TRECVID 2022
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;

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 2022 (30 main) Queries by complexity

A Person, Location, or Object

A man with a white beard
A room with blue wall
A construction site
A parked white car
A type of cloth hanging on a rack, hanger, or line
Building with columns during daytime

Person + Location

A kneeling man outdoors
Two or more persons in a room with a fireplace

Object + Action

A drone landing or rising from the ground

Person + Action

A person is mixing ingredients in a bowl, cup, or similar type of containers
A female person bending downwards
A person is in the act of swinging

Person + Object

A person wearing a light t-shirt with dark or black writing on it
A woman wearing a head kerchief
A man wearing black shorts

Object + Location

A large stone building from the outside
A piece of heavy farm equipment or machine seen outdoors
A clock on a wall in a room

AVS 2022 (30 main) Queries by complexity

Person + Action + Location

Two persons are seen while at least one of them is speaking in a non-English language outdoors

A woman is eating something outdoors

A person is biking through a path in a forest

An Asian bride and groom celebrating outdoors

Person + Action + Object

A man and a bike in the air after jumping from a ramp

A woman holding or smoking a cigarette

Two teams playing a game where one team have their players wearing white t-shirts.

Person + Object + Location

A ring shown on the left hand of a person

A man is holding a knife in a non-kitchen location

Object + Action + Location

A black bird seen on a dry area sitting, walking, or eating

Person + Action + Object + Location

Two persons wearing white outfits and black belts demonstrate martial arts in a room with floor mats

Two adults are seated in a flying paraglider in the air

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

Teams – Main Task (33 runs)

DIGITAL VIDEO RETRIEVAL at NIST

Team Name (7 Finishers)	Organization	System Type		
		Manually assisted	Fully automatic	Novelty run
VIREO	Singapore Management University; City University of Hong Kong	5	5	1
Kindai_ogu_osaka	Kindai University; Osaka Gakuin University; Osaka University		4	
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas		2	
RUCAIM3-Tencent	Renmin University of China		4	
RUCMM	Renmin University of China		4	
WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation		4	
CamiloUchile	Uchile		4	

Teams – Progress Task (28 runs)

Team Name (6 Finishers)	Organization	System Type		
		Manually assisted	Fully automatic	Novelty run
VIREO	Singapore Management University; City University of Hong Kong	5	5	
Kindai_ogu_osaka	Kindai University; Osaka Gakuin University; Osaka University		4	
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas		2	
RUCAIM3-Tencent	Renmin University of China		4	
RUCMM	Renmin University of China		4	
WasedaMeiseiSoftbank	Waseda University; Meisei University; SoftBank Corporation		4	

Evaluation Methodology

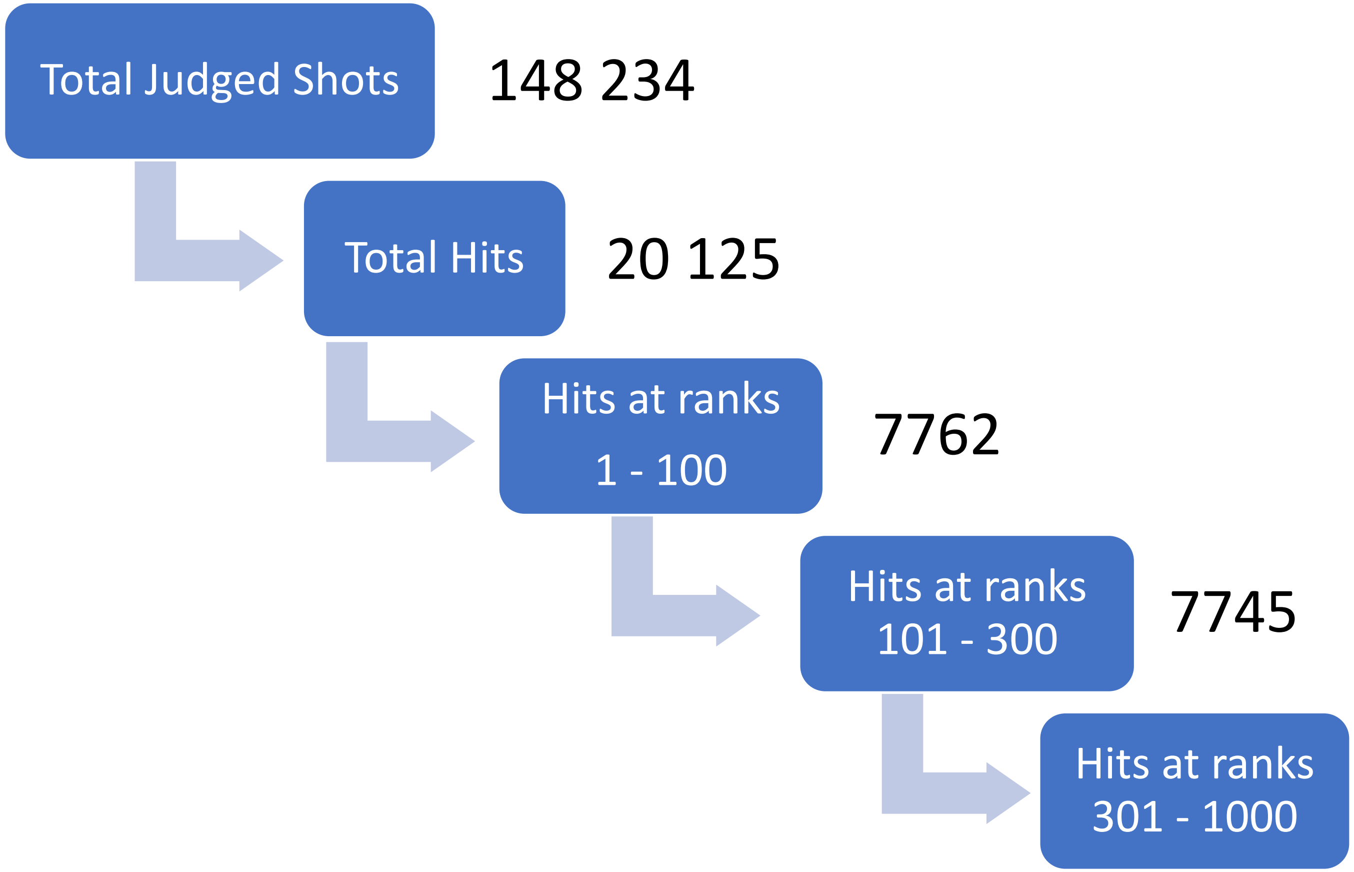
- NIST judged 100% of top (ranks 1 – 300) pooled results from all submissions and sampled 25% from the rest of pooled results (ranks 301 – 1000).
- Stats of sampled and judged clips (ranks 301 to 1000) across all runs and topics
  - At minimum, 24.3 % of any run and query results were sampled and judged
  - At maximum, 76.7 % of any run and query results were sampled and judged
  - On average, 55 % 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

