Harnessing Pre-trained Models for Ad-hoc Video Search

Fan Hu[1,2], Ruixiang Zhao[1,2], Ziyuan Wang[1,2], Jingyu Liu[1,2], Jiazhen Liu[1,2], Zhanhui Kang[3], Xirong Li[1,2]

[1]MOE Key Lab of DEKE, Renmin University of China
[2]AIMC Lab, School of Information, Renmin University of China
[3]Tencent
(TRECVID team ID: RUCMM)
https://ruc-aimc-lab.github.io/

2023.11.13

Research questions in Ad-hoc Video Search

- **How to apply multi-modal pre-training models for video search?**
- **How to apply advanced yet computationally heavy models to large-scale retrieval?**
- **How to effectively fuse the results of multiple models?**

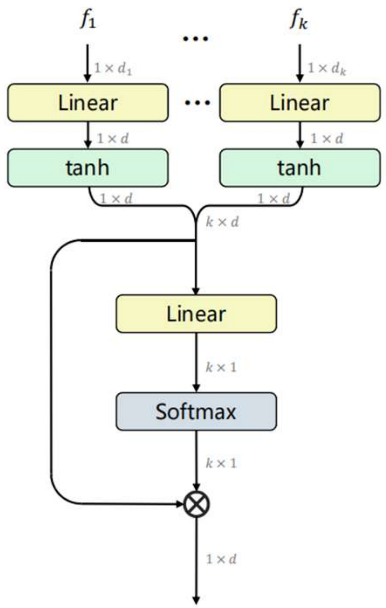
Our Solution

Feature Fusion, Multi-Grained Teaching and Learn-to-Rank late fusion

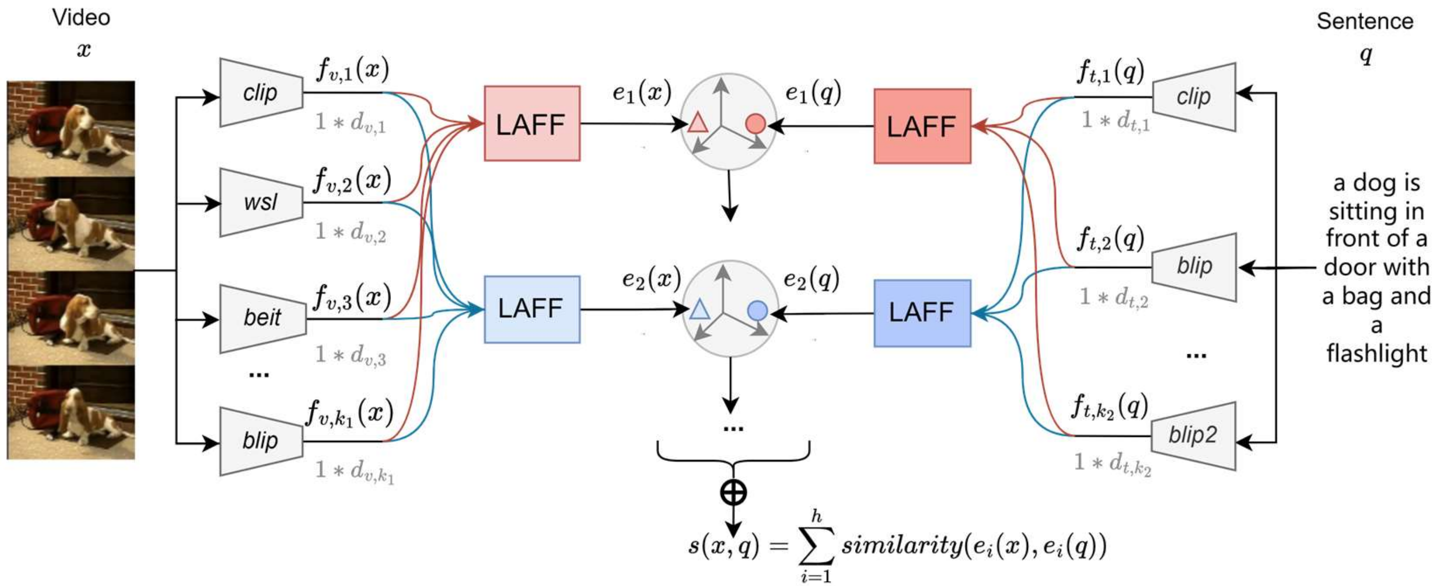
- LAFF [Hu et al., ECCV'22]
  - Focus on feature fusion of multi-modal pre-training models
- TeachCLIP [Tian et al., arXiv'23]
  - Focus on multi-grained teaching, to ensure both precision and efficiency
- Learn-to-Rank late fusion
  - Focus on weighted late fusion to boost the performance

