TRECVID 2023
Ad-hoc Video Search (AVS) Task Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble, France

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

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

Information Access Division
Information Technology Laboratory

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept 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 entities, materials, or equipment are necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST, or the U.S. Government.

Outline

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

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, and a master shot boundary reference, return a ranked list of at most 1000 shots (out of 1, 425,454) which best satisfy the query.

Queries:

- Main : New 20 to 30 queries each year
- Progress : A set of fixed 20 queries for 3 years

Testing data: ***9760*** Vimeo Creative Commons Videos (V3C2), 1300 total hours with mean video durations of 8 min. Reflects a wide variety of content, style and source devices.

Development data:

- ≈2000 hours of previous IACC.1-3 (Internet Archive) data used between 2010-2018 with concept and ad-hoc query annotations.
- V3C1 (Vimeo Creative Commons) dataset, 1000 hours, with ad-hoc query annotations (used between 2019 – 2021).

Task Parameters

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

Training data categories	Description
A	Only V3C1 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 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

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 2023 (20 main) Queries by complexity

A Person, Location, or Object

A man is seen with a baby
A woman with red hair
A golf course
A recording studio
A toy vehicle

Person + Action + Object

A person taking picture using a cell phone camera
A person wearing gloves while biking
A man riding a scooter

Person + Object + Location

A person wearing any kind of face or head mask
A man with an earring in his left ear
A red or blue scarf around someone's neck

At least two persons are working on their laptops together in the same room indoors
A man carrying a bag on one of his shoulders

Person + Action

A person opens a door and enters a location

Person + Object

A woman wearing (dark framed) glasses
A police officer wearing a helmet

Person + Location

Two or more persons are seen in front of a chain link fence
A heavy man indoors

Person + Action + Location

A child climbs an object outdoors
A man is talking in a small window located in the lower corner of the screen

2022-2024 (20 progress) Queries by complexity

A Person, Location, or Object

A woman with a ponytail
A person's Hands with a red nail polish
A building with balconies seen from the outside during daytime
A room with a wood floor
A wooden bridge
A round table

Person + Object

A man wearing a lanyard around his neck

Person + Location

A man is seen at a gas station

Person + Object + Location

A person wearing a ring in their nose
A man wearing a dark colored hooded jacket outdoors

Person + Action

A person is throwing an object away
A person is washing oneself or another thing

Object + Location

A vehicle driving under a tunnel
A big building that is being camera panned or tilted from the outside

Person + Action + Location

A person is lying on the ground outdoors
A person is rubbing part of their face using their hands

Person + Action + Object

A man holding a gun but not shooting
A person is pouring liquid into a type of container

Person + Action + Object + Location

A man holding a fishing rod while being dipped in a body of water
A person holding a long stick which is not a drum stick outdoors

Evaluated in 2023

Teams – Main Task (43 runs)

Team Name (7 Finishers)	Organization	System Type		
		Manually Assisted	Fully Automatic	Relevance Feedback
VIREO	Singapore Management University; City University of Hong Kong	6	5	
NII_UIT	National Institute of Informatics; University of Information Technology	4	4	
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas		4	
RUC_AIM3	Renmin University of China		4	
RUCMM	Renmin University of China		4	
WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation		4	
WHU_NERCMS	Wuhan University		4	4

Teams – Progress Task (58 runs)

Team Name (7 Finishers)	Organization	System Type		
		Manually Assisted	Fully Automatic	Relevance Feedback
VIREO	Singapore Management University; City University of Hong Kong	5(22), 5(23)	5(22), 5(23)	
Kindai_ogu_osaka	Kindai University; Osaka Gakuin University; Osaka University		4 (22)	
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas		2(22), 4(23)	
RUCAIM3-Tencent	Renmin University of China		4(22)	
RUCMM	Renmin University of China		4(22), 4(23)	
WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation		4(22), 4(23)	
NII_UIT	National Institute of Informatics; University of Information Technology	6(23)	2(23)	

** Runs(year)

Evaluation Methodology

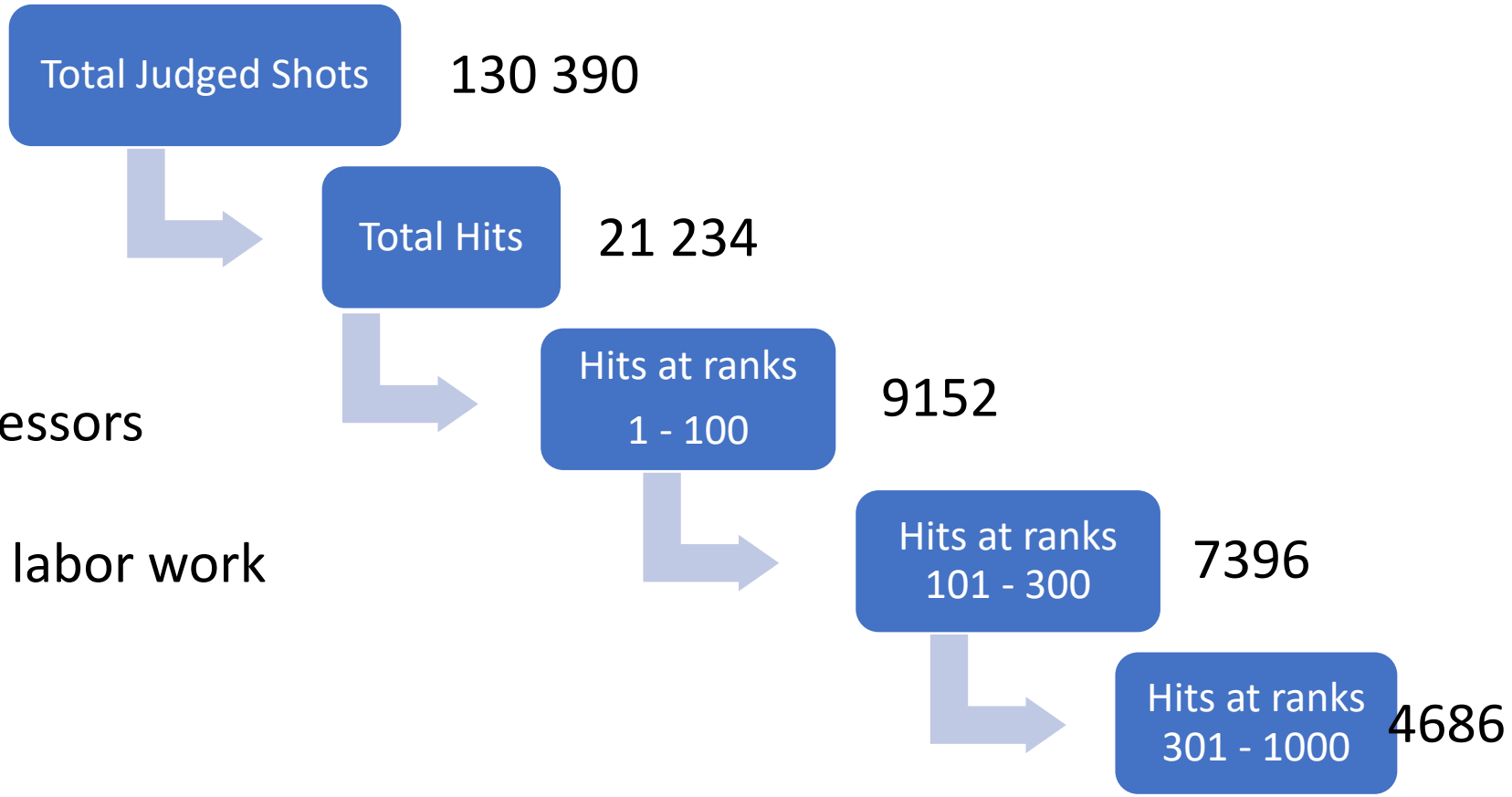
- NIST judged 100% of top (ranks 1 – 300) pooled results from all submissions and sampled 25% from the rest of pooled results (ranks 301 – 1000).
- Stats of sampled and judged clips (ranks 301 to 1000) across all runs and topics
  - At minimum, 22.5 % of any run and query results were sampled and judged
  - At maximum, 86 % of any run and query results were sampled and judged
  - On average, 63 % of any run and query results were sampled and judged
- 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.

