VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

NII-UIT at TRECVID 2023: Ad-hoc Video Search

[1] *University of Information Technology (UIT), Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City (VNU-HCM), Vietnam*
[3] *National Institute of Informatics (NII), Tokyo, Japan*

Overview

- AVS Task Introduction
- Challenges
- Our Approach
- Experimental Results

Introduction - Ad-hoc Video Search

A group of people are playing a football game

Input
A textual query

Output
A ranking list of videoID-frameID

Video-text retrieval examples on the MSR-VTT dataset. The red box indicates the item is retrieved correctly.

Dataset

Test set

V3C2
3.0TB
9,760
1300 hours, 52 minutes, 48 seconds
7 minutes, 59 seconds
1,425,454

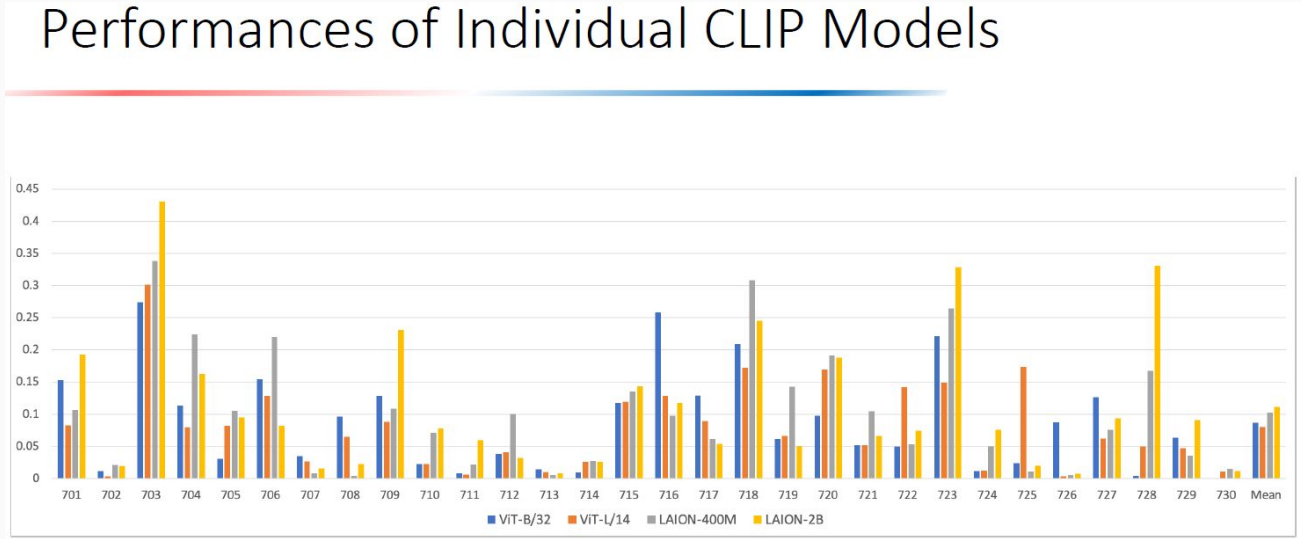
731 A man is seen with a baby
732 A woman with red hair
733 A golf course
734 A recording studio
735 A toy vehicle
736 A person opens a door and enters a location
737 A woman wearing (dark framed) glasses
738 A police officer wearing a helmet
739 Two or more persons are seen in front of a chain link fence
740 A heavy man indoors
741 A red or blue scarf around someone's neck
742 A child climbs an object outdoors
743 A man is talking in a small window located in the lower corner of the screen
744 A person taking picture using a cell phone camera
745 A person wearing gloves while biking
746 A man riding a scooter
747 At least two persons are working on their laptops together in the same room indoors.
748 A man carrying a bag on one of his shoulders (excluding backbags)
749 A person wearing any kind of face or head mask
750 A man with an earring in his left ear