Technique 1 LAFF based Video Retrieval

- ***LAFF***



LAFF block



The framework of LAFF based Video Retrieval

It supports feature fusion at both text and video ends to exploit diverse (off-the-shelf) features.

Video/ Text Feature

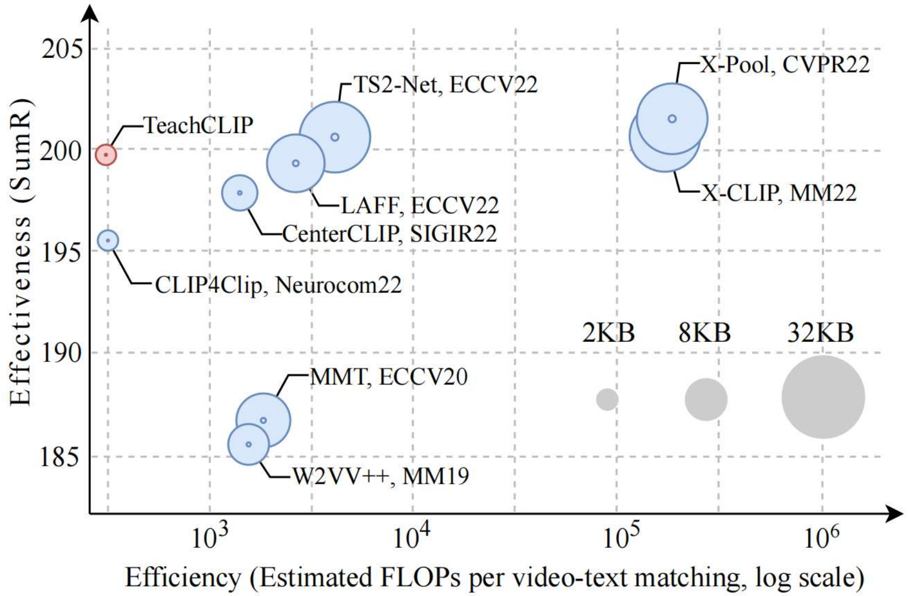
Seven video features & three text features

Video Features	Dimensionality
wsl	2048
ircsn	2048
beit	1024
clip	512
blip	256
blip2	**256**
video-llama	**768**

Text Features	Dimensionality
CLIP(ViT-L/14)@336	512
BLIP(ViT-B)	256
BLIP-2(Vit-G)	**256**

Compared to TV22, we have added two new video features and one text feature, extracted by recent multi-modal pre-training models BLIP-2 and Video-LLaMA.

Technique 2 Multi-Grained Teaching for Efficient Retrieval



Effectiveness, efficiency and video-feature storage footprint of present-day (CLIP based) text-to-video retrieval models. Dataset: MSRVTT-1k

X-CLIP and X-Pool

- recent advanced heavy **fine-grain** models.
- not suitable for **large-scale retrieval.**
- with **high precision**.