Evaluation Methodology

- Top submitted results were *double judged* if at least 10 runs submitted them, and assessor judged them as false positive.
- submitted results were *double judged* if keyframes of close neighbourhood (+/- 5) shots are visually similar but judged differently.
- Extended inferred average precision (xinfAP[1]) was calculated using the judged and unjudged pool by sample_eval[2] tool.
- Compared runs in terms of **mean** extended *inferred average precision* across all evaluated queries.

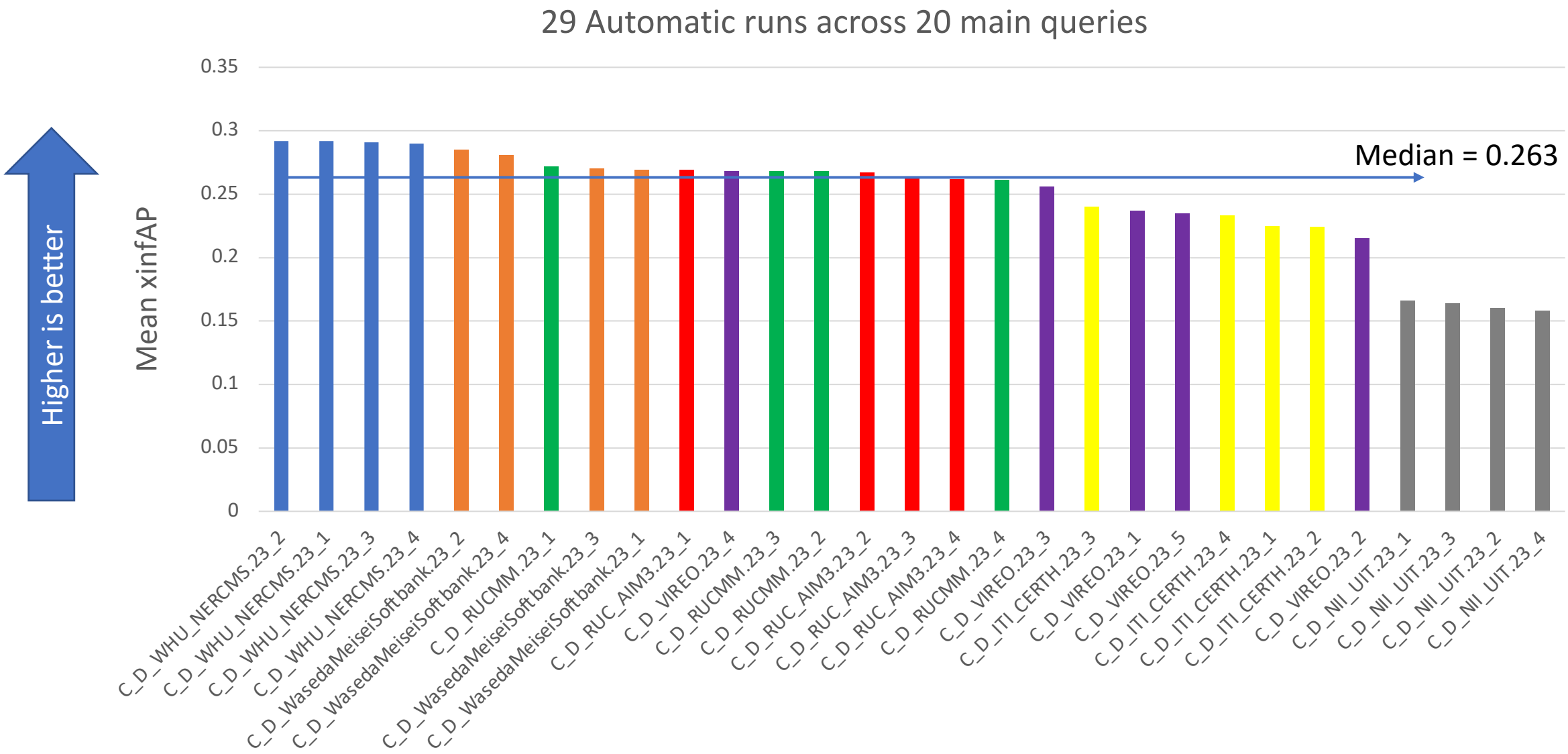
[1] J.A. Aslam, V. Pavlu and E. Yilmaz, Statistical Method for System Evaluation Using Incomplete Judgments Proceedings of the 29th ACM SIGIR Conference, Seattle, 2006.
[2] https://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/

Human Judgments



5 human assessors
30 queries
500 hours of labor work

Sorted Overall Scores – Automatic Runs



Statistical Significance (top F runs)

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

- F_M_C_D_WasedaMeiseiSoftbank.23_4
  - F_M_C_D_WasedaMeiseiSoftbank.23_1
  - F_M_C_D_WasedaMeiseiSoftbank.23_3
- F_M_C_D_WasedaMeiseiSoftbank.23_2
  - F_M_C_D_WasedaMeiseiSoftbank.23_1
  - F_M_C_D_WasedaMeiseiSoftbank.23_3
- F_M_C_D_WHU_NERCMS.23_1
  - F_M_C_D_RUC_AIM3.23_1

