TRECVID 2020
Ad-hoc Video Search: Task Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble, France

George Awad
Retrieval Group, Information Access Division, Information Technology Laboratory, NIST;
Georgetown University

National Institute of Standards and Technology
U.S. Department of Commerce

Information Access Division
Information Technology Laboratory

Outline

Task Definition
Dataset
Topics (Queries)
Participating Teams
Evaluation & Results
General Observation

NIST disclaimer: Certain commercial products or company names are identified here to describe our study adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products or names identified are necessarily the best available for the purpose.

Task Definition

Goal: promote progress in content-based video retrieval based on end user **ad-hoc (generic) textual queries** that include searching for persons, objects, locations, actions and their combinations.

Task: Given a test collection, a query (***surprise/fixed (progress)***), and a master shot boundary reference, return a ranked list of at most 1000 shots (out of 1,082,657) which best satisfy the need.

Testing data: 7475 Vimeo Creative Commons Videos (V3C1), 1000 total hours with mean video durations of 8 min. Reflects a wide variety of content, style and source device. Fixed testing data since 2019.

Development data: ≈2000 hours of previous IACC.1-3 data used between 2010-2018 with concept and ad-hoc query annotations.

Vimeo Creative Commons Collection

Partition	V3C1	V3C2	V3C3	Total
File Size	2.4TB	3.0TB	3.3TB	8.7TB
Number of Videos	7'475	9'760	11'215	28'450
Combined Video Duration	1000 hours, 23 minutes, 50 seconds	1300 hours, 52 minutes, 48 seconds	1500 hours, 8 minutes, 57 seconds	3801 hours, 25 minutes, 35 seconds
Mean Video Duration	8 minutes, 2 seconds	7 minutes, 59 seconds	8 minutes, 1 seconds	8 minutes, 1 seconds
Number of Segments	1,082,659	1,425,454	1,635,580	4,143,693

The Vimeo Creative Commons Collection (V3C)* consists of '**free**' video material sourced from the web video platform **vimeo.com**. *It is designed to contain a wide range of content which is representative of what is found on the platform in general*. All videos in the collection have been released by their creators under a **Creative Commons License** which allows for unrestricted redistribution.

* Rossetto, L., Schuldt, H., Awad, G., & Butt, A. (2019). V3C – a Research Video Collection. *Proceedings of the 25th International Conference on MultiMedia Modeling*.

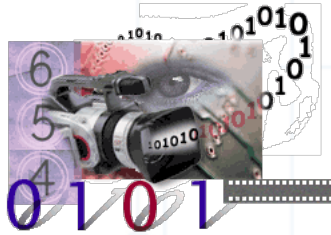
AVS 2020 (20 main) Queries by complexity

Query	Person	Action	Object	Location
Find shots of a person paddling kayak in the water	✓	✓	✓	✓
Find shots of people dancing or singing while wearing costumes outdoors	✓	✓	✓	✓
Find shots of people or cars moving on a dirt road	✓	✓	✓	✓
Find shots of one or more persons exercising in a gym	✓	✓		✓
Find shots of one or more persons standing in a body of water	✓	✓		✓
Find shots of someone jumping while snowboarding	✓	✓	✓	
Find shots of one or more people drinking wine	✓	✓	✓	
Find shots of a person wearing a necklace	✓		✓	
Find shots of a woman sitting on the floor	✓		✓	
Find shots of one or more people skydiving	✓	✓		
Find shots of a little boy smiling	✓	✓		
Find shots of group of people clapping	✓	✓		
Find shots of a woman with short hair indoors	✓			✓
Find shots of two or more people under a tree	✓			✓
Find shots showing an aerial view of buildings near water in the daytime			✓	✓
Find shots of sailboats in the water			✓	✓
Find shots of a man in blue jeans outdoors	✓		✓	✓
Find shots of a church from the inside			✓	✓
Find shots of train tracks during the daytime			✓	
Find shots of a long-haired man	✓			

2019-2021 (20 progress) Queries by complexity

Query	Person	Action	Object	Location
Find shots of a person holding an opened umbrella outdoors	✓	✓	✓	✓
Find shots of two people talking to each other inside a moving car	✓	✓	✓	✓
Find shots of people walking across (not down) a street in a city	✓	✓		✓
Find shots of a shark swimming under the water		✓	✓	✓
Find shots of a person reading a paper including newspaper	✓	✓	✓	
Find shots of fishermen fishing on a boat	✓	✓	✓	
Find shots of a person jumping with a motorcycle	✓	✓	✓	
Find shots of a person jumping with a bicycle	✓	✓	✓	
Find shots of one or more women models on a catwalk demonstrating clothes	✓	✓		
Find shots of people doing yoga	✓	✓		
Find shots of a person sleeping	✓	✓		
Find shots of people hiking	✓	✓		
Find shots of bride and groom kissing	✓	✓		
Find shots of a person skateboarding	✓	✓		
Find shots of people queuing	✓	✓		
Find shots of two people kissing who are not bride and groom	✓	✓		
Find shots of a man in a clothing store	✓			✓
Find shots of a person in a bedroom	✓			✓
Find shots of a person's shadow			✓	
Find shots showing electrical power lines			✓	

Task Parameters



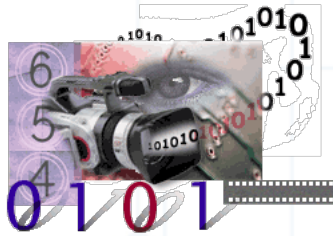
DIGITAL VIDEO RETRIEVAL at NIST

System Types	Description
Fully Automatic (F)	System uses official query directly
Manually-Assisted (M)	Query built manually
Relevance-Feedback (R)	Allow judging top-30 results up to 3 iterations

Training data categories	Description
A	Only IACC training data
D	Other training data sources
E	Only training data collected ***automatically*** using the query text
F	Only training data collected ***automatically*** using a query ***built manually*** from the official query text

→ Novelty (optional) run type was introduced to encourage retrieving non-common relevant shots easily found across systems.
→ Explainability of result items were allowed as extra optional information with the submitted shots

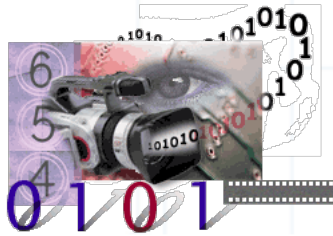
Teams – Main Task (39 runs)



Team (9 Finishers / 25)	Organization	System Type			
		M	F	R	N
VIdeoREtrievalGrOup	City University of Hong Kong	4	4		1
FIU_UM	Florida International University; University of Miami		2		1
Kindai_ogu	Kindai University; Osaka Gakuin University		1		
DVA_Researchers	Indian Institute of Space Science and Technology (IIST), Thiruvananthapuram Development and Educational Communication Unit (DECU), Indian Space Research Organisation (ISRO)		1		
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas		1		
RUC_AIM3	Renmin University of China		4		
RUCMM	Renmin University of China		4		
WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation	4	4		
ZY_BJLAB	XinHuaZhiYun Technology CO,. Ltd.	4	4		

N : Novelty runs

Teams – Progress Task (74 runs)



DIGITAL VIDEO RETRIEVAL at NIST

Team 12 Finishers	Organization	System Type			
		M	F	R	N
VIdeoREtrievalGrOup	City University of Hong Kong	6	8		
FIU_UM	Florida International University; University of Miami		6		
Kindai_ogu	Kindai University; Osaka Gakuin University		5		
SIRET (2019)*	Charles University	4			
ATL (2019)*	Alibaba group; ZheJiang University		4		
Inf (2019)*	Carnegie Mellon University; Monash University; Renmin University; Shandong University		4		
EURECOM (2019)*	EURECOM		3		
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas		1		
RUC_AIM3	Renmin University of China		4		
RUCMM	Renmin University of China		8		
WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation	8	5		
ZY_BJLAB	XinHuaZhiYun Technology CO,. Ltd.	4	4		