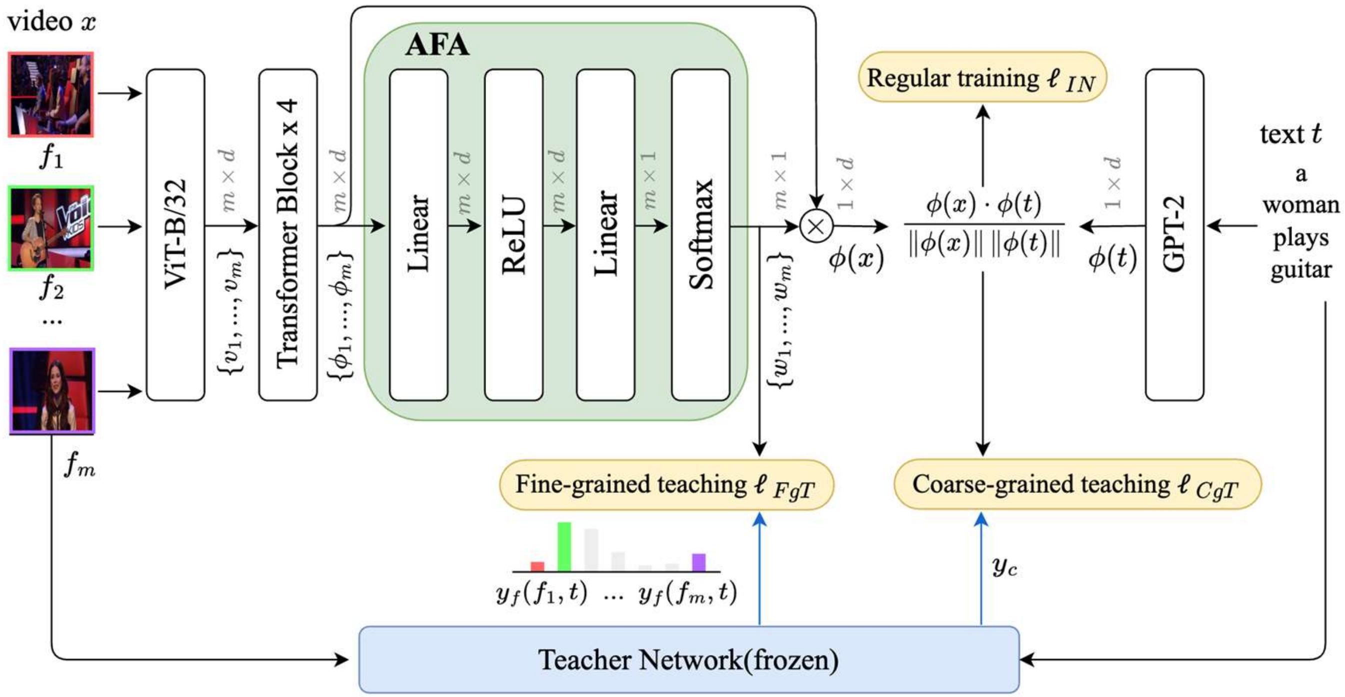
CLIP4Clip:

- coarse-grained CLIP-based model.
- efficient for **large-scale retrieval**.
- not at the same precision as X-CLIP and X-Pool.

TeachCLIP: The use of state-of-the-art teacher models for fine-grained and coarse-grained teaching ensures the precision and efficiency of student model.