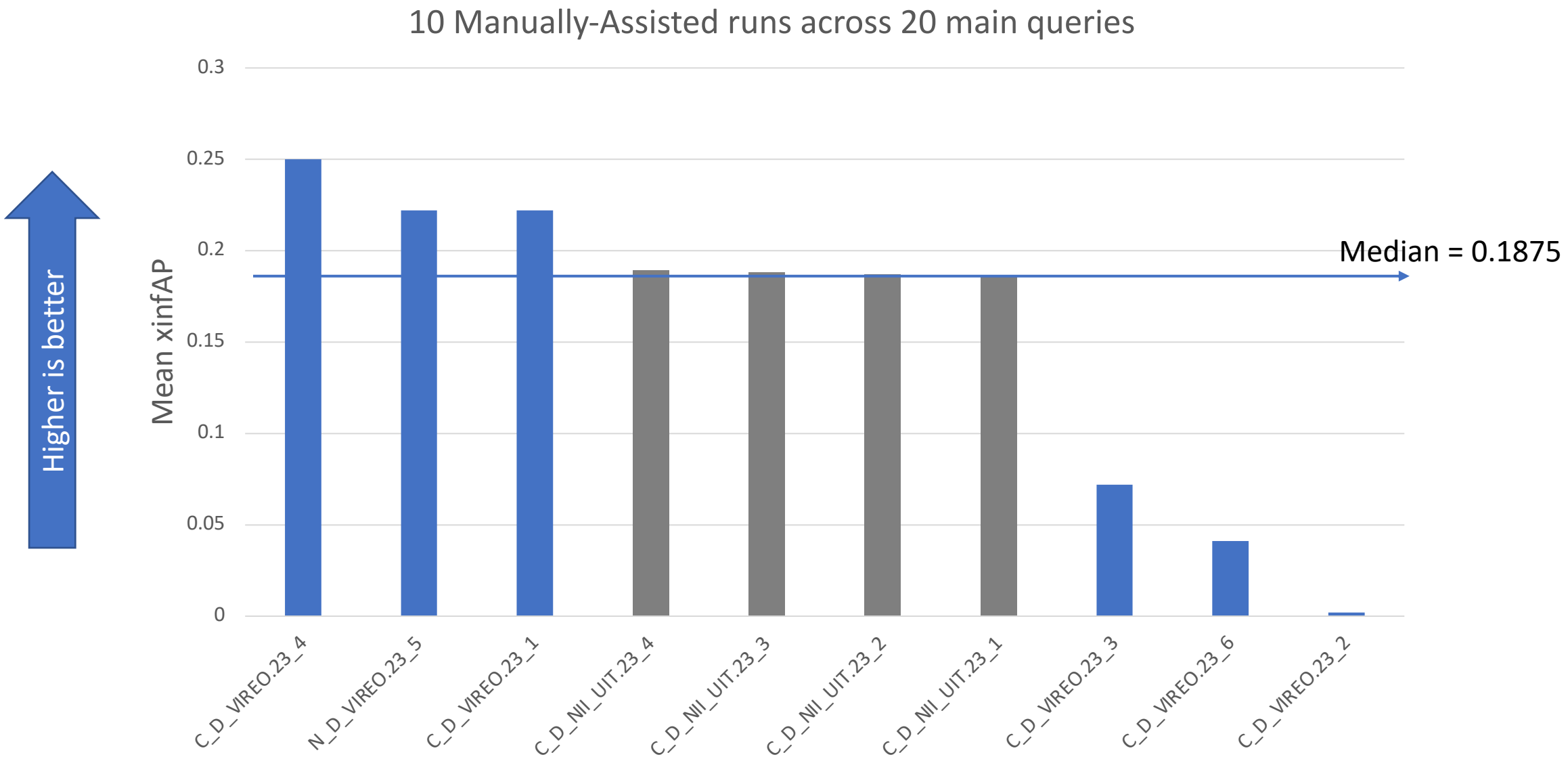
Waseda team's runs 2 and 4 and both statistically sig. than runs 1 and 3.

No statistical diff. between Waseda team's run 2 & 4, and 1 & 3.

WHU_NERCMS run 1 is statistically sig. than RUC_AIM3 run 1.

Top 4 runs of WHU_NERCMS are not statistically sig. from each other.

Sorted Overall Scores – Manually Assisted



Statistical Significance (top M runs)

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

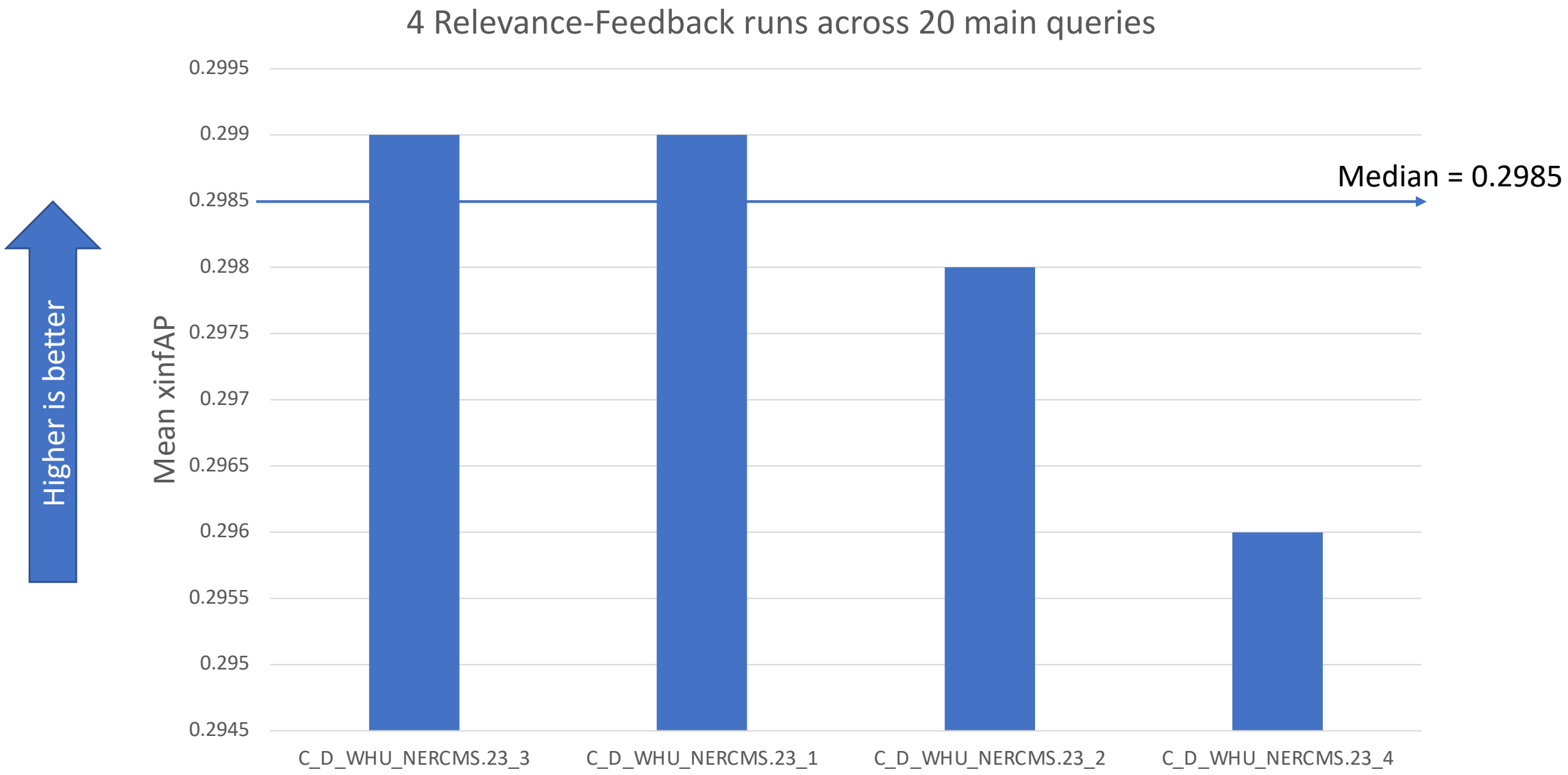
- M_M_C_D_VIREO.23_4
  - M_M_C_D_NII_UIT.23_1
    - M_M_C_D_VIREO.23_3
      - M_M_C_D_VIREO.23_6
        - M_M_C_D_VIREO.23_2
  - M_M_C_D_NII_UIT.23_2
    - M_M_C_D_VIREO.23_3
      - M_M_C_D_VIREO.23_6
        - M_M_C_D_VIREO.23_2
  - M_M_C_D_NII_UIT.23_3
    - M_M_C_D_VIREO.23_3
      - M_M_C_D_VIREO.23_6
        - M_M_C_D_VIREO.23_2
  - M_M_C_D_NII_UIT.23_4
    - M_M_C_D_VIREO.23_3
      - M_M_C_D_VIREO.23_6
        - M_M_C_D_VIREO.23_2
- M_M_N_D_VIREO.23_5
  - M_M_C_D_VIREO.23_3
    - M_M_C_D_VIREO.23_6
      - M_M_C_D_VIREO.23_2