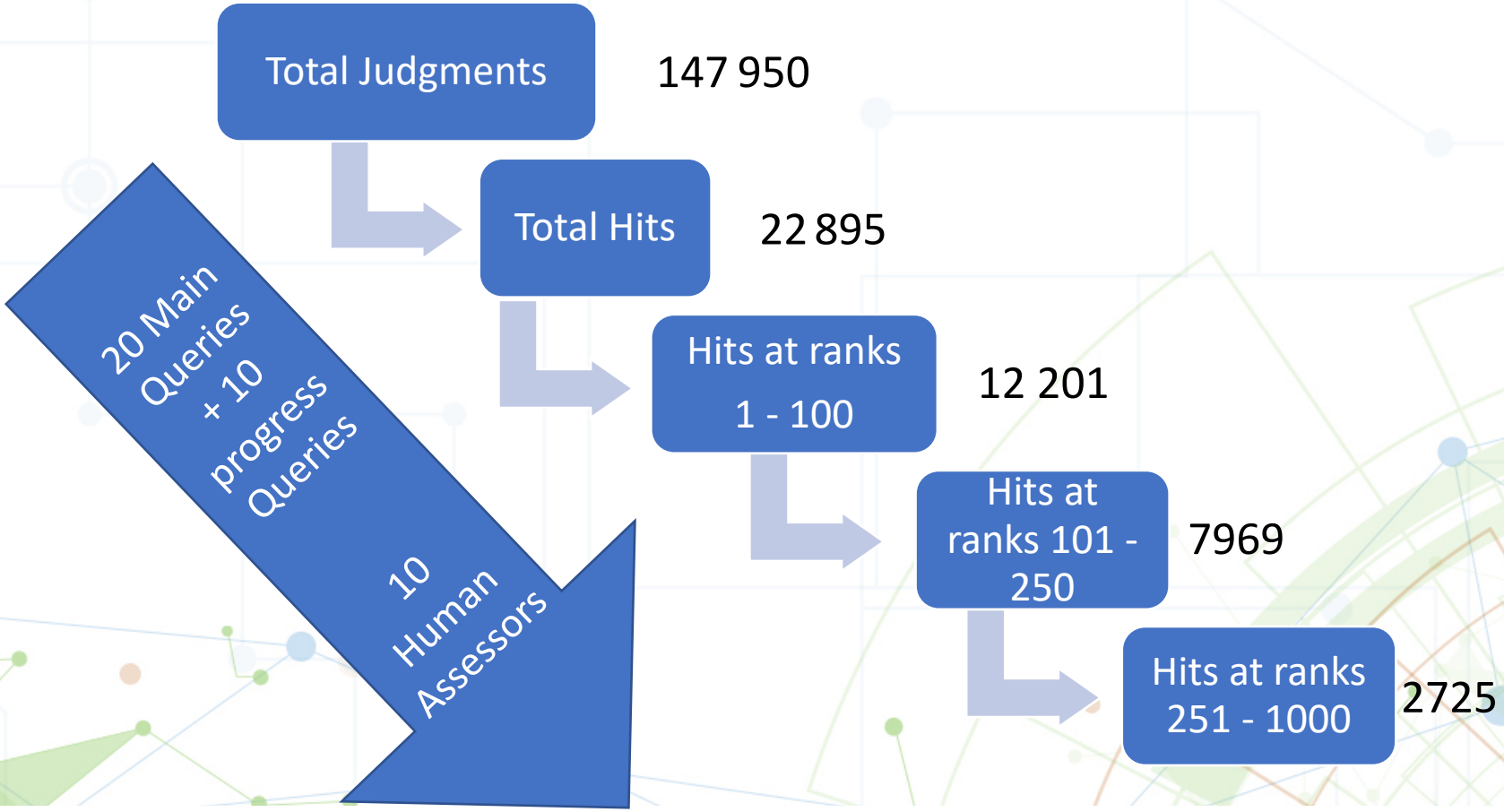
*: Teams submitted *only* progress runs in 2019

Evaluation Methodology

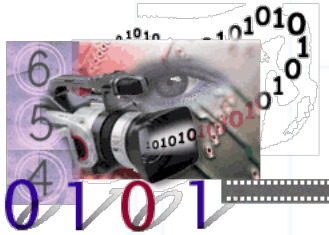
- NIST judged 100% of top (ranks 1 – 250) pooled results from all submissions and sampled (11.1%) the rest of pooled results (ranks 251 – 1000).
- Stats of sampled and judged clips from rank 251 to 1000 across all runs and topics
  - min= 10.0 %
  - max = 88.5 %
  - mean = 53.2 %
- One assessor per query, watched complete shot while listening to the audio.
- Each query assumed to be binary: absent or present for each master reference shot.
- Top submitted results were *double judged* if at least 10 runs submitted them, and assessor judged them as false positive.
- Extended inferred average precision (xinfAP) was calculated using the judged and unjudged pool by sample_eval[1] tool.
- Compared runs in terms of **mean** extended *inferred average precision* across the all evaluated queries.