Technique 2 Multi-Grained Teaching for Efficient Retrieval

- ***TeachCLIP for Ad-hoc Video Search***

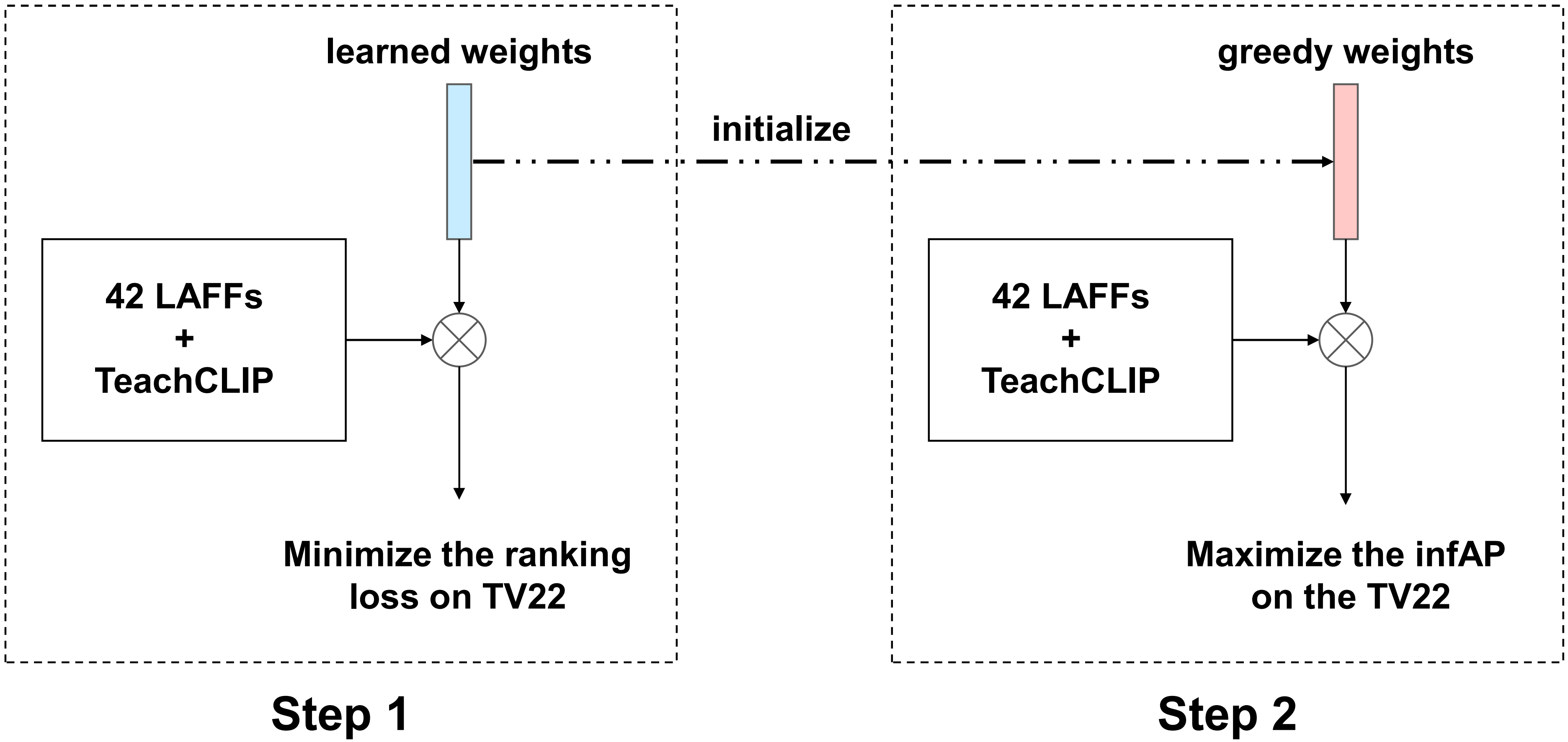


Teacher:
- X-CLIP, pretrained on MSR-VTT + TGIF + VATEX

Student:
- CLIP4Clip based network

Training datasets:
- MSR-VTT + TGIF + VATEX

Technique 3 Learn-to-Rank late fusion



Choice of (Pre-)Training Data

Three public datasets for training

Dataset	#Videos	#Sentences
MSR-VTT (CVPR2016)	10,000	200,000
TGIF (CVPR2016)	100,855	124,534
VATEX (ICCV2019)	32,239	259,909

#1 a crowd at a music festival
#2 a concert with people on the stage

Pre-training video-text datasets

Dataset	Frame/Video Num	Sentence Num
V3C1-PC	1,605,335/219,530	436,203
ChinaOpen	52,170/1,516,598	52,170
WebVid	2,291,129/44,847,987	2,291,129

Internal experiments

- ***The influence of different pre-training datasets***

  Model: LAFF with all text and video features.

Pre-training	TV19	TV20	TV21	MEAN(TV19-21)	TV22
V3C1-PC	0.255	0.345	0.352	0.317	**0.258**
WebVid	0.235	0.333	0.317	0.295	0.230
ChinaOpen	0.252	0.337	0.335	0.308	0.252
ChinaOpen, V3C1-PC	0.252	0.341	0.347	0.313	0.251
WebVid, V3C1-PC	0.256	0.349	0.351	**0.319**	0.251
ChinaOpen, WebVid, V3C1-PC	0.248	0.337	0.337	0.307	0.247