- M_M_C_D_VIREO.23_1
  - M_M_C_D_VIREO.23_3
    - M_M_C_D_VIREO.23_6
      - M_M_C_D_VIREO.23_2

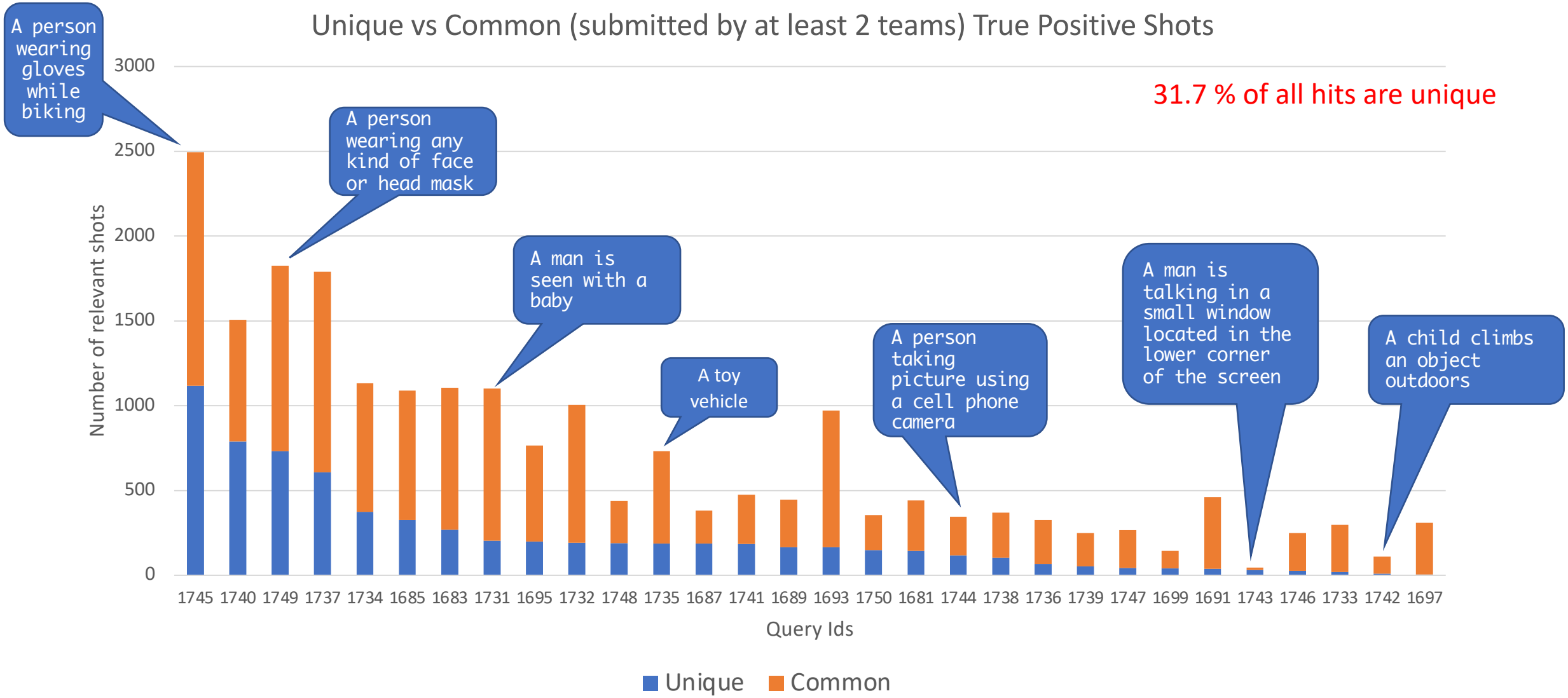
No statistical diff. between VIREO runs 1, 4, and 5

No statistical diff. between NII_UIT runs 1,2,3 and 4

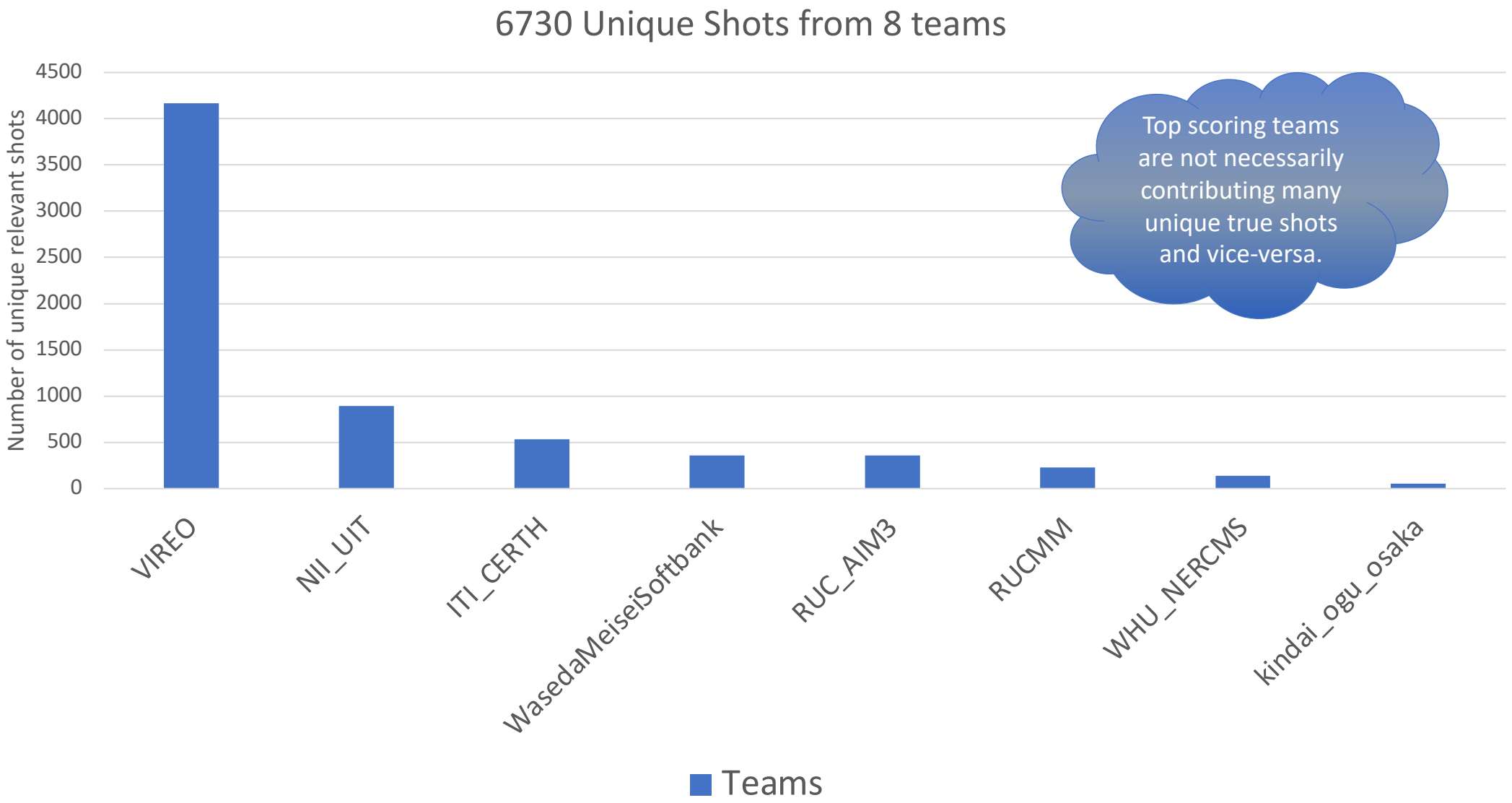
Sorted Overall Scores – Relevance Feedback