[1]https://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/

Human Judgments

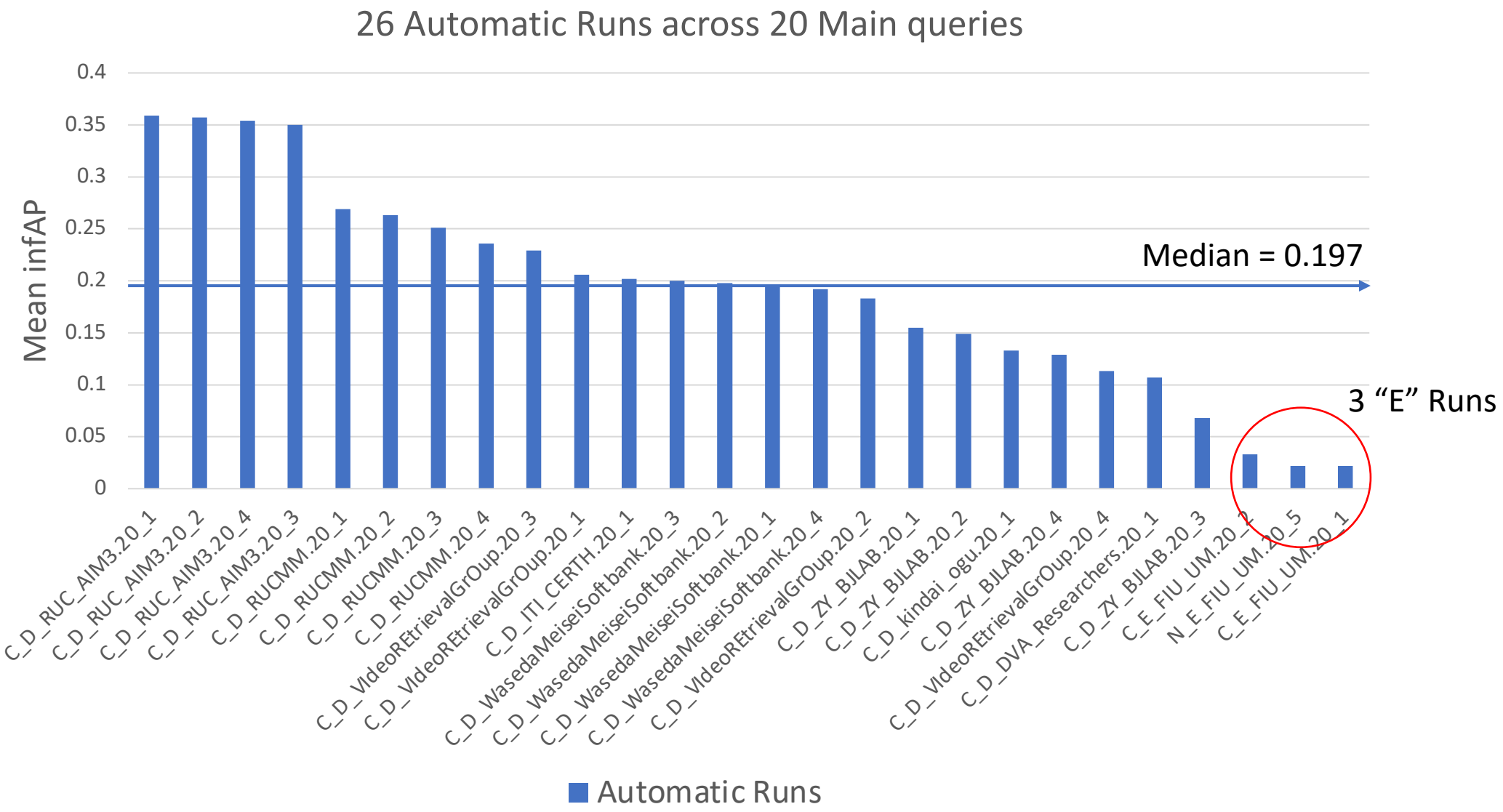


DIGITAL VIDEO RETRIEVAL at NIST

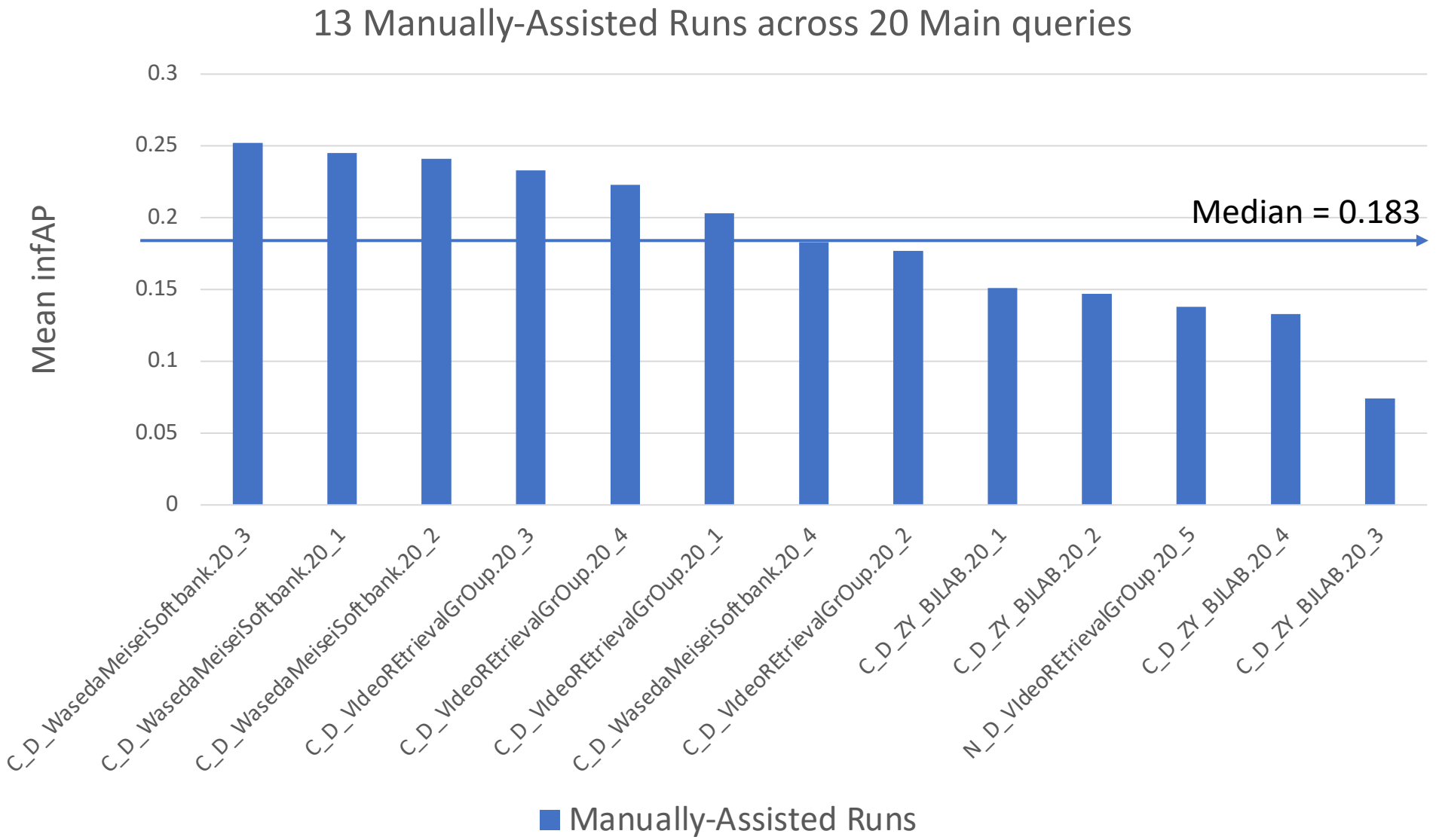


Main Task Results

Sorted Overall Scores

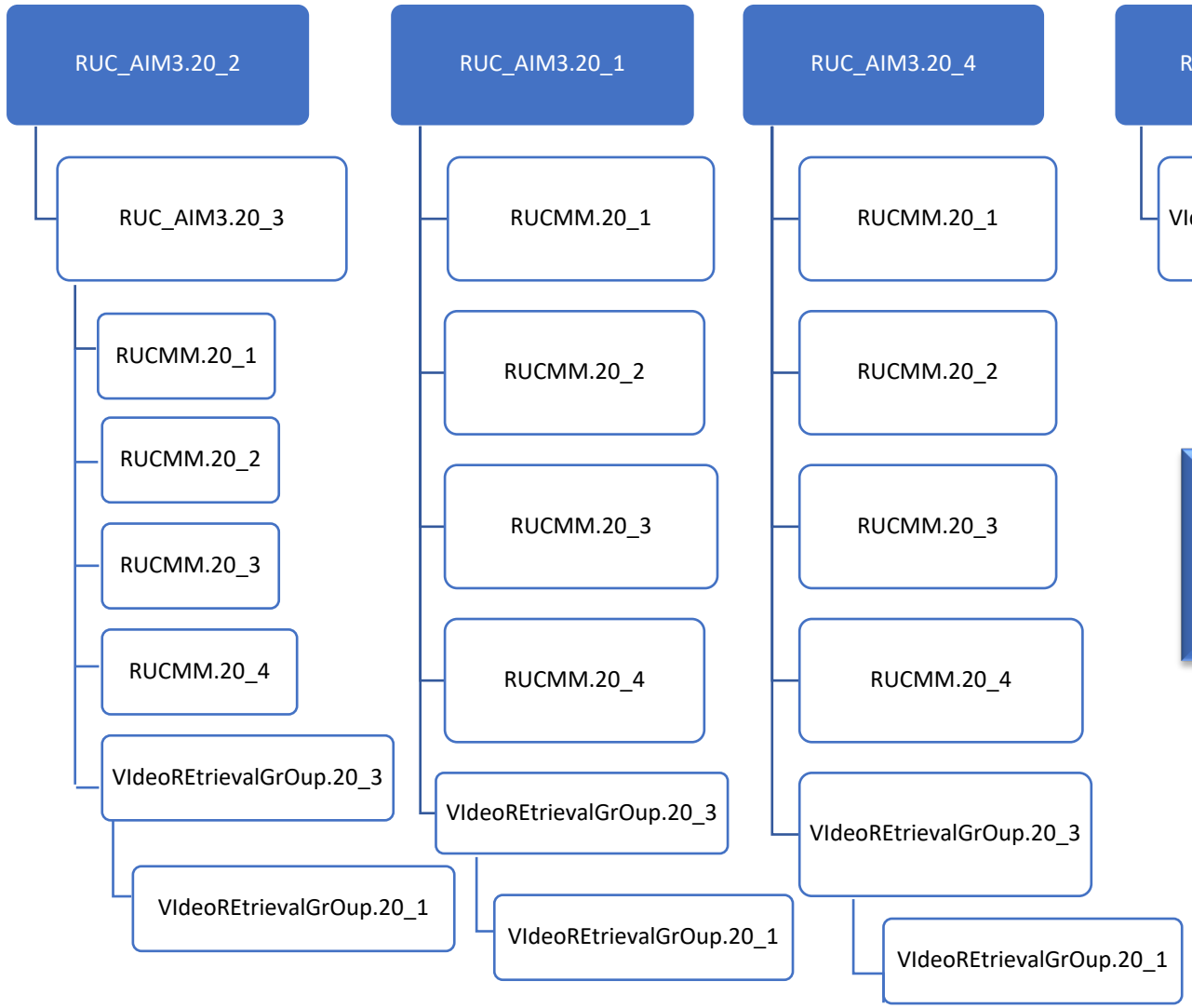


Sorted Overall Scores



Statistical Significance

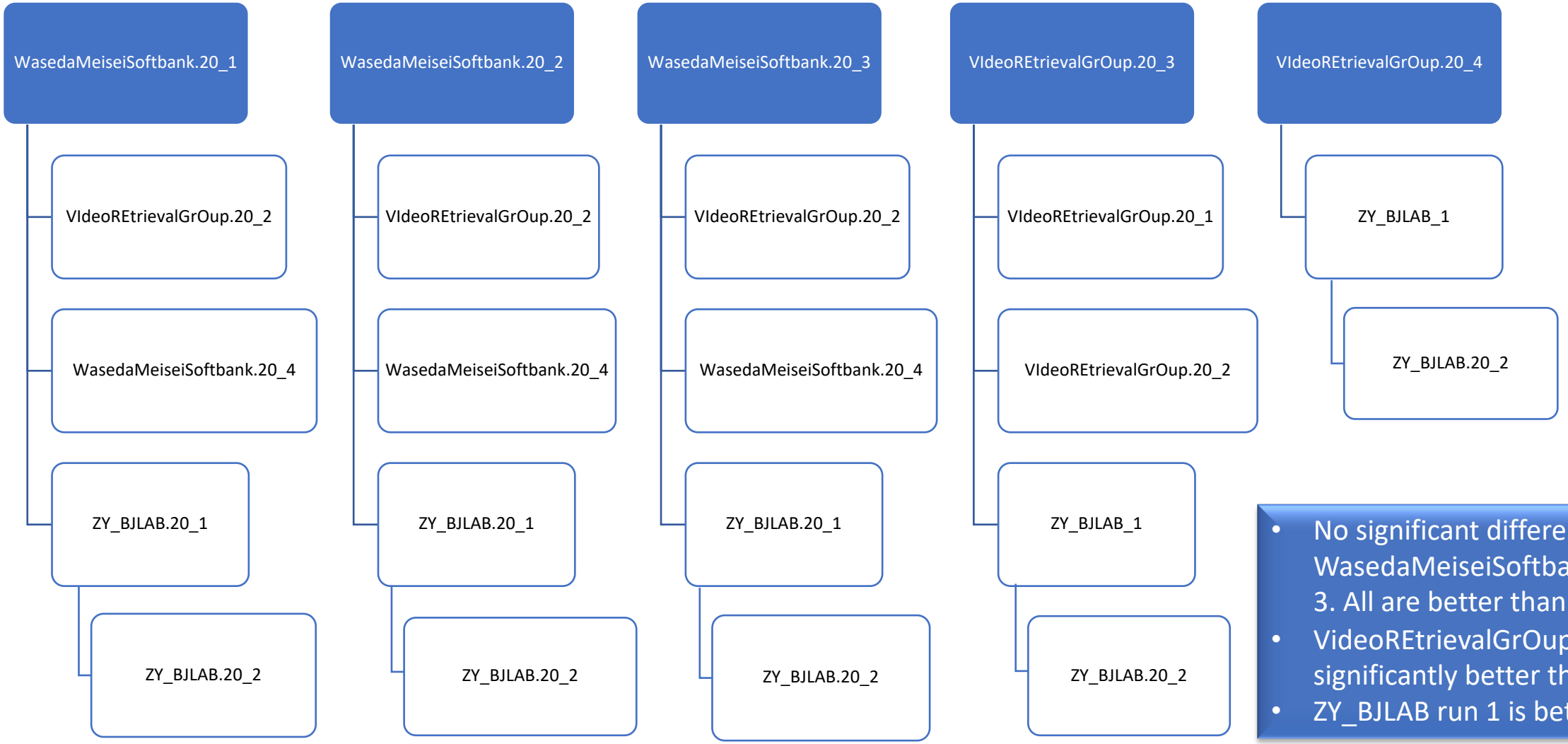