Internal experiments

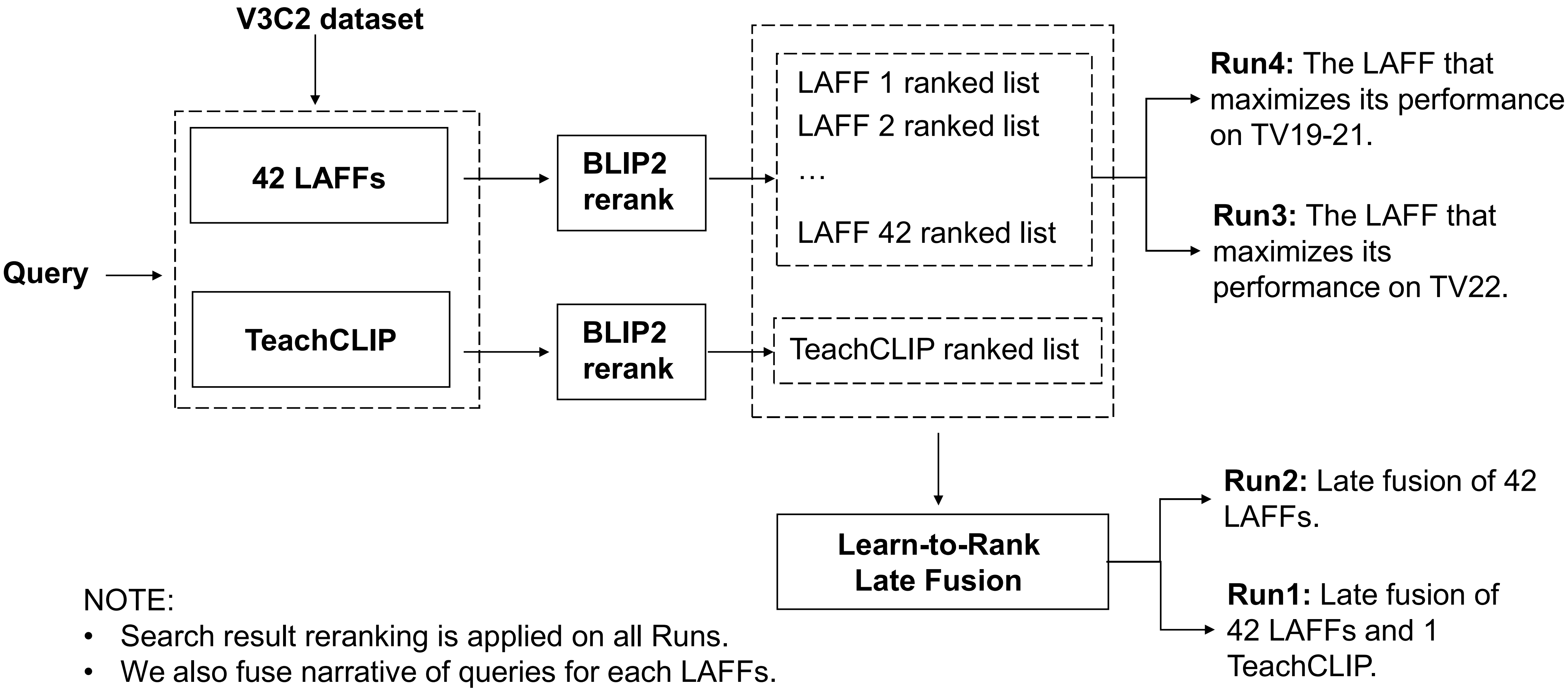
- ***TeachCLIP (TV19-21 performance)***

Pre-training	TV19	TV20	TV21	MEAN(TV19-21)	TV22
TeachCLIP	0.115	0.149	0.166	0.143	0.134

- ***Learn-to-rank Late fusion (TV22 performance)***

Fusion Models	Best Model	Step 1	Step 1+Step 2
42 LAFF	0.258	0.276	**0.279**
42 LAFF + 1 TeachCLIP	0.258	0.278	**0.282**