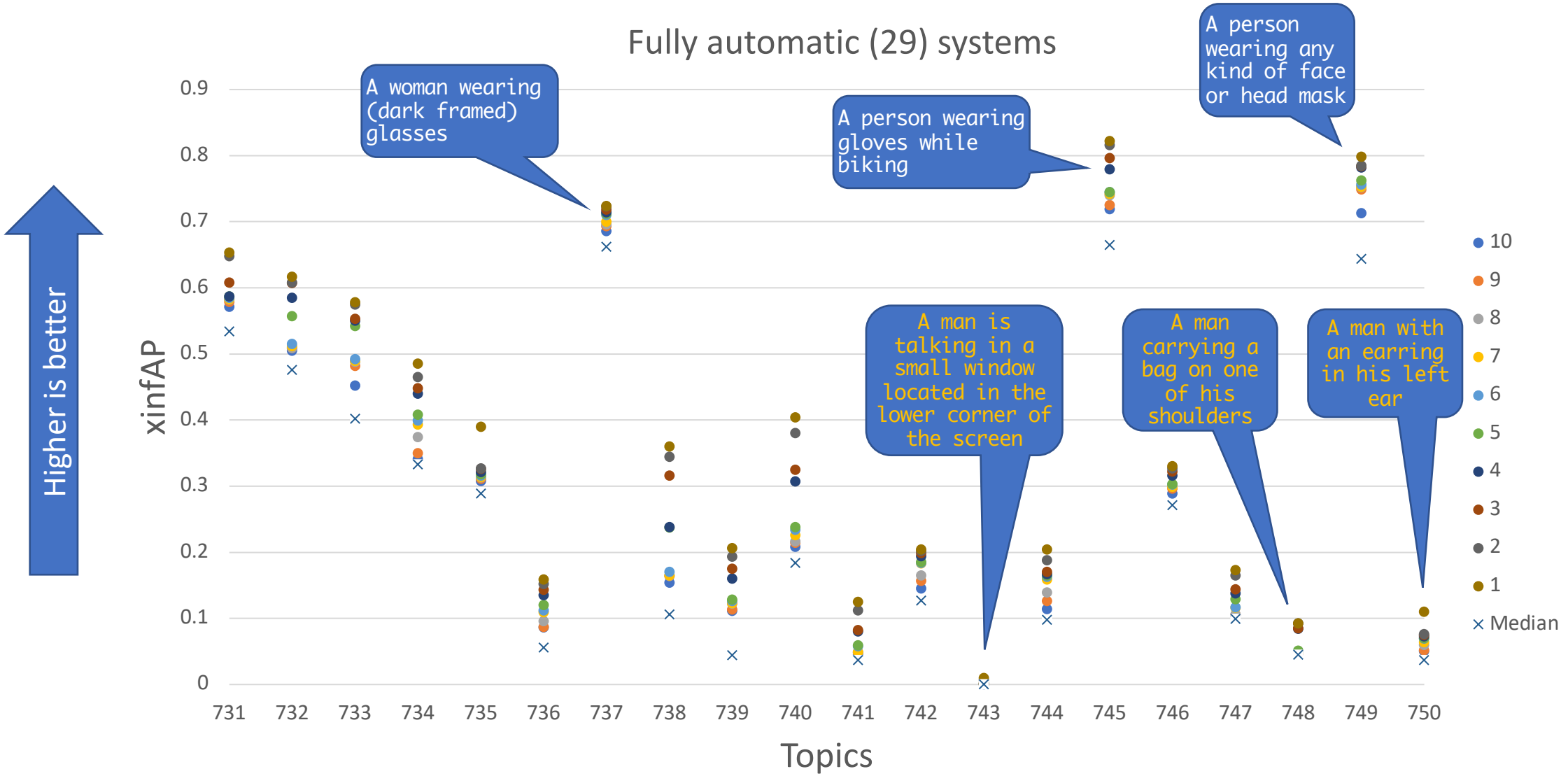
Hits Per Topic



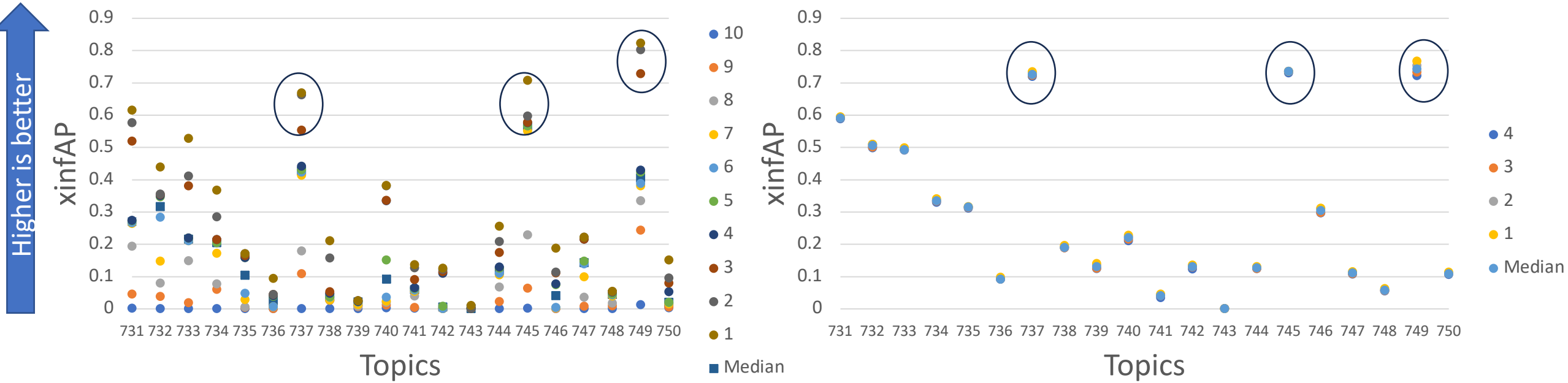
Sorted Unique Hits by Team



Top runs per query (Main Task)