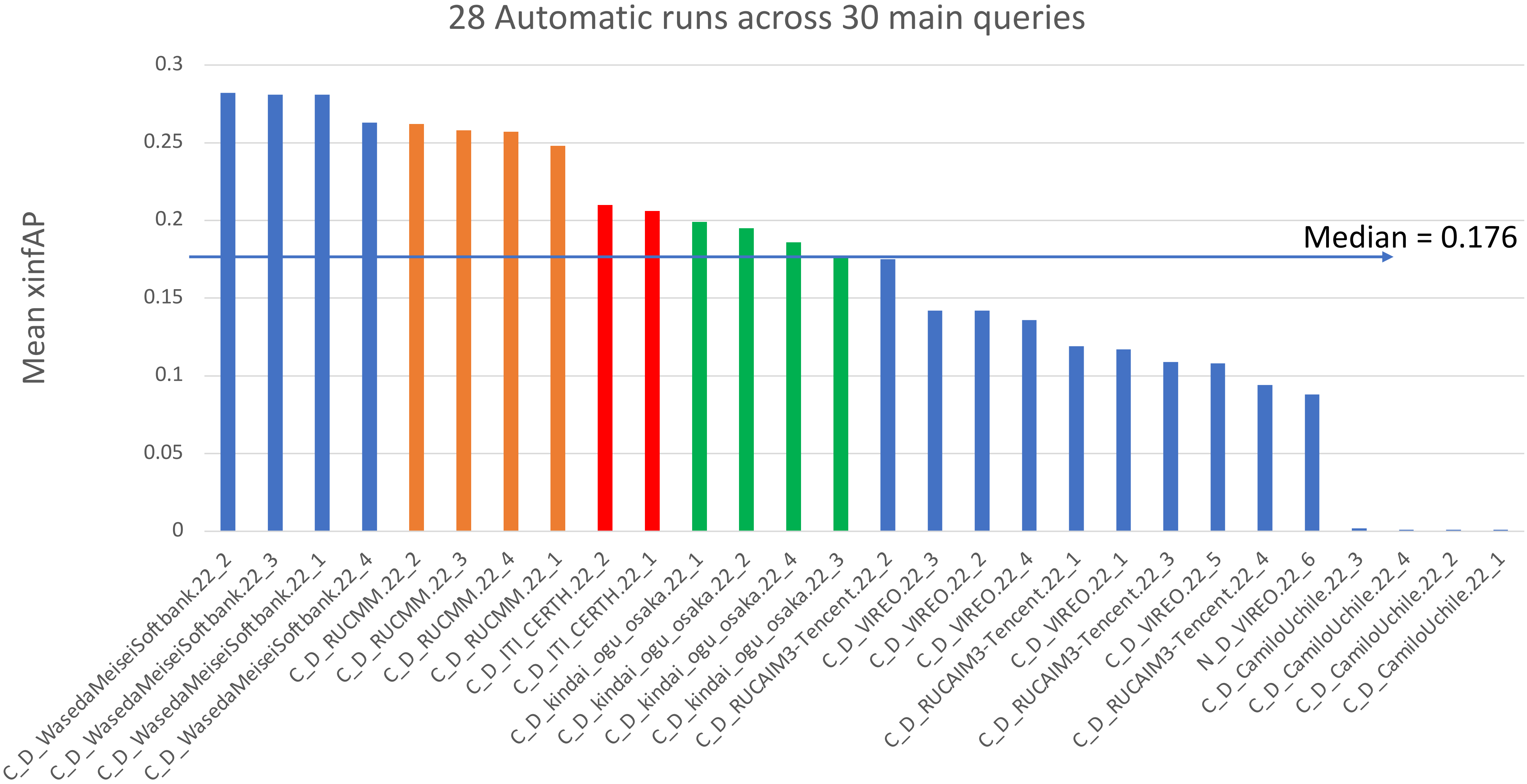


Total Judged Shots	148 234
Total Hits	20 125
Hits at ranks 1 - 100	7762
Hits at ranks 101 - 300	7745
Hits at ranks 301 - 1000	4618

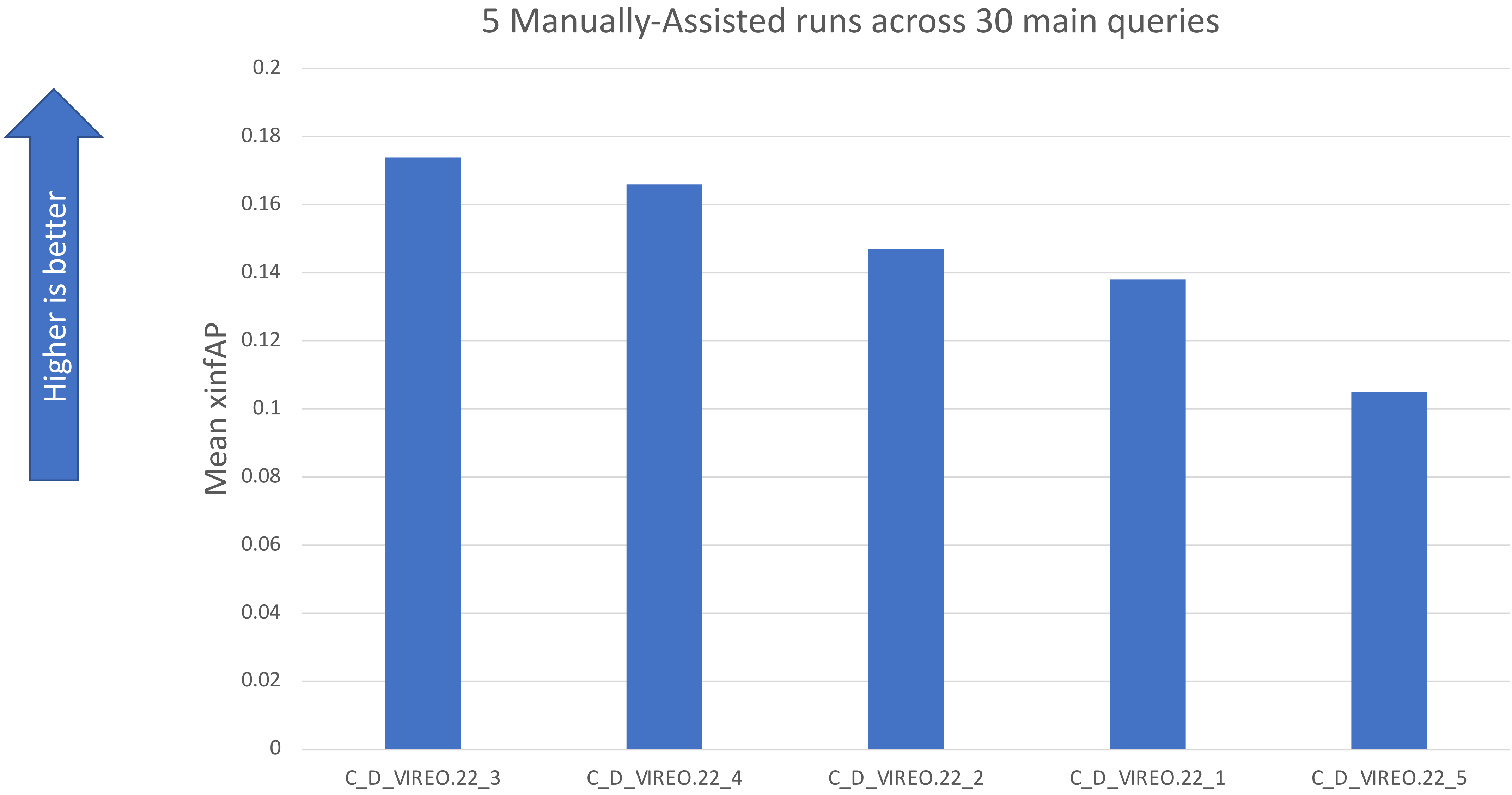
5 human assessors
30 queries
500 hours of labor work