Top 10 automatic runs - randomization test ($p < 0.05$)



- No Significance Difference between RUC_AIM3 runs 1, 2 & 4.
- RUC_AIM3 runs 2 are significantly better than run 3
- No Significance Difference between RUCMM runs 1, 2, 3 & 4.
- VideoREtrievalGrOup run 3 is significantly better than run 1

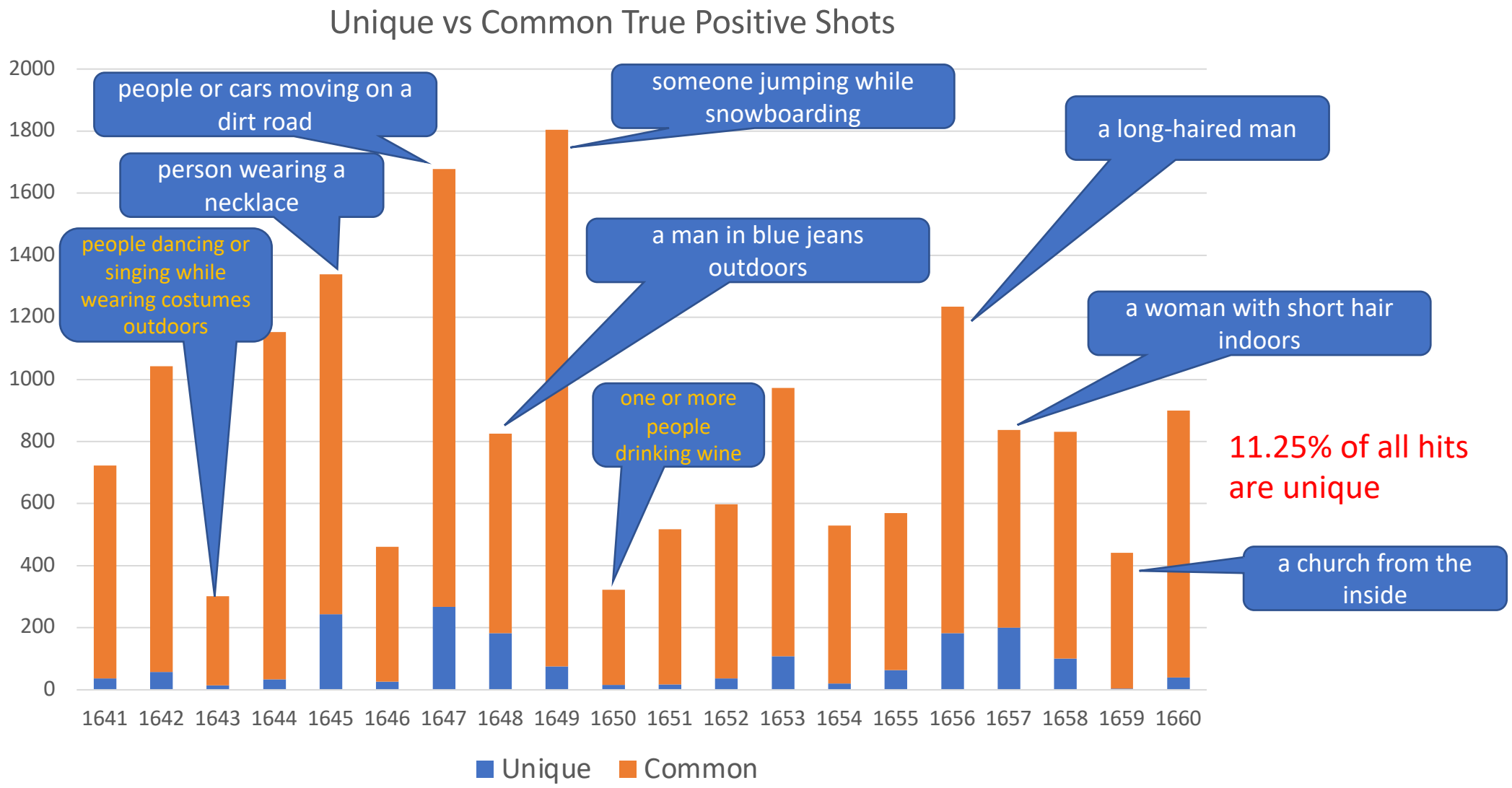
Statistical Significance

Top 10 manually-assisted runs - randomization test ($p < 0.05$)