Top runs per query (Main Task)



Easy vs Hard Queries

Top 5 easiest queries (based on avg infAP of runs scored >= 0.5)

Query
A person wearing gloves while biking
A man is seen with a baby
A person wearing any kind of face or head mask
A woman wearing (dark framed) glasses
A woman with red hair

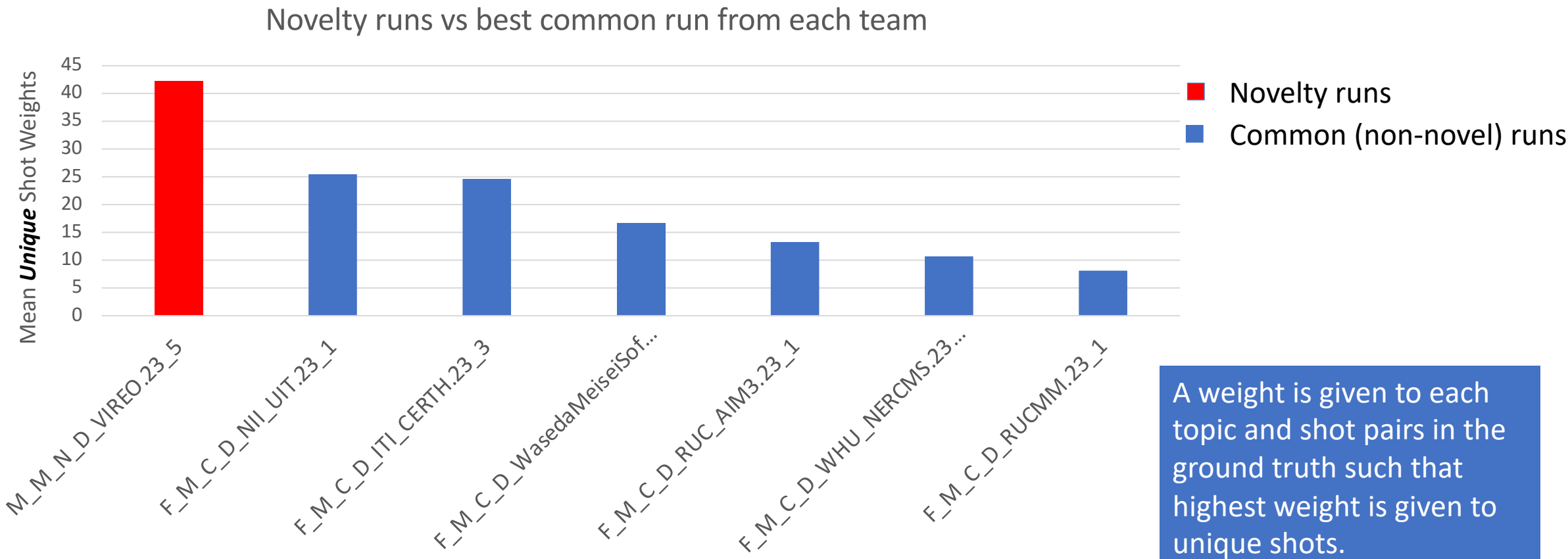
Top 5 hardest queries (based on avg infAP of runs scored < 0.5)

Query
A man is talking in a small window located in the lower corner of the screen
A man carrying a bag on one of his shoulders
A red or blue scarf around someone's neck
A man with an earring in his left ear
A person opens a door and enters a location

Informal method of declaring easy/hard topic:

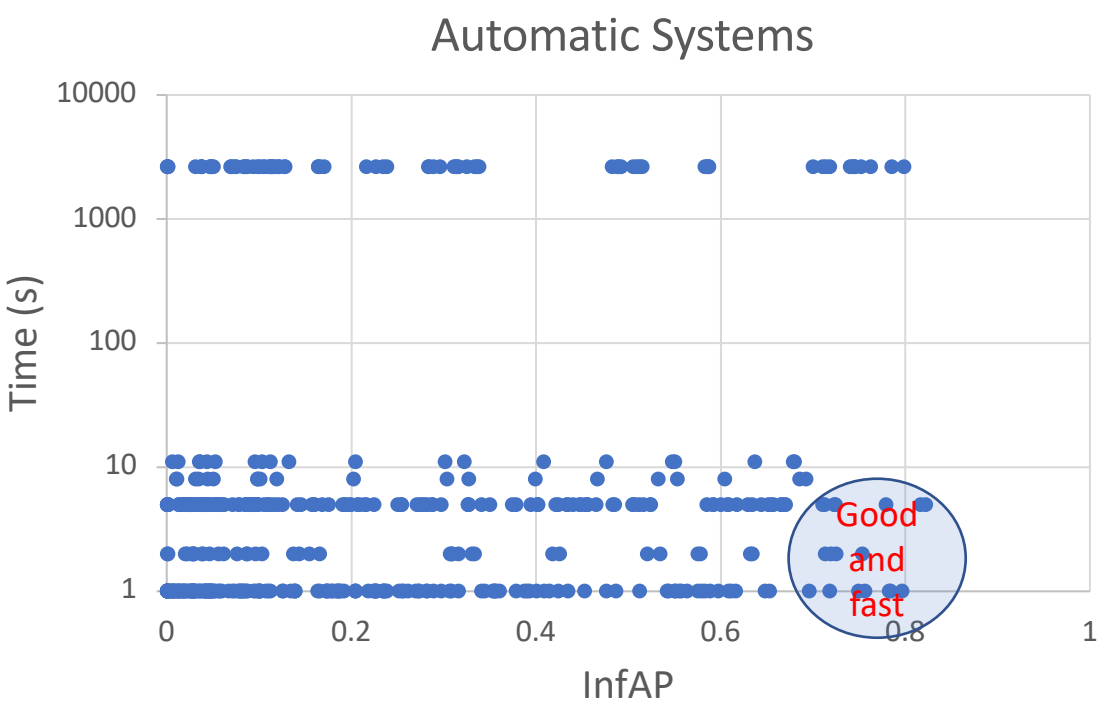
- Sorted number of runs scored above / below 0.5 for any topic.