Sorted Overall Scores – Automatic Runs

Higher is better



Sorted Overall Scores – Manually Assisted



Statistical Significance (top 10 runs)

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

No statistical diff. between WasedaMeiseiSoftbank runs 1, 3, & 2.

No statistical diff. between all RUCMM runs.

ITI_CERTH run 2 is better than run 1

F_M_C_D_WasedaMeiseiSoftbank.22_4	F_M_C_D_ITI_CERTH.22_2	F_M_C_D_ITI_CERTH.22_1	
🟩	🟩	🟩	F_M_C_D_WasedaMeiseiSoftbank.22_1
🟩	🟩	🟩	F_M_C_D_WasedaMeiseiSoftbank.22_3
🟩	🟩	🟩	F_M_C_D_WasedaMeiseiSoftbank.22_2
	🟩	🟩	F_M_C_D_RUCMM.22_4
	🟩	🟩	F_M_C_D_WasedaMeiseiSoftbank.22_4
	🟩	🟩	F_M_C_D_RUCMM.22_2
	🟩	🟩	F_M_C_D_RUCMM.22_3
	🟩	🟩	F_M_C_D_RUCMM.22_1
		🟩	F_M_C_D_ITI_CERTH.22_2
			F_M_C_D_ITI_CERTH.22_1

Statistical Significance (top 10 runs)

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

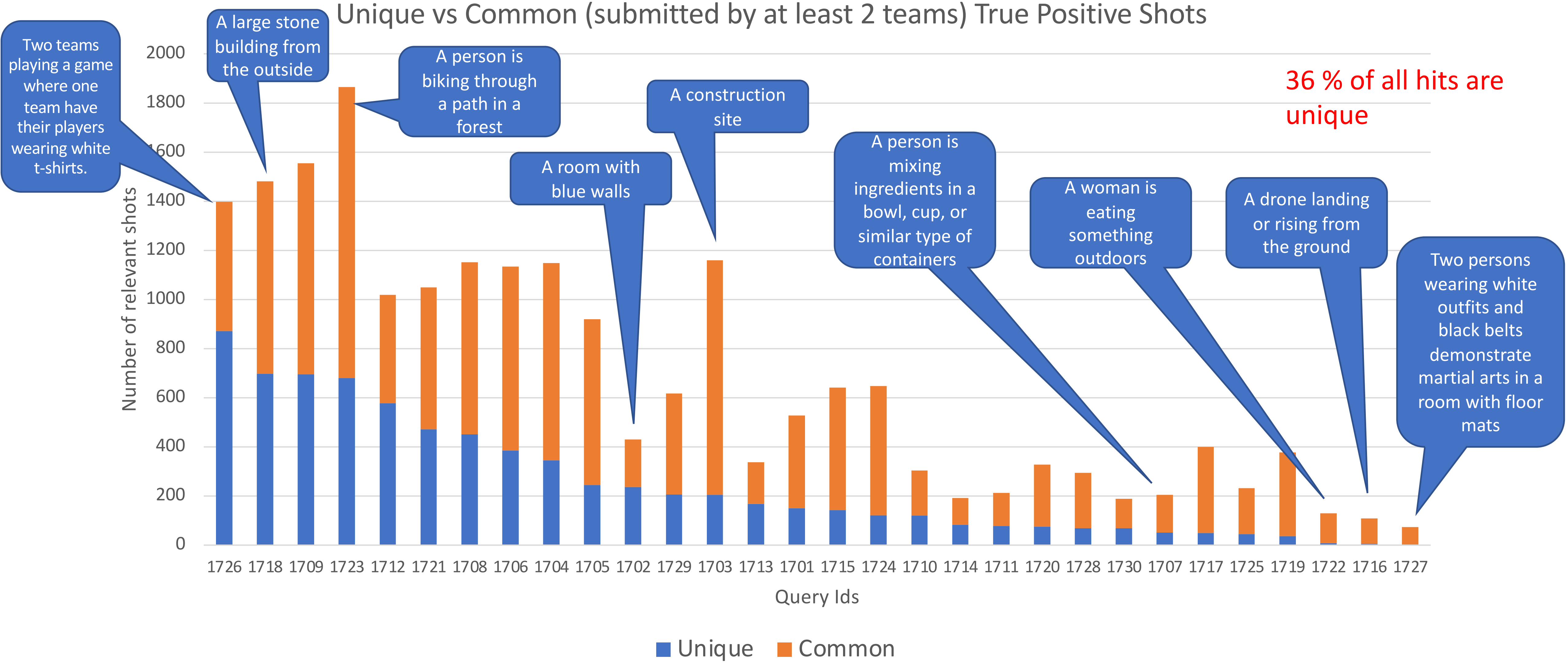
VIREO run 3 is better than all other VIREO runs.

VIREO run 4 is better than runs 1, 2, & 5.

