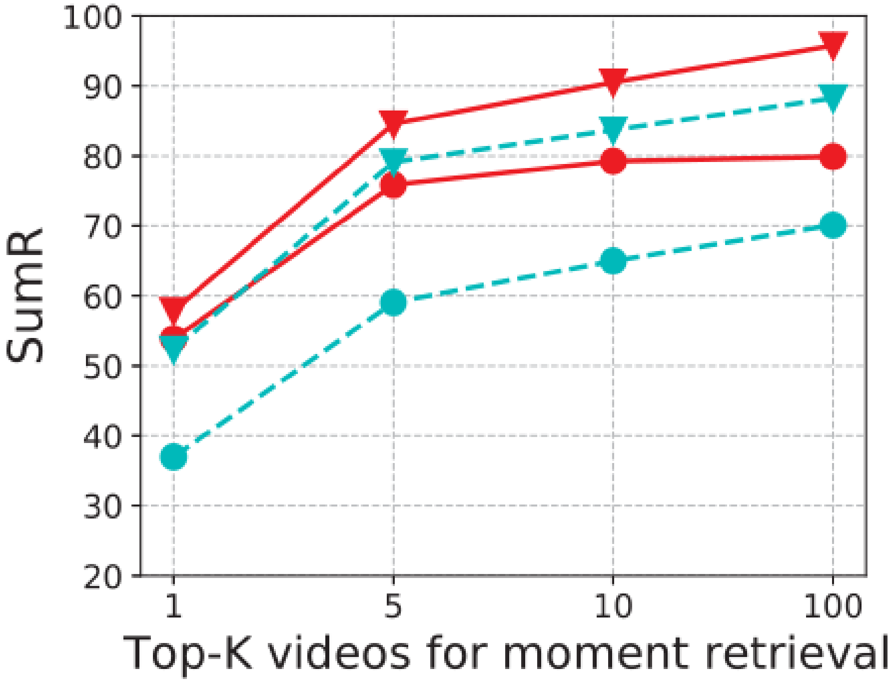


Experiments

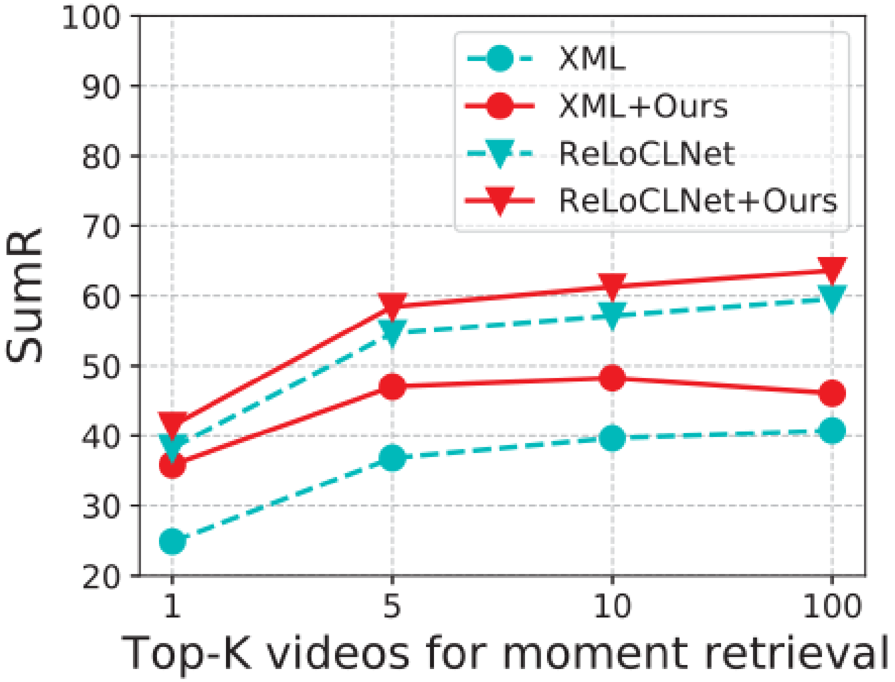
- R1: How does the proposed method perform compared with baseline methods?
- R2: How the effects of the different components in our method?
- R3: How much does our model improve the performance of VCMR methods?
- R3: How the complexity of the proposed method compared with baseline methods?

PRVR for VCMR

- We replace the first stage of two VCMR models, which brings performance improvement to both models.



(a) IoU=0.5



(b) IoU=0.7

Experiments

- R1: How does the proposed method perform compared with baseline methods?
- R2: How the effects of the different components in our method?
- R3: How much does our model improve the performance of VCMR methods?
- R4: How the complexity of the proposed method compared with baseline methods?

Comparison on Model Complexity

- In terms of FLOPs, our model is at the mid-level. In terms of memory consumption, our model requires more memory than the majority of compared models.

	W2VV	HGR	HTM	CE	W2VV++	VSE++	DE	DE++	RIVRL	XML	ReLoCLNet	Ours
FLOPs (G)	0.42	2.96	0.06	0.06	0.4	0.20	5.24	5.30	8.64	0.80	0.96	1.22
Memory (MiB)	1231	8555	1225	1435	1281	1299	5837	3515	4809	2451	2673	5349

- Retrieval efficiency: 0.2 seconds for retrieval videos from 20,000 candidate untrimmed videos.

Conclusions

- In this work, we have proposed a novel T2VR subtask termed **PRVR**. Different from the conventional T2VR where a query is usually full relevant to the corresponding video, it is typically partially relevant in PRVR.
- Towards PRVR, we have **formulated it as a MIL problem**, and propose **MS-SL** which computes the similarity on both clip scale and frame scale in a **coarse-to-fine** manner.
- Extensive experiments on three datasets have verified the effectiveness of our method for PRVR, and have shown that it can also be used for improving VCMR.

Homepage of paper: http://danieljf24.github.io/prvr/

Partially Relevant Video Retrieval

Jianfeng Dong[1] Xianke Chen[1] Minsong Zhang[1] Xun Yang[2] Shujie Chen[1] Xirong Li*[3] Xun Wang*[1]

[1]School of Computer and Information Engineering, Zhejiang Gongshang University
[2]School of Information Science and Technology, University of Science and Technology of China
[3]Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China

Paper Data Code

```
@inproceedings{ dong2022prvr,
        title = {Partially Relevant Video Retrieval},
        author = {Jianfeng Dong and Xianke Chen and Minsong Zhang and Xun Yang and Shujie Chen and Xirong Li and Xun Wang},
        booktitle = {Proceedings of the 30th ACM International Conference on Multimedia},
        year = {2022}
        }
```


E-mail: dongjf24@gmail.com a397283164@163.com