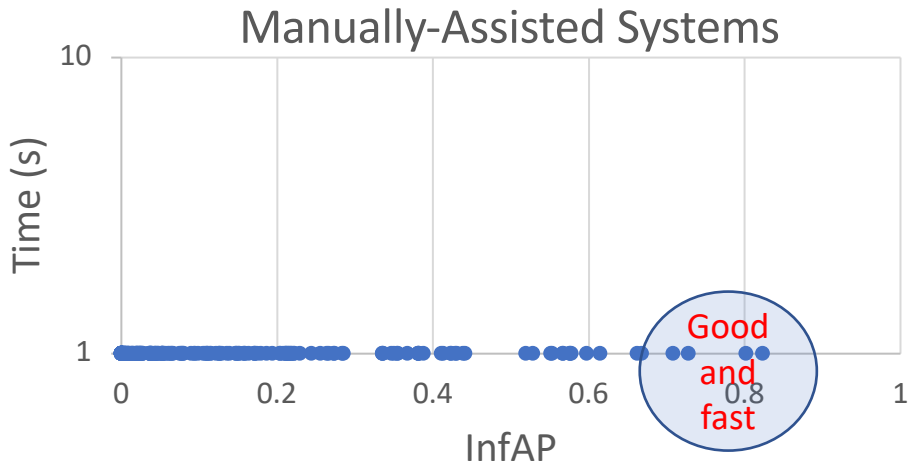
Novelty Scores

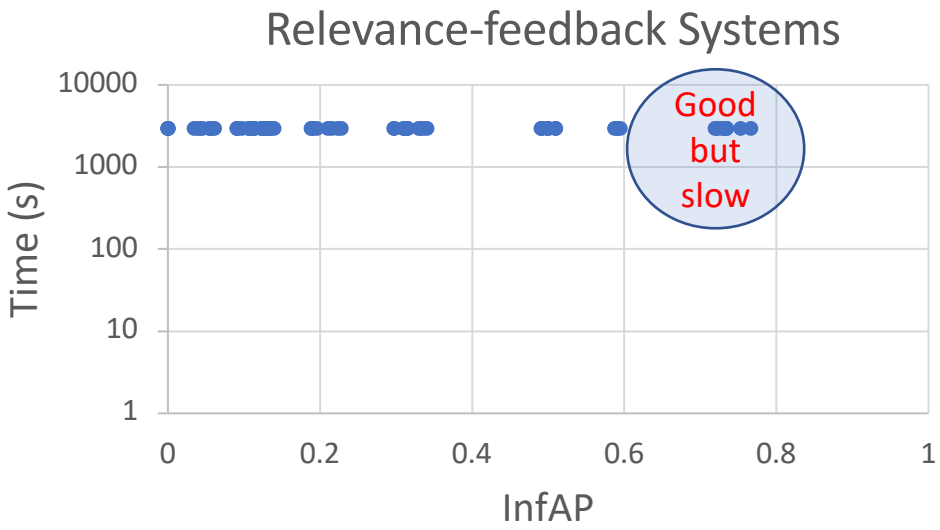


A weight is given to each topic and shot pairs in the ground truth such that highest weight is given to unique shots.

Efficiency


Automatic Systems
Time (s)
10000
1000
100
10
1
0
0.2
0.4
0.6
0.8
1
InfAP
Good
and
fast


Manually-Assisted Systems
Time (s)
10
1
0
0.2
0.4
0.6
0.8
1
InfAP
Good
and
fast

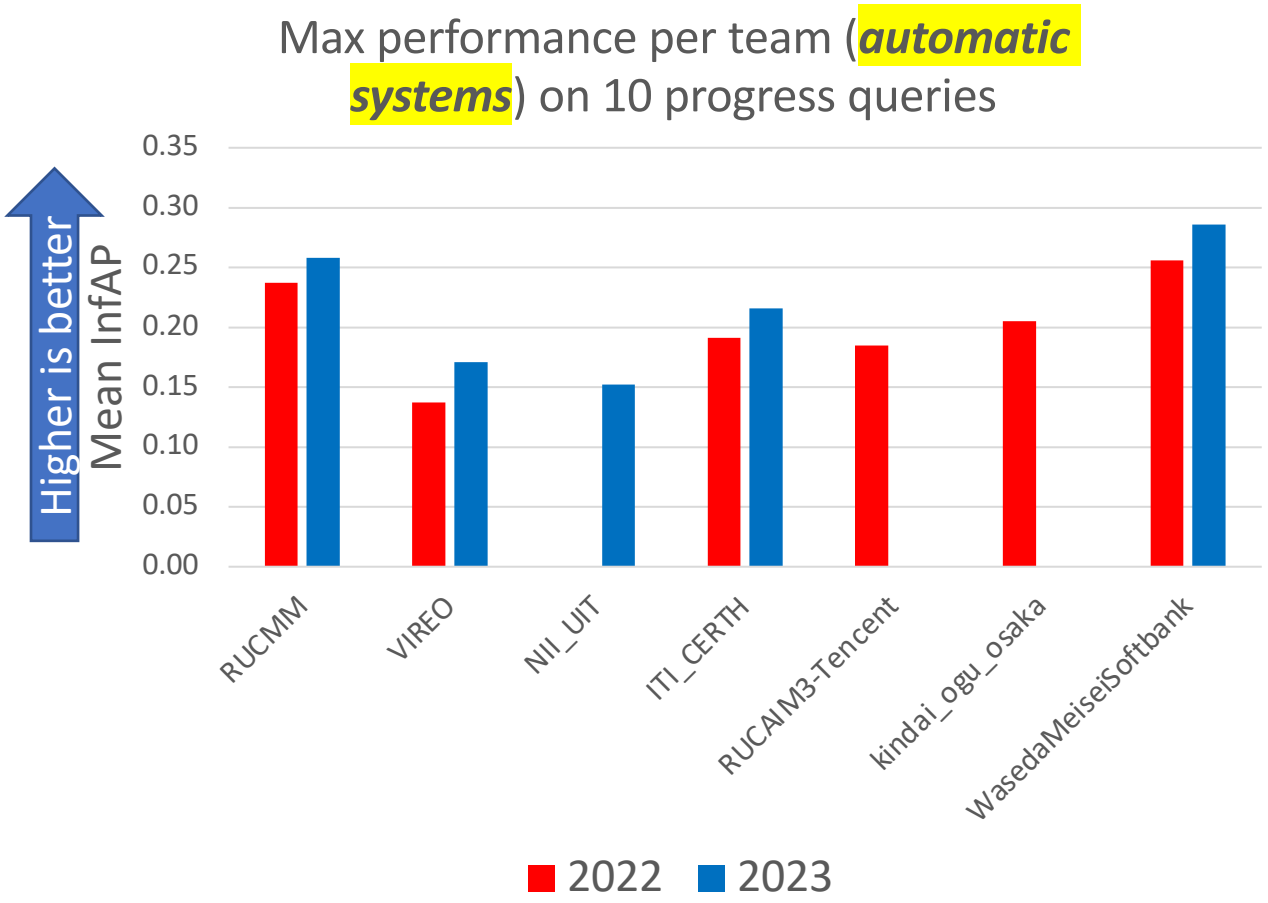

Relevance-feedback Systems
Time (s)
10000
1000
100
10
1
0
0.2
0.4
0.6
0.8
1
InfAP
Good
but
slow