VIREO runs 1 & 2 are better than run 5

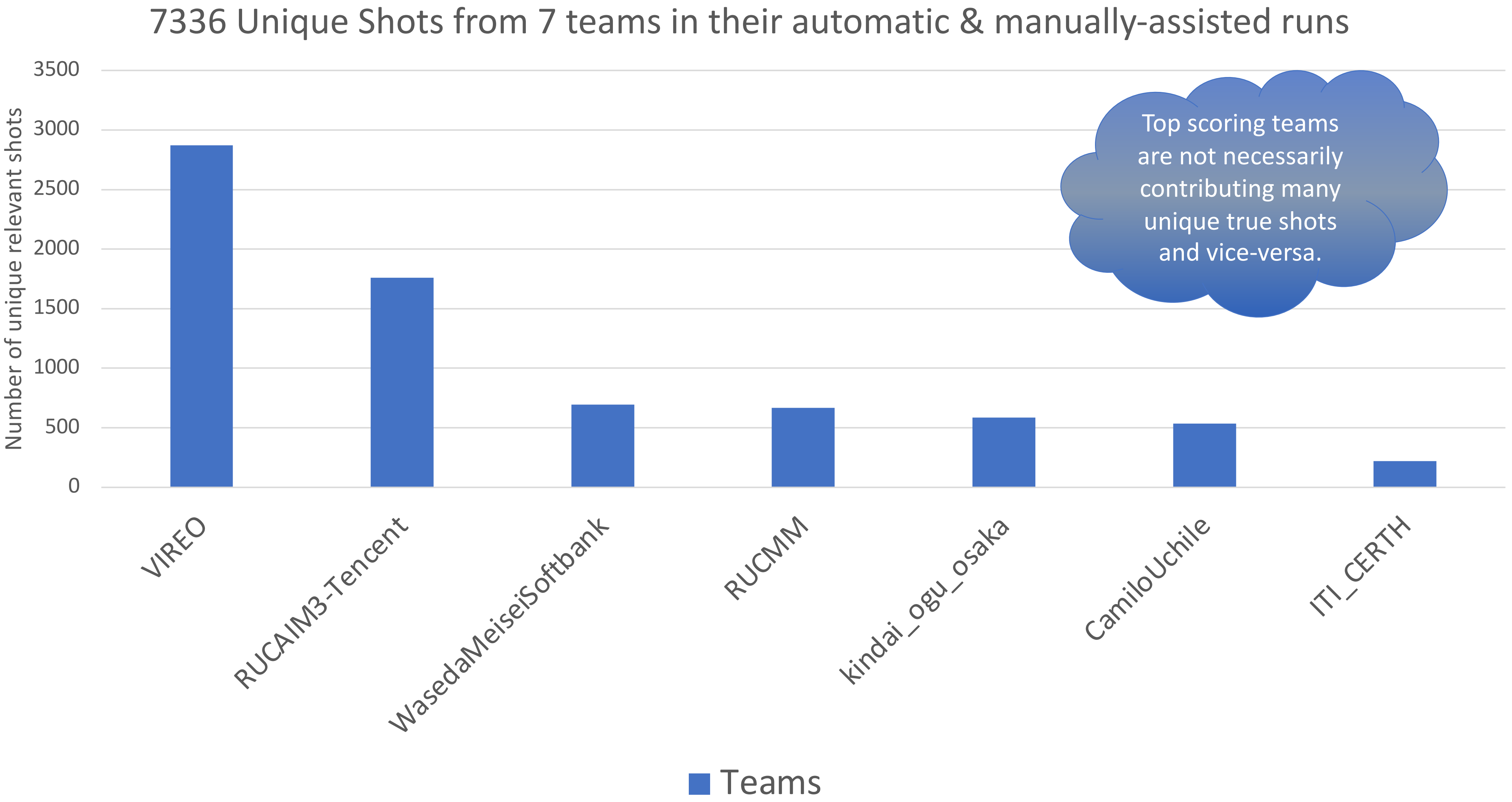
	M_M_C_D_VIREO.22_4	M_M_C_D_VIREO.22_1	M_M_C_D_VIREO.22_2	M_M_C_D_VIREO.22_5
M_M_C_D_VIREO.22_3	🟩	🟩	🟩	🟩
M_M_C_D_VIREO.22_4		🟩	🟩	🟩
M_M_C_D_VIREO.22_1				🟩
M_M_C_D_VIREO.22_2				🟩
M_M_C_D_VIREO.22_5				

Hits Per Topic



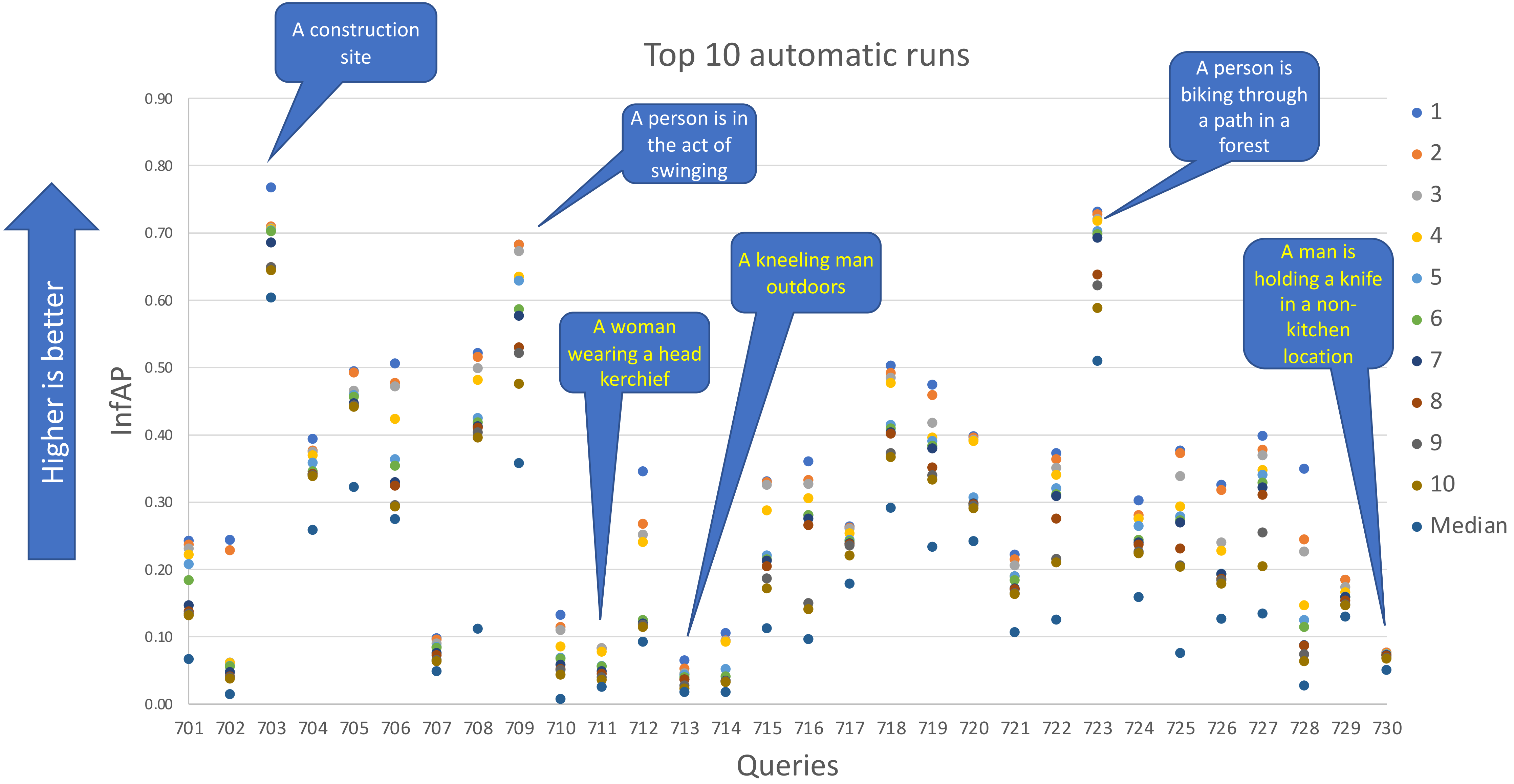
36 % of all hits are unique

Sorted Unique Hits by Team



Top scoring teams are not necessarily contributing many unique true shots and vice-versa.