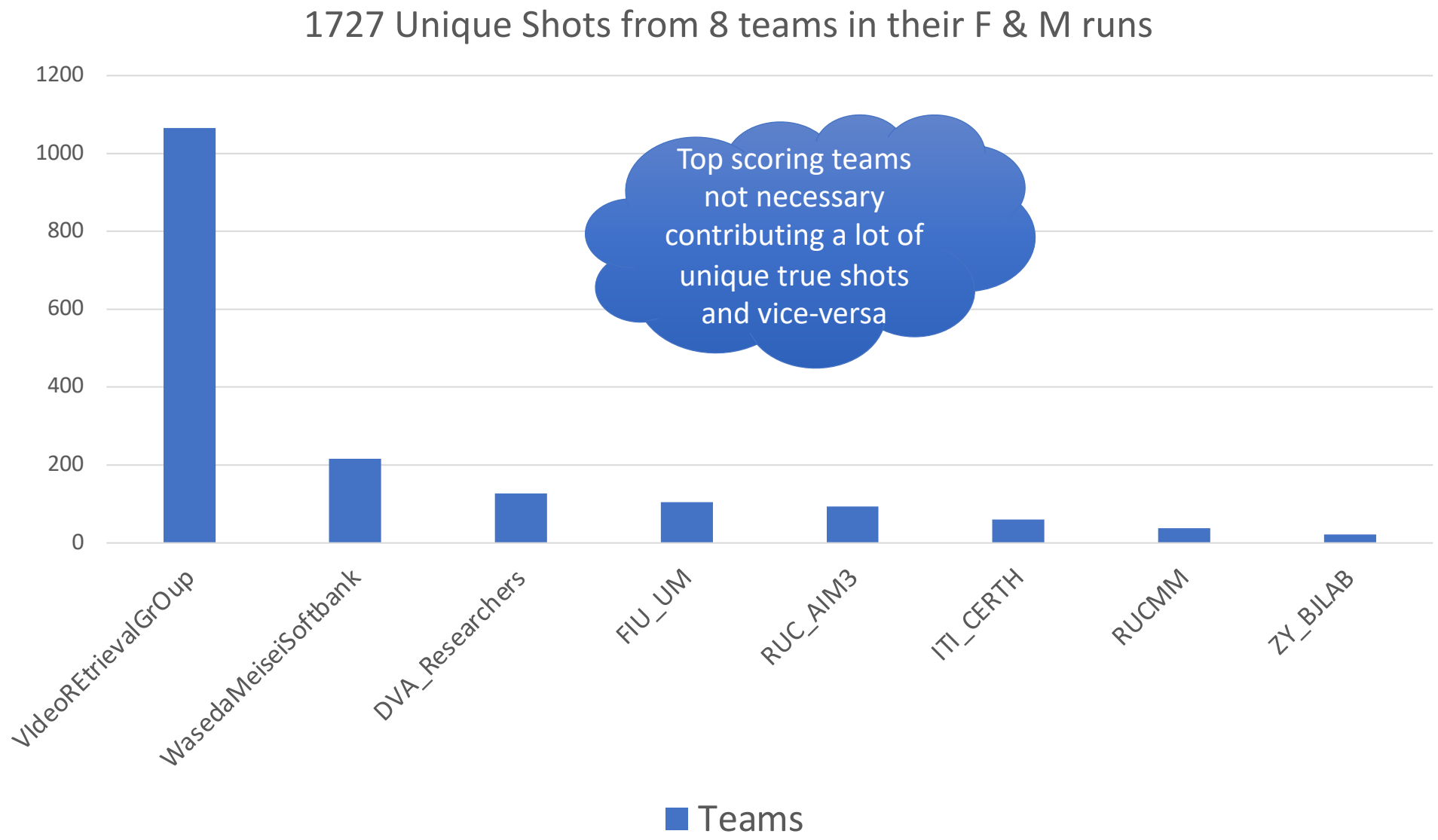


- No significant difference between WasedaMeiseiSoftbank runs 1, 2 and 3. All are better than run 4
- VideoRetrievalGrOup run 3 is significantly better than runs 1 and 2.
- ZY_BJLAB run 1 is better than run 2

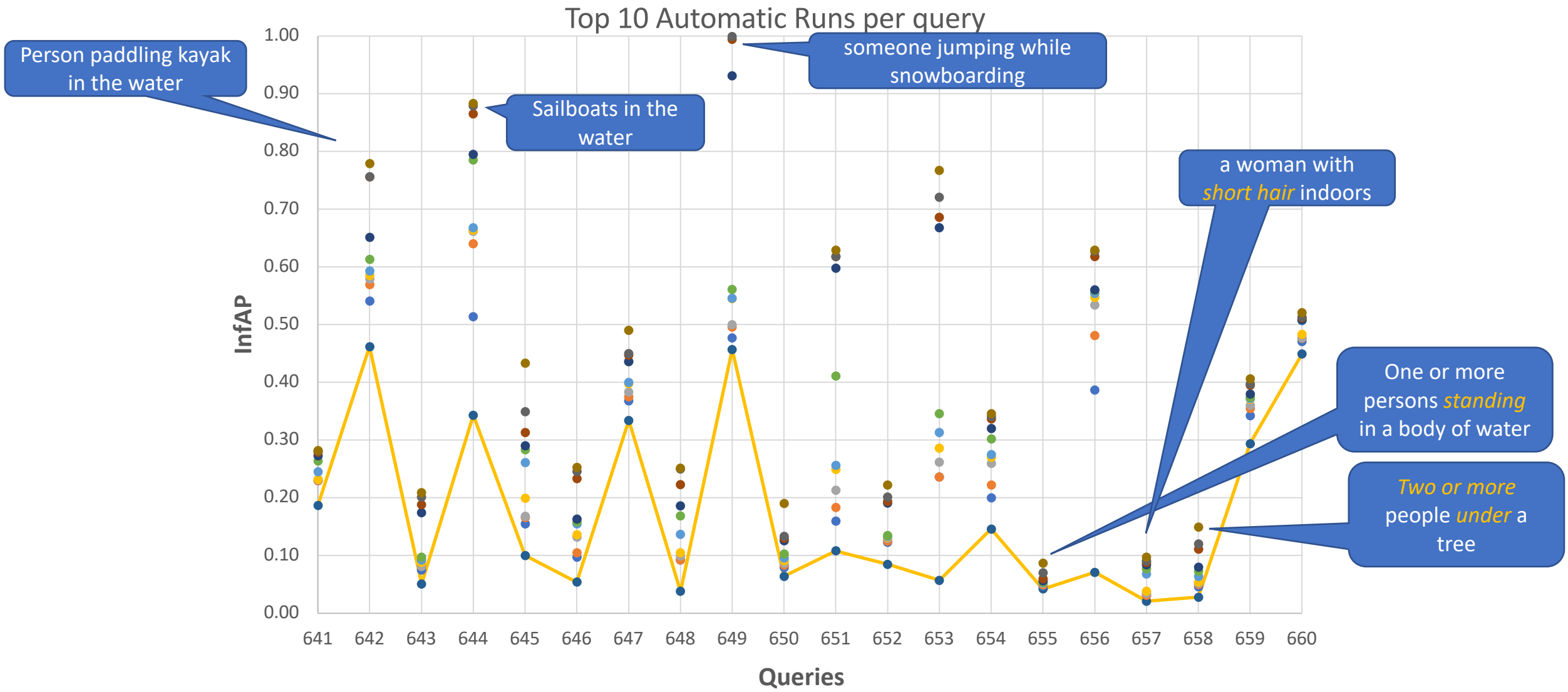
Hits Per Topic (Main Task)