Progress Task Plan

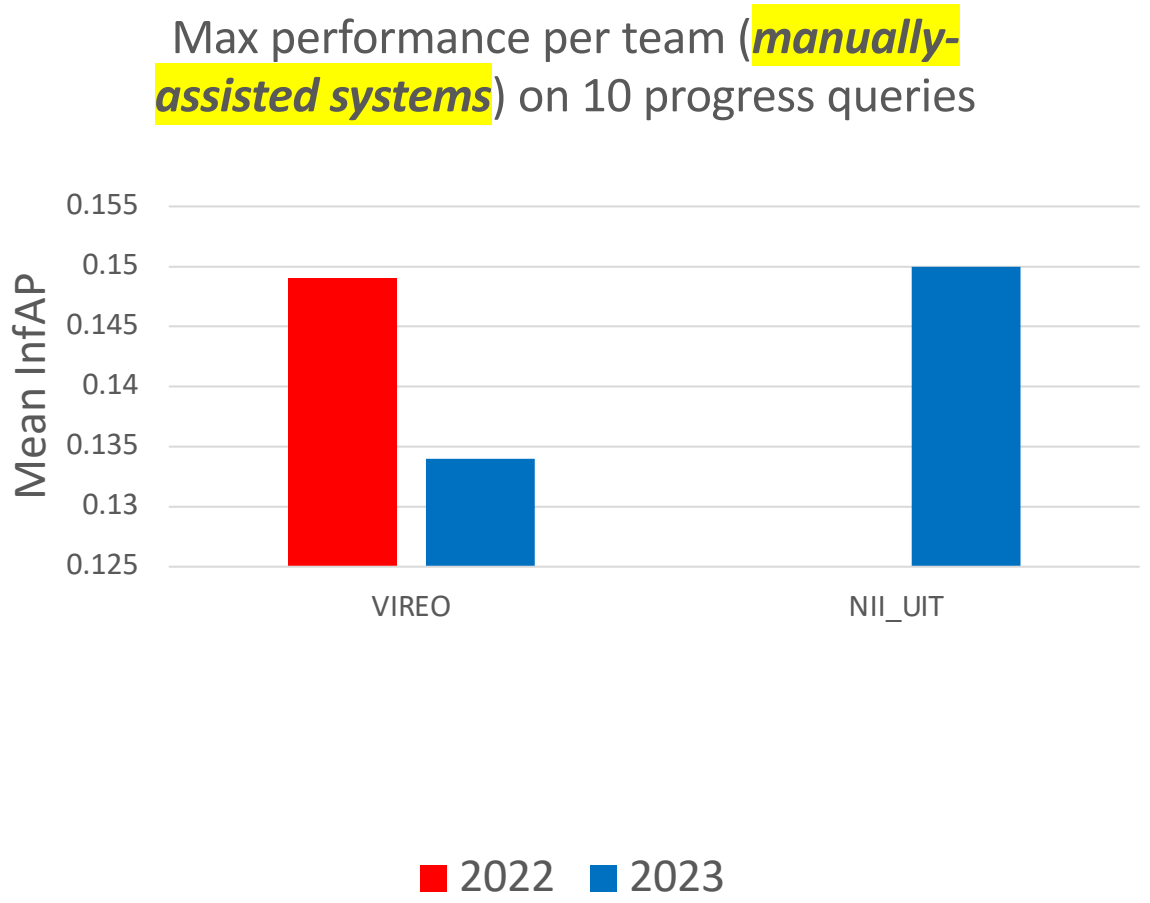
		Evaluation year		
Submission year		2022	2023	2024
	2022	***Systems:*** Submit 20 fixed progress queries		
	2023		***Systems:*** Submit 20 fixed progress queries ***NIST:*** Eval 10 queries (set A)	
	2024			***Systems:*** Submit 20 fixed progress queries ***NIST:*** Eval 10 queries (set B)

Goals : Evaluate 10 (set A) common queries submitted in 2 years (2022 - 2023)
Evaluate 10 (set B) common queries submitted in 3 years (2022 - 2024)

Progress set-A results (2022-2023)



Majority of automatic systems performed better in 2023 compared to 2022.



Top M run of VIREO in 2023 did not exceed 2022 top run

Samples of false positives

A woman with a ponytail

A man wearing a lanyard around his neck

A person wearing a ring in their nose

A toy vehicle

A person wearing gloves while biking

A man with an earring in his left ear

2023 Main Approaches

- Use of multiple text-image / text-video common latent embeddings: VSE++, CLIP and its various variants: SLIP, BLIP, BLIP-2, LaCLIP, OpenCLIP, TeachCLIP …
- Query expansion with ChatGPT
- Use of generative Text-to-Image generative approaches
- Transformer-based extension of a cross-modal deep network architecture
- Top-K Feedback and a new algorithm Quantum-Theoretic Interactive Ranking Aggregation (QT-IRA)
- No more concept bank approaches but “dual task” (interpretable embeddings)
- Use of multiple text-image / text-video annotated collections: MSR-VTT, TGIF, Vatex, Flickr8k/30k, MS-COCO, Conceptual Captions, …
- Large number of combinations and fusion (normalization | averaging)
- Lightweight Attentional Feature Fusion
- Hard to distinguish between data / features effects and algorithmic effects

2023 Task Observations

- Submissions
  - 7 teams finished the main task (43 runs) including 5 teams submitting to the progress task (30 runs).
  - 29 automatic systems, 10 manually-assisted systems, and 4 relevance feedback systems joined the main task.
  - Run training types are dominated by “D” runs. No “E” or “F” runs.
  - No teams submitted “optional” explainability results with their runs!
  - Only 1 Novelty system submitted. Better than common runs on novelty metric.
- Performance
  - Above 2022 in general.
  - Few automatic systems are good and fast (< 10 sec).
  - High similarity between F, M, and R systems in terms of query performance relatively to each other.
  - Top scoring teams not necessary contributing a lot of unique true shots and vice-versa.
  - About 32% of all hits are unique. 68% are common hits across the runs.
  - 16.2% of all judged shots across all queries are true positives.
  - Hard queries are the ones asked for unusual combinations of facets (compared to well-known concepts)
  - For low performance queries, usually all systems are condensed in small range.
  - For mid to high performance queries, the top 10 runs vary in their range of performance.

Interactive Video Retrieval

During the Video Browser Showdown (VBS)

At MMM 2024
30th International Conference on Multimedia Modeling, January 2024, Amsterdam, The Netherlands

- 10 Ad-Hoc Video Search (AVS) topics : Each AVS topic has several/many target shots (from V3C1 + V3C2 datasets) 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.
- 10 Q/A tasks, described via text and needs a textual answer.
- Registration for the task is now closed

AVS Task Discussion

- 2023 AVS reflections:
  - Feedback (Queries , Schedule, Data, Trends, Lessons Learned, etc)
- 2024 AVS Plan:
  - 20 new queries
  - 20 progress (NIST will evaluate 10 (compare progress between 2022 - 2024))
  - Needed updates to existing task?
    - Add
    - Remove
    - Change
  - Suggestions for attracting more teams
- Opportunities:
  - New tasks
  - New collaborations, venues