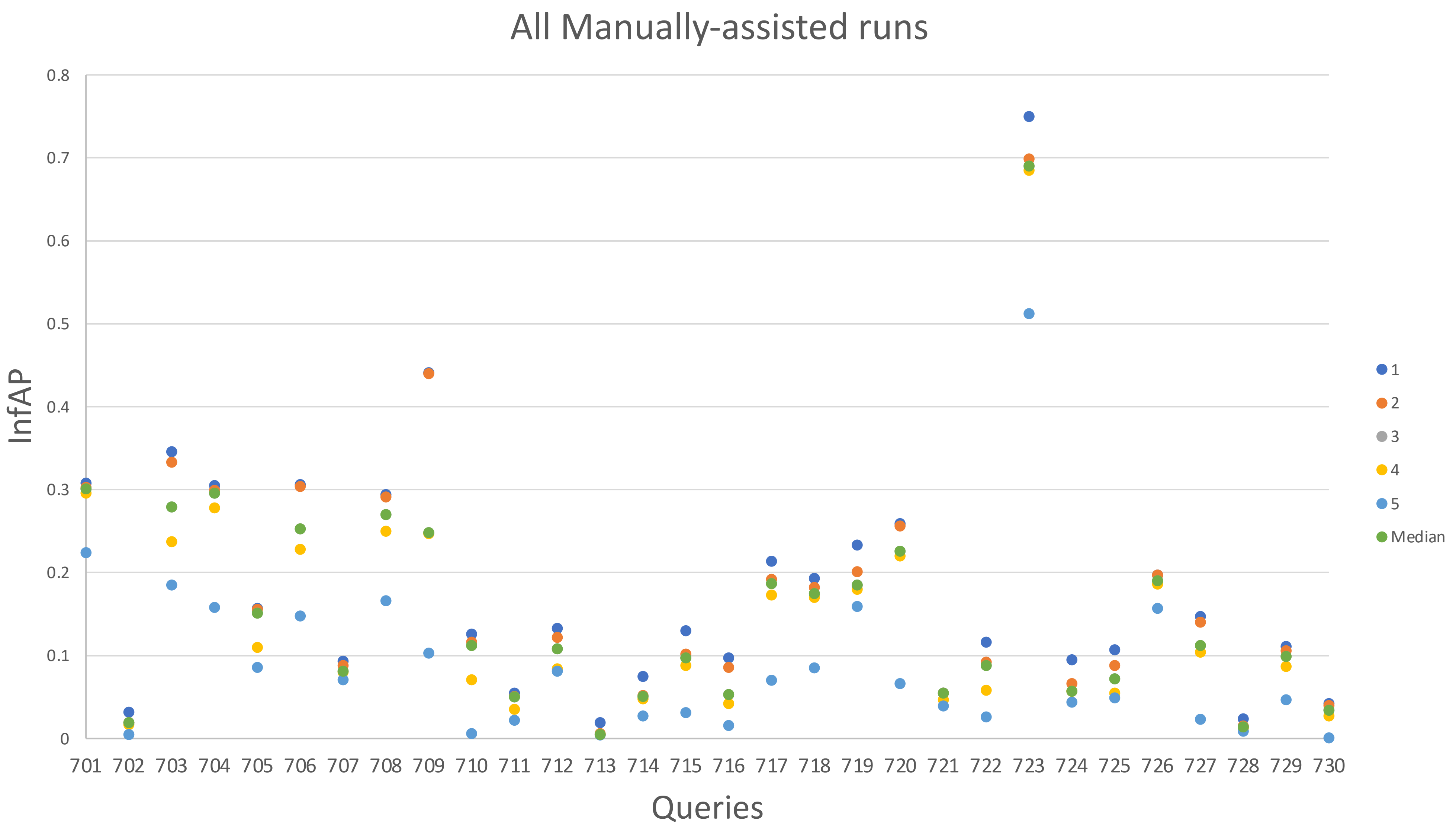
Top runs per query (Main Task)



Top runs per query (Main Task)

NIST

Higher is better



Easy vs Hard Queries

Top 5 easiest queries (based on avg infAP of runs scored >= 0.38)

Query
A person is biking through a path in a forest
A construction site
A person is in the act of swinging
A female person bending downwards
A type of cloth hanging on a rack, hanger, or line

Top 5 hardest queries (based on avg infAP of runs scored < 0.38)

Query
A kneeling man outdoors
Two or more persons in a room with a fireplace
A woman wearing a head kerchief
A room with blue wall
A person wearing a light t-shirt with dark or black writing on it