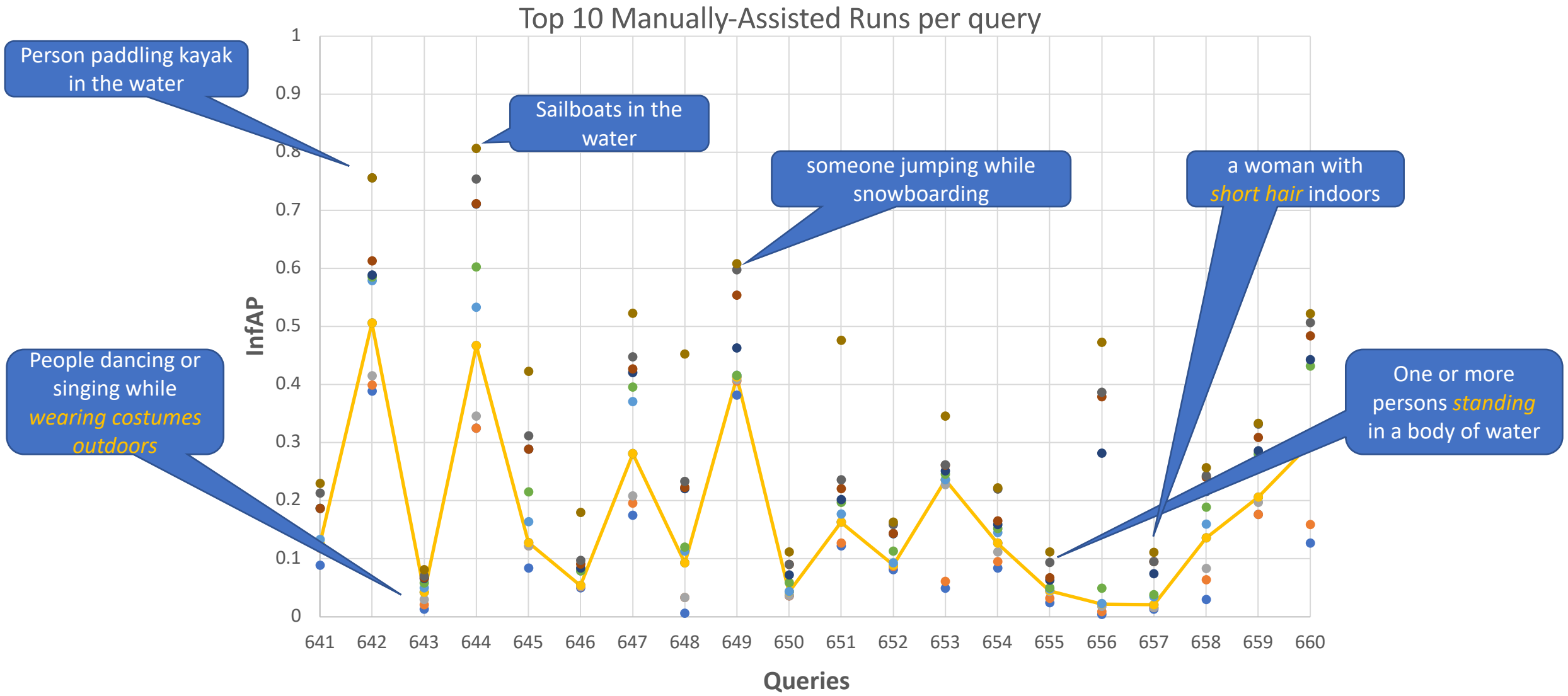
Sorted Unique Hits by Team



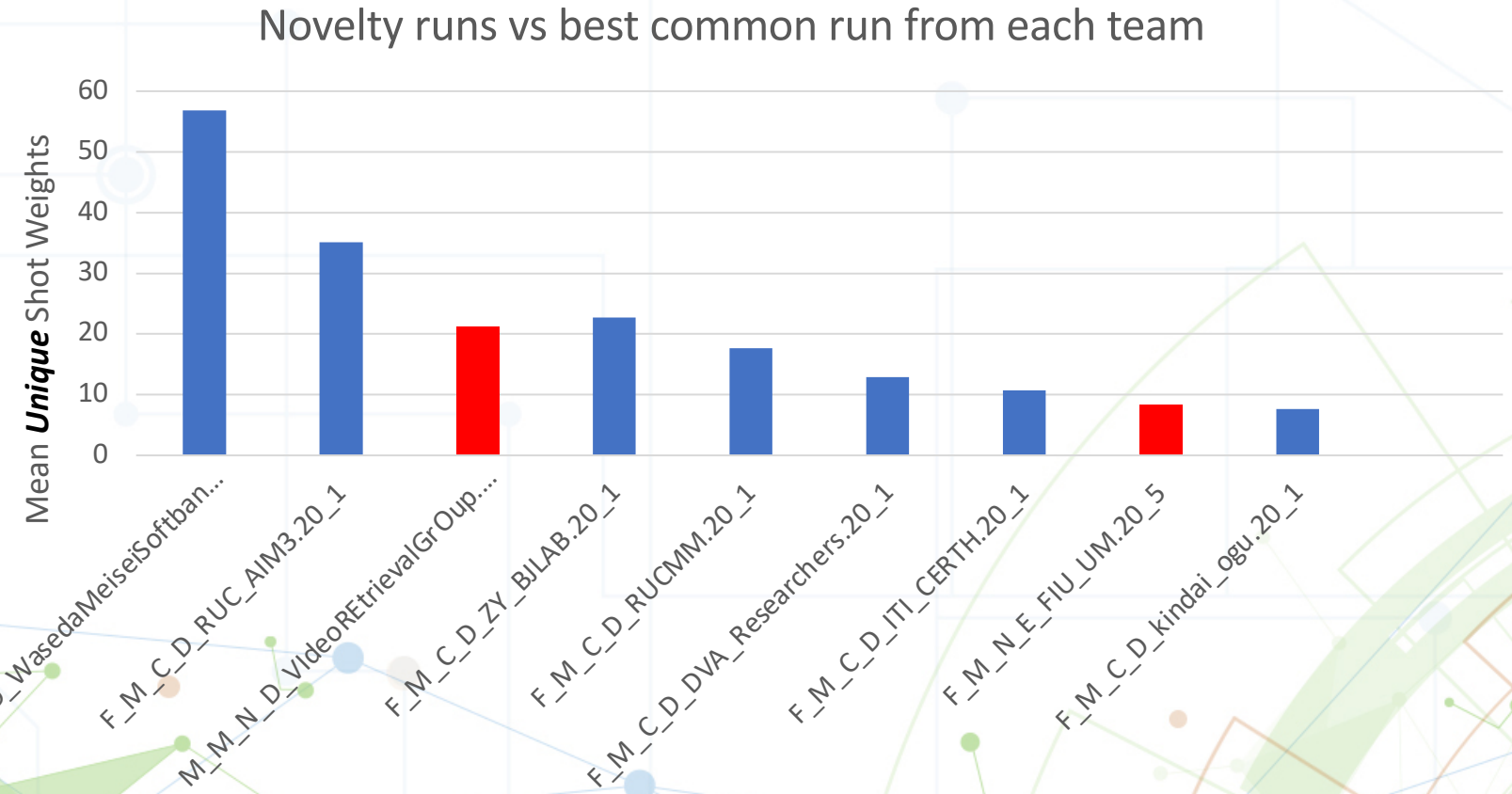
Top runs per query (Main Task)



Top runs per query (Main Task)