RUCMM Video Search Engine@TV23

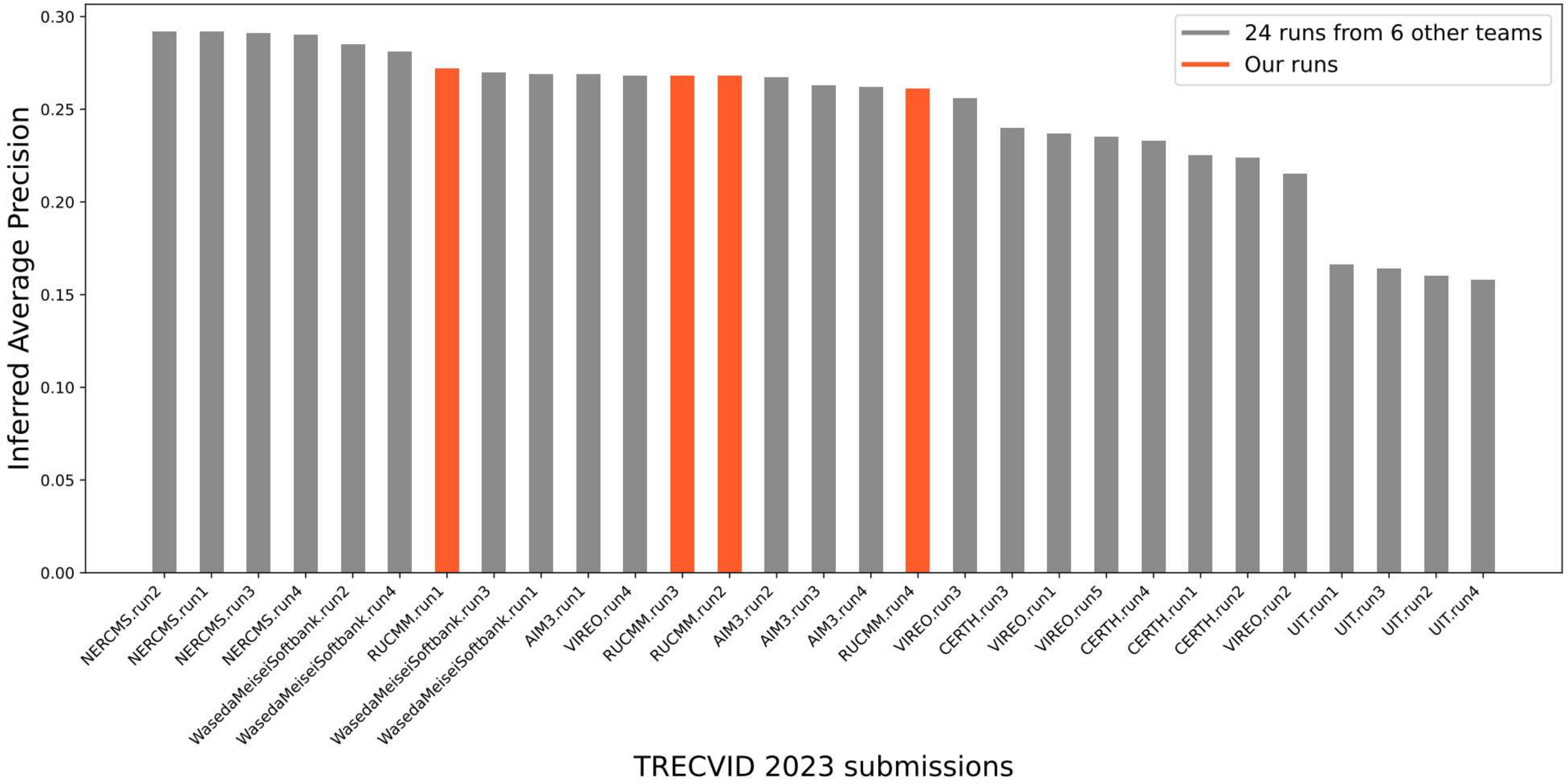


NOTE:

- Search result reranking is applied on all Runs.
- We also fuse narrative of queries for each LAFFs.