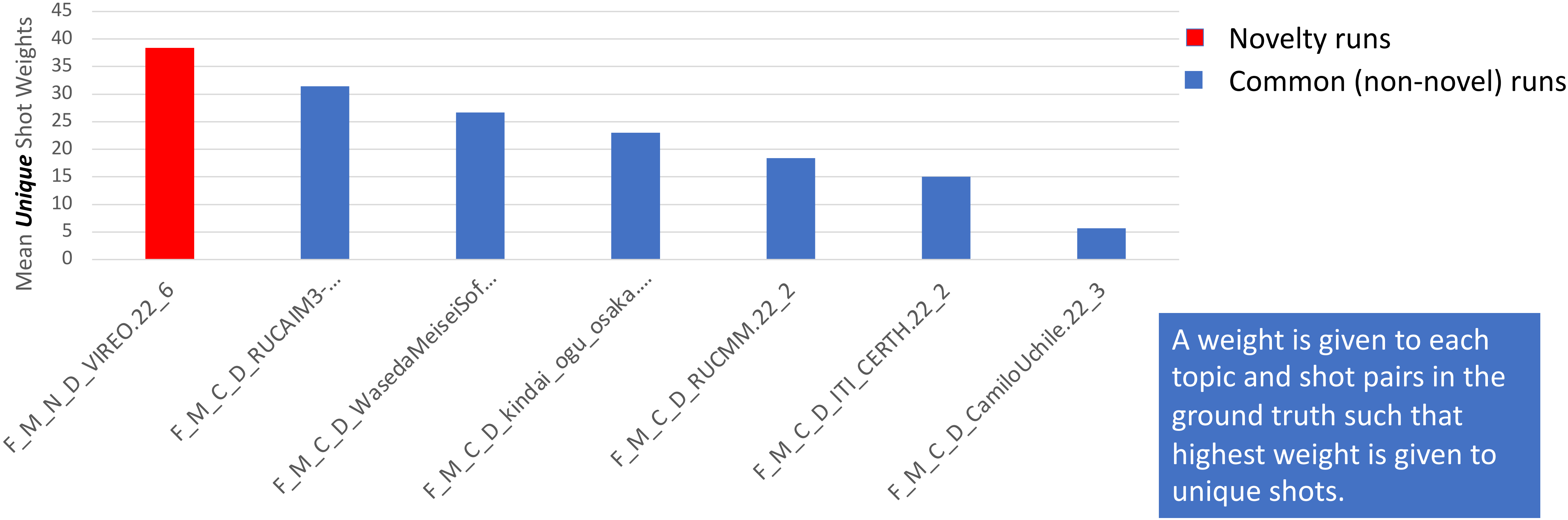
Informal method of declaring easy/hard topic:

1- Threshold of 0.38 xinfAP is calculated as the mid point between all topics score range.

2- Sorted number of runs scored above / below 0.38 for any topic.

Novelty Scores

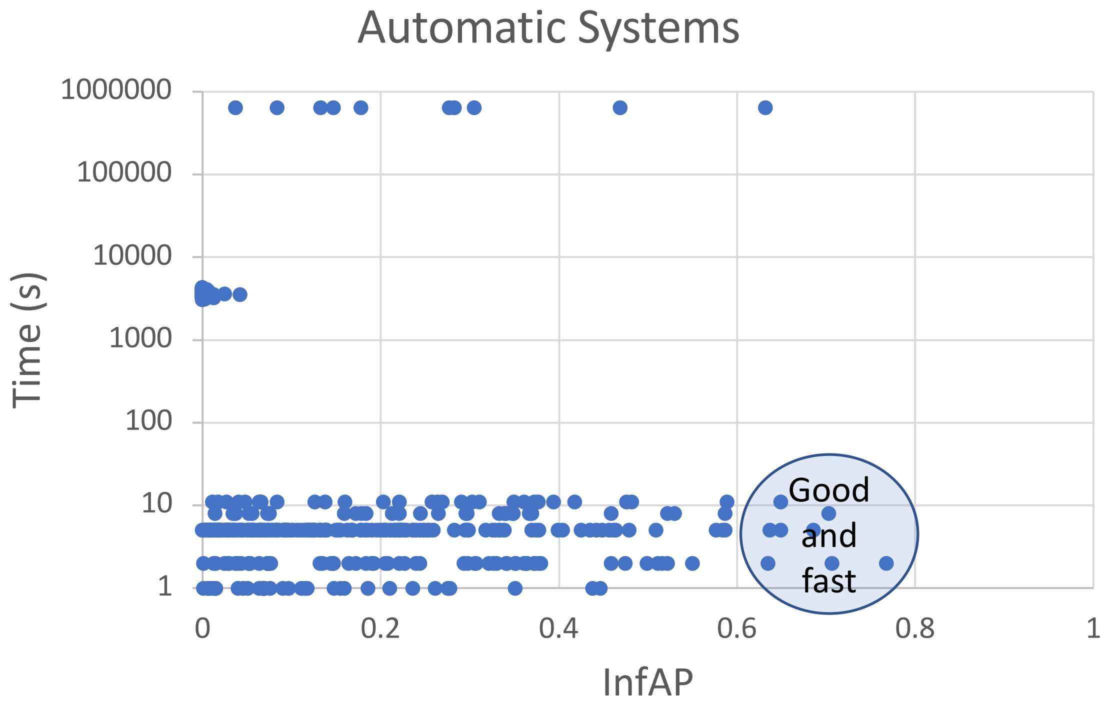
Novelty runs vs best common run from each team

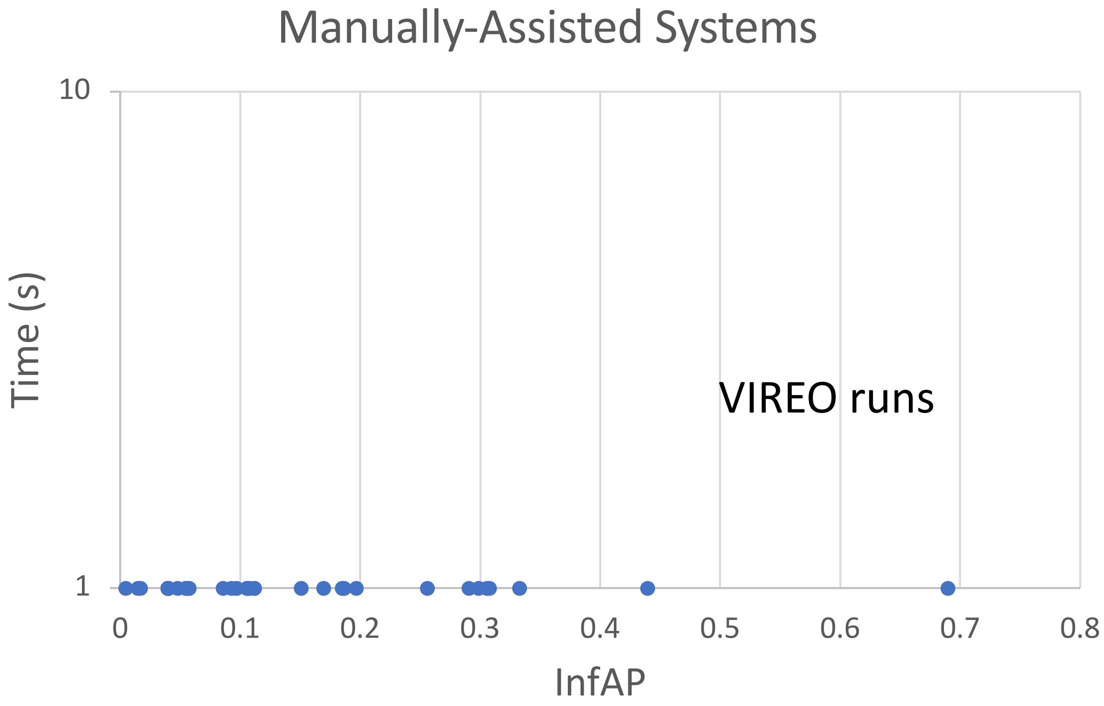


A weight is given to each topic and shot pairs in the ground truth such that highest weight is given to unique shots.