Novelty Scores

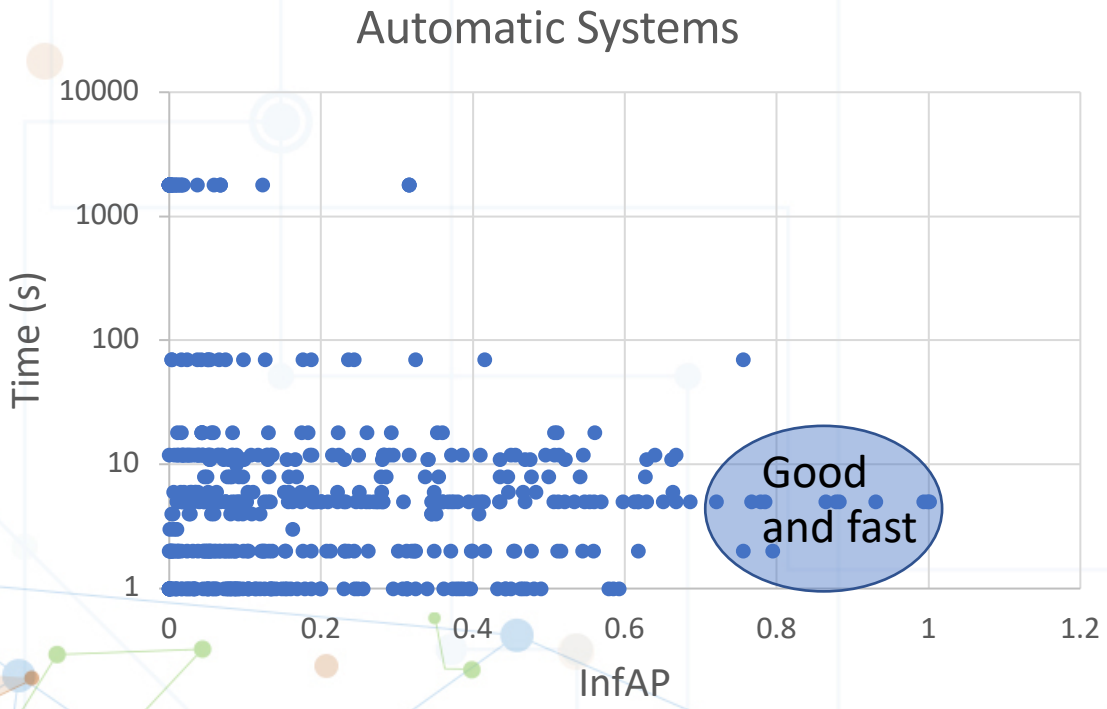


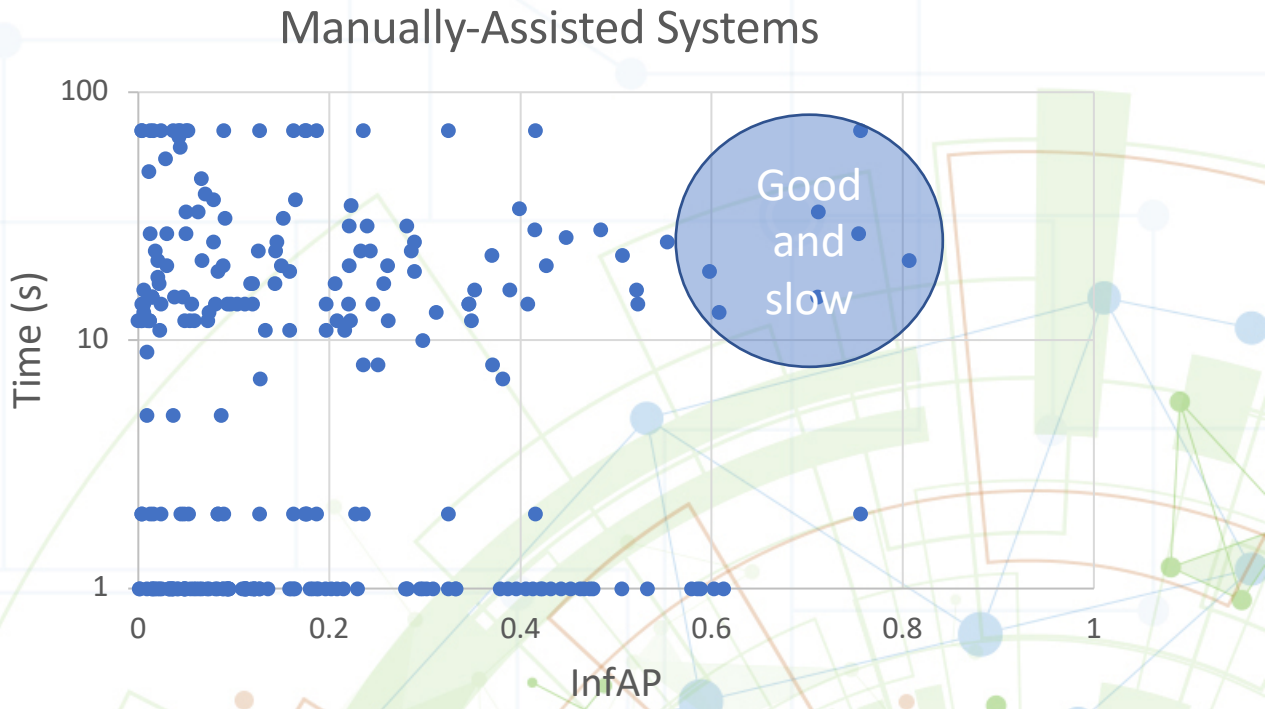
Novelty runs

Common runs

A weight is given to each topic and shot pairs in the ground truth such that highest weight is given to unique shots.

Efficiency

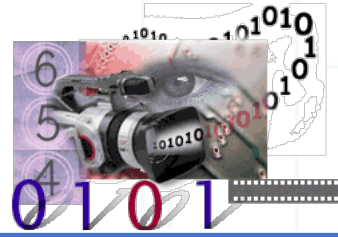




DIGITAL VIDEO RETRIEVAL at NIST

National Institute of Standards and Technology
U.S. Department of Commerce

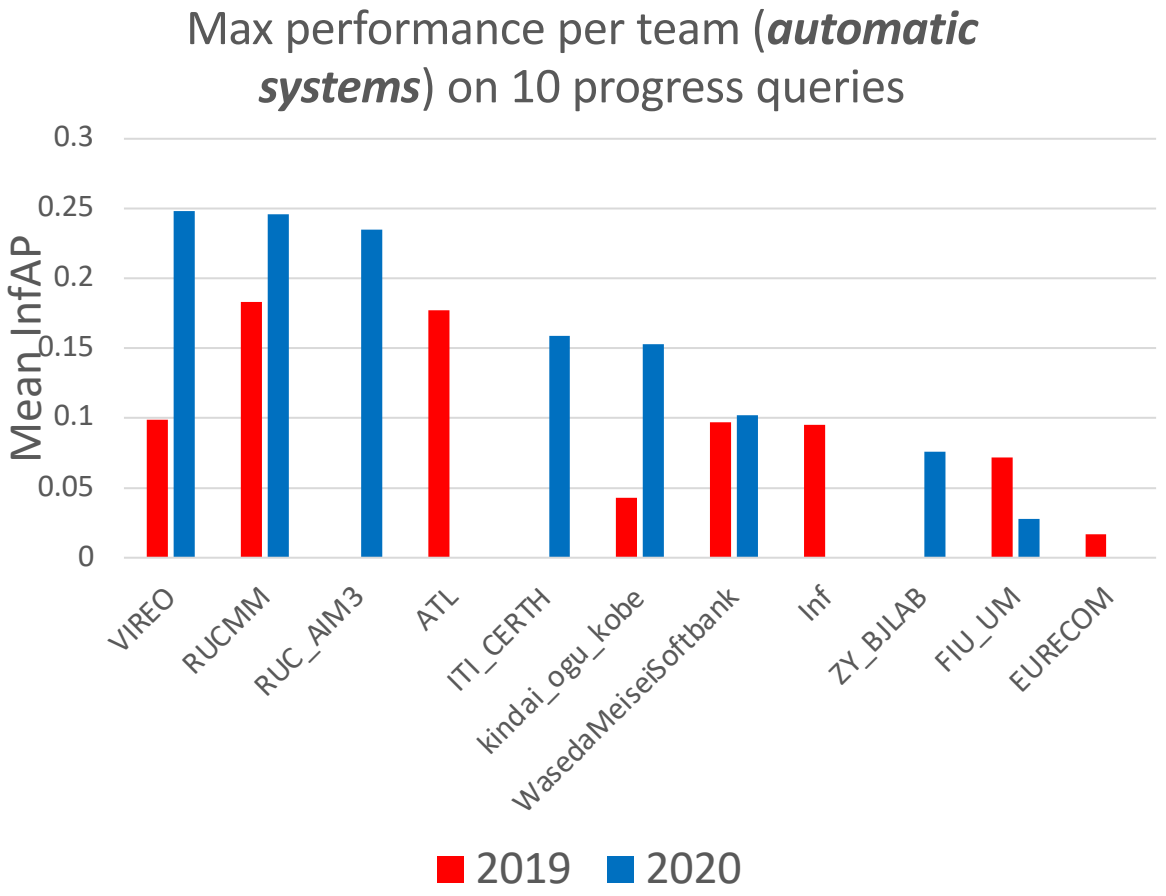
Progress Task

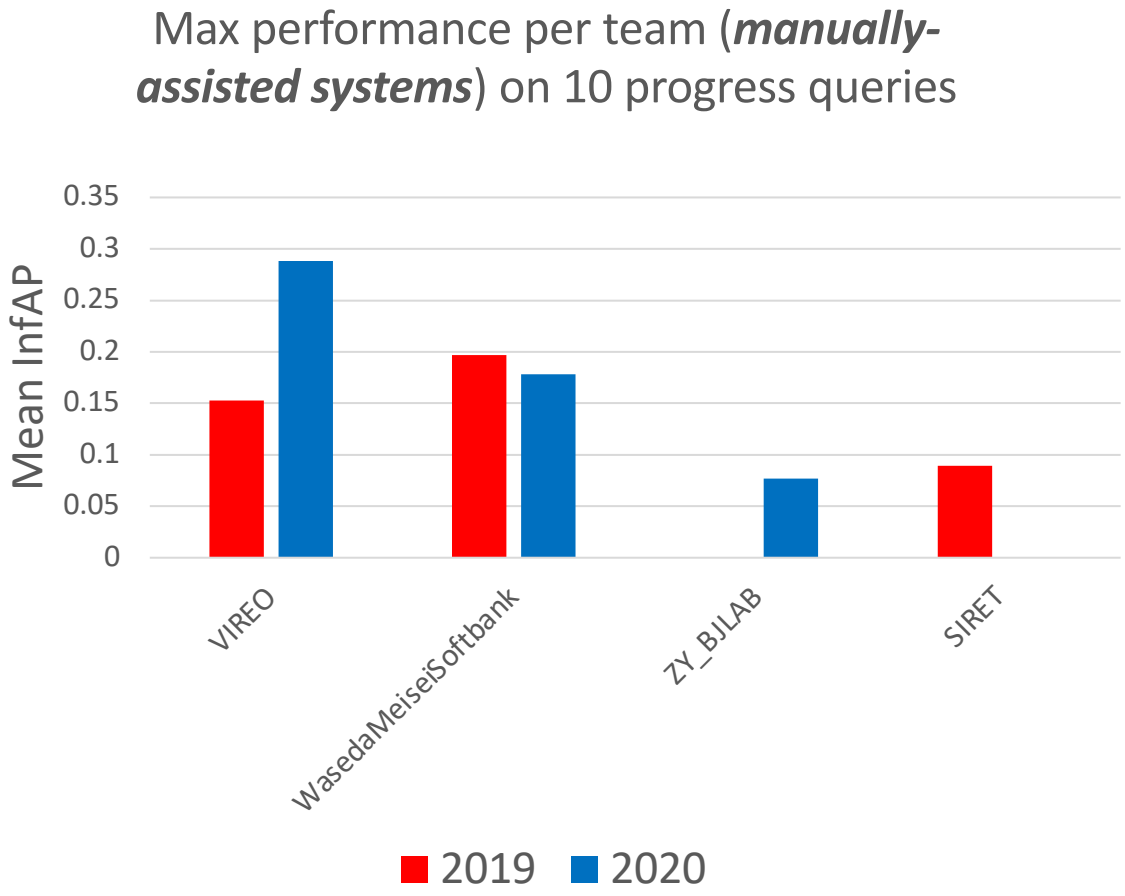


		Evaluation year		
		2019	2020	2021
Submission year	2019	***Systems:*** Submit 20 fixed progress queries		
	2020		***Systems:*** Submit 20 fixed progress queries ***NIST:*** Eval 10 queries (set A)	
	2021			***Systems:*** Submit 20 fixed progress queries ***NIST:*** Eval 10 queries (set B)