Benchmark evaluation

Our submissions ranked the 3rd



Retrospective experiments

- ***Is Learn-to-Rank Late Fusion Effective?***

Fusion Models	TV22		TV23	
	Average	Learn-to-rank	Average	Learn-to-rank
42 LAFF	0.276	**0.279**	0.249	**0.252**
42 LAFF + 1 TeachCLIP	0.277	**0.282**	0.251	**0.254**

NOTE: Search result reranking is not applied.

Learn-to-rank late fusion is more effective than average fusion.

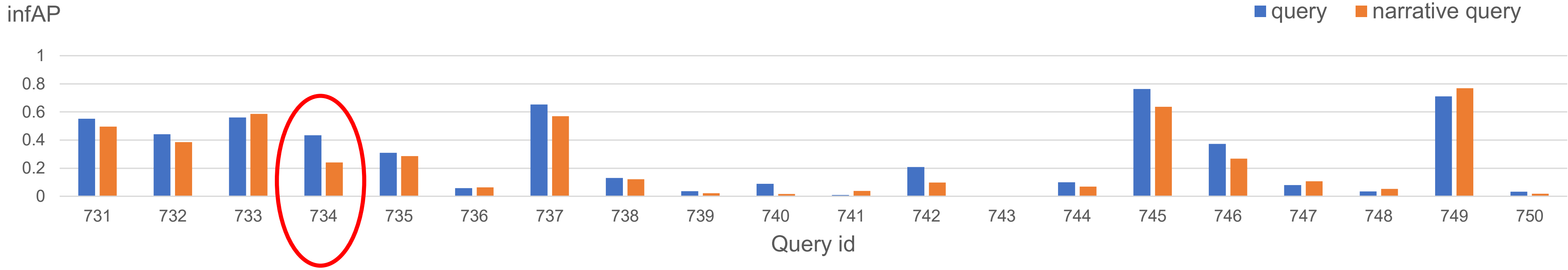
Retrospective experiments

- ***Is narrative queries Effective?***

Strategy	TV23
Query	**0.279**
Narrative	0.242
Query+Narrative	0.268

734: A **recording studio**.
734narrative: A location that can be identified as a **studio** where **recordings** can take place.

- ***For '734', narrative makes the main information unclear.***



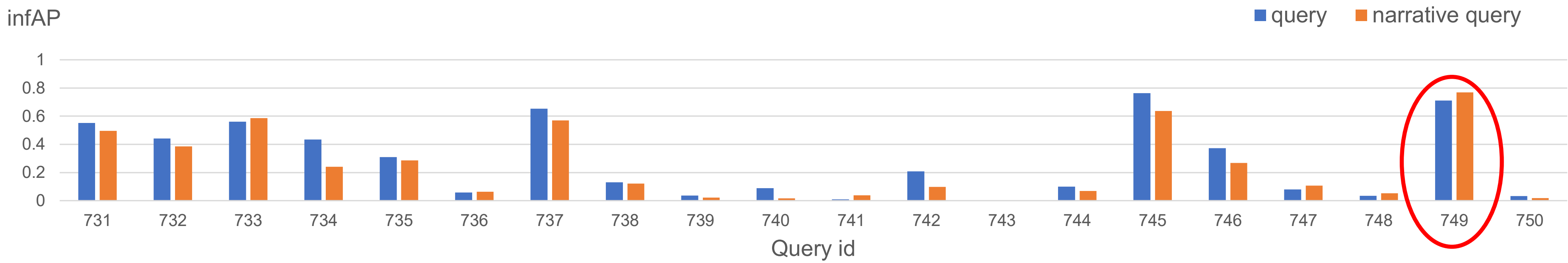
Retrospective experiments

- ***Is narrative queries Effective?***

Strategy	TV23
Query	**0.279**
Narrative	0.242
Query+Narrative	0.268

749: A person wearing any kind of **face or head mask**.
749narrative: A person is **seen** while wearing a type of **face mask or head mask**

- ***For '749', narrative provides extra information.***



Query structured understanding may be a future research direction

Conclusions

- LAFF is an effective feature fusion block for video retrieval.
- TeachCLIP may not have good standalone performance, but it can boost performance through late fusion.
- Learn-to-Rank late fusion could effectively fuse retrieval results.

https://github.com/ruc-aimc-lab

Hufan_hf@ruc.edu.cn