Efficiency

DIGITAL VIDEO RETRIEVAL at NIST



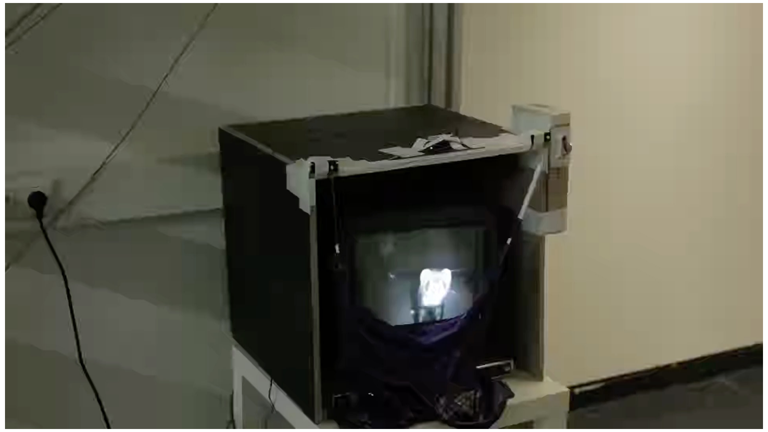


Progress Task Plan

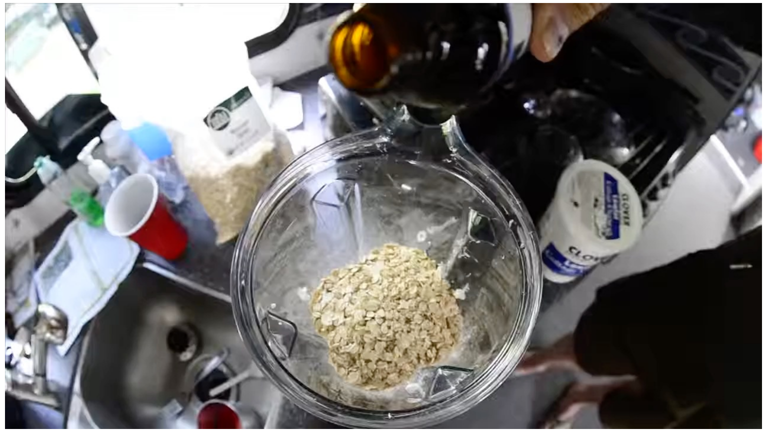
		Evaluation year		
Submission year		2022	2023	2024
	2022	***Systems:*** Submit 20 fixed progress queries		
	2023		***Systems:*** Submit 20 fixed progress queries ***NIST:*** Eval 10 queries (set A)	
	2024			***Systems:*** Submit 20 fixed progress queries ***NIST:*** Eval 10 queries (set B)

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)

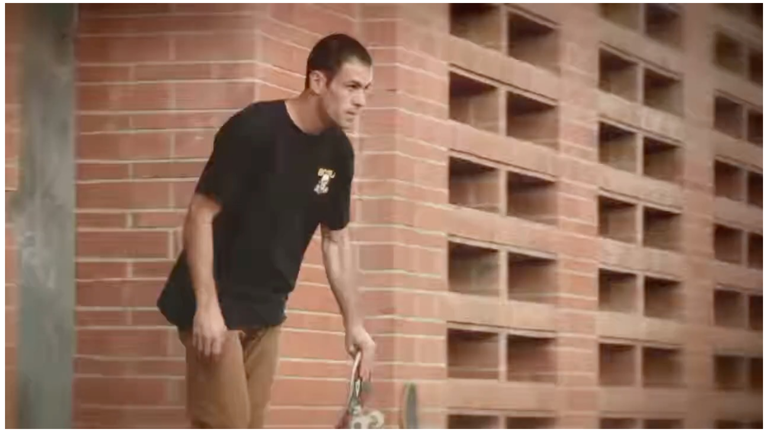
Samples of frequent false positives



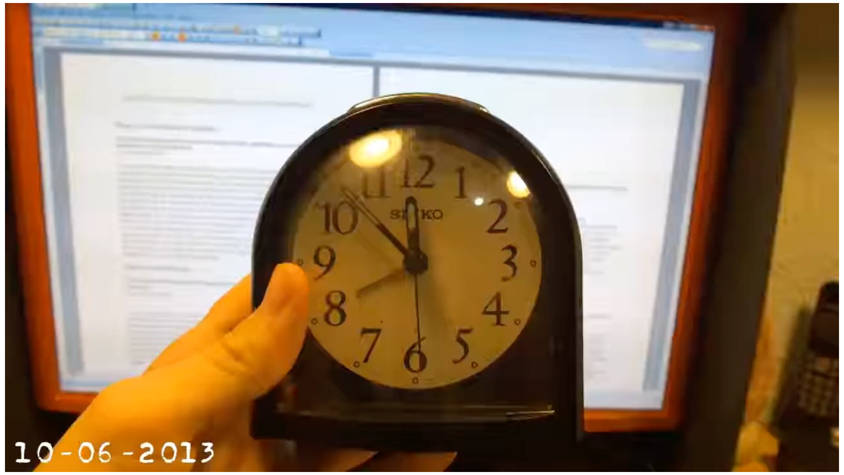
A type of cloth hanging on a rack, hanger, or line



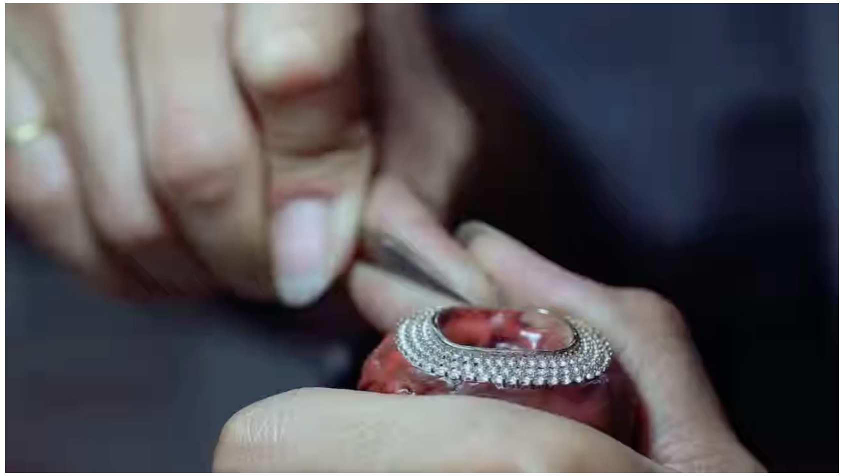
A person is mixing ingredients in a bowl, cup, or similar type of containers