Goals : Evaluate 10 (set A) common queries submitted in 2 years (2019, 2020)
Evaluate 10 (set B) common queries submitted in 3 years (2019, 2020, 2021)
Evaluate 20 common queries submitted in 3 years (2019 , 2020, 2021)
Ground truth for 20 common queries can be released only in 2021

Progress subtask results (2019-2020)





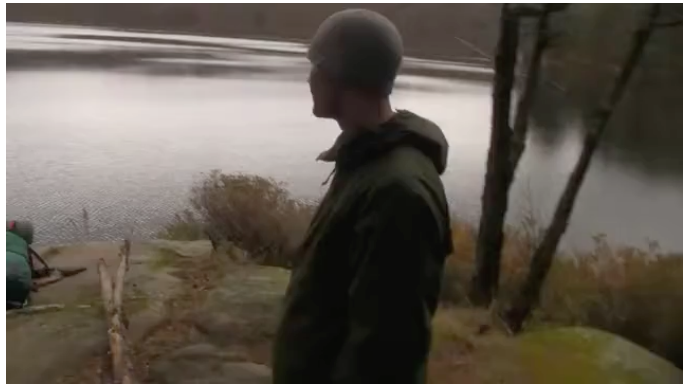
Samples of (tricky/failed) results

Find shots showing an aerial view of buildings near water in the daytime

Find shots of people dancing or singing while wearing costumes outdoors

Find shots of a woman sitting on the floor

Find shots of one or more people skydiving



Find shots of one or more persons standing in a body of water

Find shots of one or more persons exercising in a gym

All images are from the V3C1 dataset (Creative Commons Videos)

Easy vs Hard Queries

Query	Rank of easy queries (infAP >= 0.5)	Rank of hard queries (infAP < 0.5)	Person	Action	Object	Location
Find shots of a person paddling kayak in the water	1 ✓		✓	✓	✓	✓
Find shots of people dancing or singing while wearing costumes outdoors		1 ✗	✓	✓	✓	✓
Find shots of people or cars moving on a dirt road		13 ✗	✓	✓	✓	✓
Find shots of one or more persons exercising in a gym		7 ✗	✓	✓		✓
Find shots of one or more persons standing in a body of water		11 ✗	✓	✓		✓
Find shots of someone jumping while snowboarding	3 ✓		✓	✓	✓	
Find shots of one or more people drinking wine		10 ✗	✓	✓	✓	
Find shots of a person wearing a necklace		4 ✗	✓		✓	
Find shots of a woman sitting on the floor		9 ✗	✓		✓	
Find shots of one or more people skydiving	6 ✓		✓	✓		
Find shots of a little boy smiling		5 ✗	✓	✓		
Find shots of group of people clapping	7 ✓		✓	✓		
Find shots of a woman with short hair indoors		8 ✗	✓			✓
Find shots of two or more people under a tree		12 ✗	✓			✓
Find shots showing an aerial view of buildings near water in the daytime		6 ✗			✓	✓
Find shots of sailboats in the water	2 ✓				✓	✓
Find shots of a man in blue jeans outdoors		2 ✗	✓		✓	✓
Find shots of a church from the inside		3 ✗			✓	✓
Find shots of train tracks during the daytime	5 ✓				✓	
Find shots of a long-haired man	4 ✓		✓			

Easy ✓

Hard ✗

2020 Main Approaches

- Still “concept-based” and “concept-free”(visual-textual embedding spaces) approaches but clear trend toward the latter
- Clear advantage for “embedding space” approaches, especially for fully automatic search and even overall
- Concept bank often used as a complement
- Training data for semantic spaces: MSR and TRECVid VTT tasks, TGIF, IACC.3, Flickr8k, Flickr30k, MS COCO, Conceptual Captions, VATEX …
  $\rightarrow$ Arms race?

2020 Main Approaches

- Renmin University of China "RUC_AIM3" (presentation to follow):
  - Fully automatic (0.359): two-branch framework with global (VSE++) and fine-grain matching with Hierarchical Graph Reasoning (HGR)
- Renmin University of China "RUCMM" (presentation to follow):
  - Fully automatic (0.269): "dual encoding network" with Word to Visual Word (W2VV++) and BERT as text encoder similar to TRECVid 2019 plus Sentence Encoder Assembly (SEA) by multi-space multi-loss learning
- City University of Hong Kong (presentation to follow):
  - Fully automatic (0.229): dual-task model learns feature embedding and concept decoding simultaneously
  - Manually assisted (0233): same with user screening the concept list and removing unrelated or unspecific concepts

2020 Main Approaches

- Waseda University; Meisei University; SoftBank Corporation:
  - Fully automatic (0.200): visual-semantic embedding (VSE++)
  - Manually assisted (0.252): concept-based retrieval similar to previous years' concept bank approach and fusion with VSE
- Centre for Research and Technology Hellas:
  - Fully automatic (0.202): attention-based cross-modal deep network inspired by the dual encoding approach
- State Key Laboratory of Media Convergence Production Technology and Systems (ZY_BJLAB):
  - Fully automatic (0.202): search video retrieval using multi-modal video representations from collaborative experts.

2020 Task Observations

- **2nd year on AVS using V3C1 dataset** (sub-collection from a bigger V3C dataset).
- Continued the planned 2019-2021 progress subtask.
- 9 teams finished the main task and 12 (8+4) teams finished the progress task.
- 26 automatic systems and 13 manually-assisted systems submitted runs in the main task.
- 74 total systems (22 manually-assisted and 52 automatic) are submitted for the 2019-2020 progress subtask.
- **Run training types are dominated by "D" (non IACC.3 training data) runs. Only 3 "E" (no-annotation) runs and no "R" (relevance-feedback) systems submitted.**
- **No teams submitted explainability results with their runs!**
- Only 2 Novelty systems submitted. Common systems performed higher on the novelty metric.
- **Majority of 2020 systems performed higher than their 2019 systems in the progress subtask**
- Few automatic systems are good and fast, while few manually-assisted systems are good and slow.
- There is high similarity between automatic and manually-assisted in terms of query performance relatively to each other.
- Among high scoring topics, there is more room for improvement among systems.
- Among low scoring topics, most systems scores are collapsed in small narrow range.
- Absolute number of hits are comparable to 2019. Overall performance are higher than 2019 (same dataset, different queries)
- **Top scoring teams didn't necessarily report unique relevant shots (thus they are good in ranking relevant shots).**
- Hard queries are the ones asked for unusual combinations of facets (compared to well-known concepts)
- **Task is still challenging!**

Interactive Video Retrieval

During the Video Browser Showdown (VBS)

At MMM 2021
27th International Conference on Multimedia Modeling,
June 22-24, 2021 Prague, Czech Republic

- 10 Ad-Hoc Video Search (AVS) topics : Each AVS topic has several/many target shots that should be found.
- 10 Known-Item Search (KIS) tasks, which are selected completely random on site. Each KIS task has only one single 20 s long target segment.
- Registration for the task is now closed

Images borrowed from https://videobrowsershowdown.org/pics/

Agenda

EST Time

7:30 – 7:50 AM	• RUC_AIM3 • *RUC_AIM3 at TRECVID 2020: Ad-hoc Video Search*
7:50 – 8:10 AM	• RUCMM • *Sentence Encoder Assembly for Ad-hoc Video Search*
8:10 – 8:30 AM	• VIdeoREtrievalGrOup • *Concept versus Embedding search*
8:30 - 9:00 AM	• Break
9:00 - 9:20 AM	• AVS Task Discussion