A man wearing black shorts

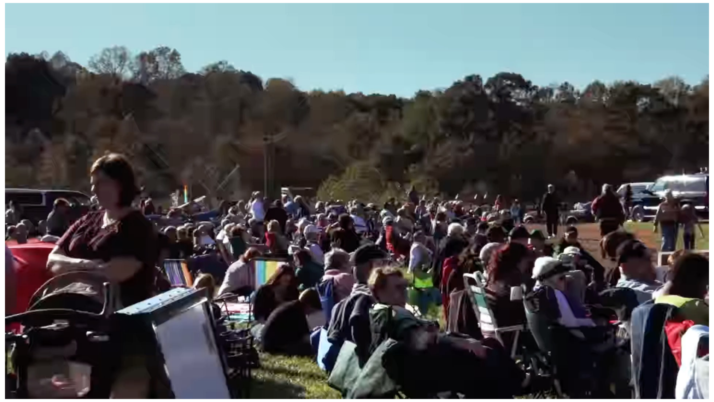


A clock on a wall in a room

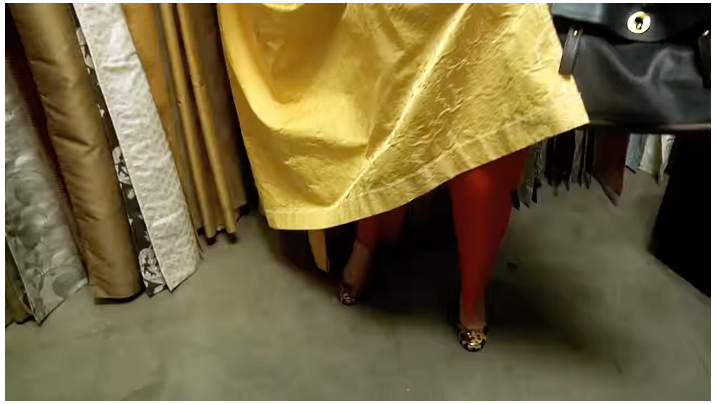


A ring shown on the left hand of a person

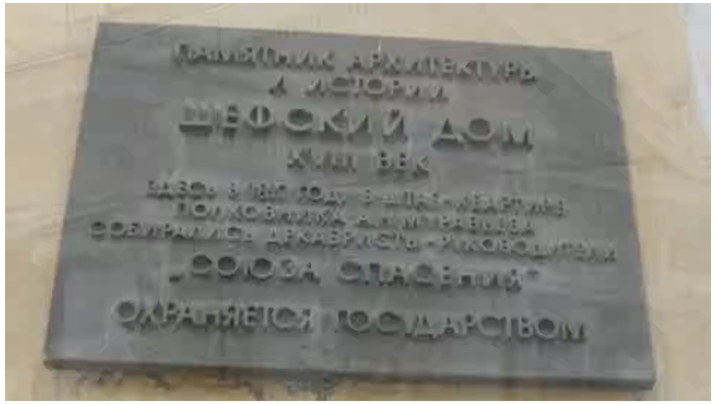
Samples of hard true positives



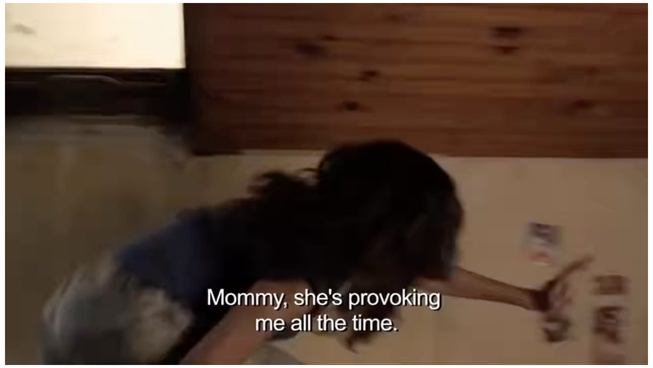
A parked white car



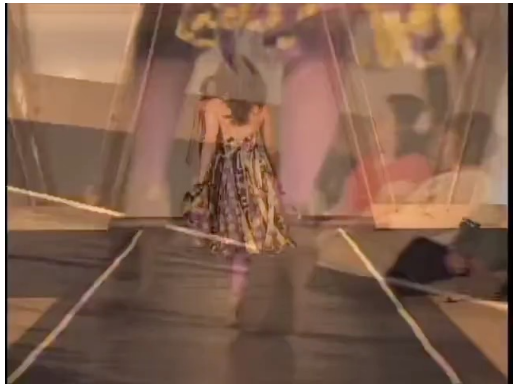
A type of cloth hanging on a rack, hanger, or line



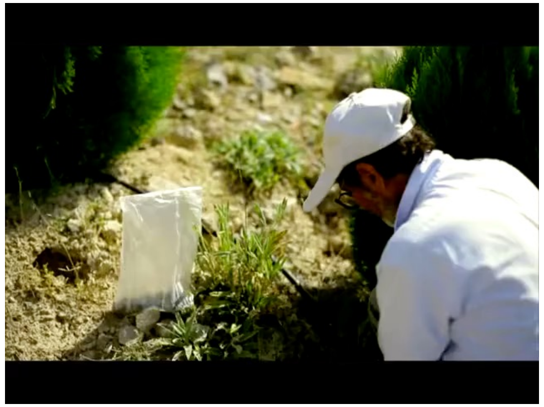
Building with columns during daytime



A female person bending downwards



A person is in the act of swinging



A kneeling man outdoors

2022 Main Approaches

- Use of multiple text-image / text-video common latent embeddings: VSE++, GSMN, CLIP, SLIP, …
- Use of multiple text-image / text-video annotated collections: MSR-VTT, TGIF, Flickr8k/30k, MS-COCO, Conceptual Captions, …
- Use of multiple visual and textual feature extractors
- Triplet loss with margin for embedding space learning
- Large number of combinations and fusion (normalization | averaging)
- Lightweight Attentional Feature Fusion
- Stacked Cross Attention Network
- Bidirectional Negation Learning (for query with negative cues)
- Dual SoftMax with “background queries”
- No more concept bank approaches but “dual task” (interpretable embeddings)
- Hard to distinguish between data / features effects and algorithmic effects

2022 Task Observations

- Submissions
  - 7 teams finished the main task including 6 teams submitting to the progress task with 28 runs.
  - 28 automatic systems and 5 manually-assisted systems submitted runs in 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
  - Below 2021 & 2020 in general. However, queries are different and meant to search for more fine grained information.
  - Few automatic systems are good and fast (< 10 sec).
  - High similarity between automatic and manually-assisted 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 36% of all hits are unique. 64% are common hits across the runs.
  - 13.5% 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 2023
29th International Conference on Multimedia Modeling,
January 2023, Bergen, Norway

- 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.
- Registration for the task is now closed