Challenges

1. **Large-scale dataset**
   a. Resources constraints, Overlapped shots;
2. **Query Ambiguity:**
   a. Subjectively created by judges;
   b. Simple query: Lack of specificity, for example:
   Query 735 "a toy vehicle" ⇒ System can return a wide range of irrelevant results.
3. **Multimodal Data**
4. **Which tasks, which models?**

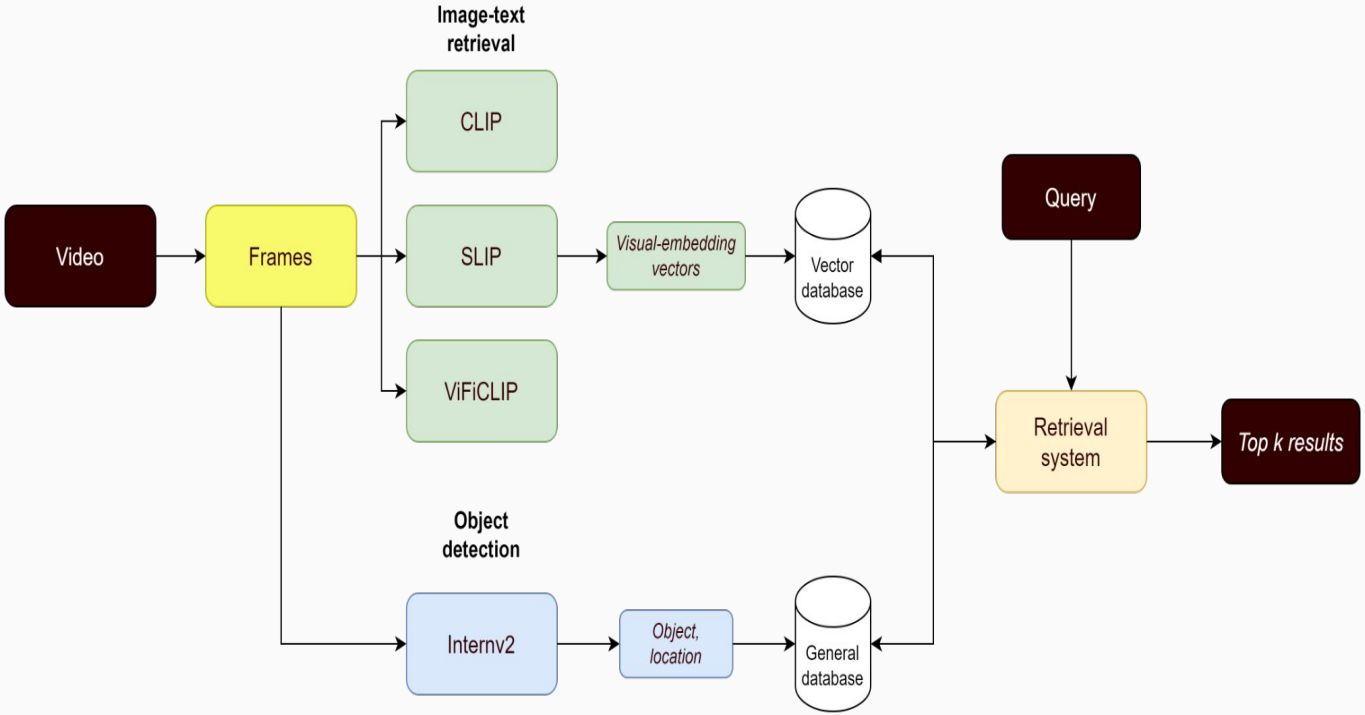
Results using the features used in our framework on 2022 query

CLIP B/32	CLIP L/14	CLIP L/14 DataComp	CLIP H/14 Laion2B	CLIP RN50x16	CLIP RN50x4	CLIP-RN101	SLIP base	SLIP small	BLIP	CLIP-bnl	CLIP-finetuned	XCLIP	ViFi-CLIP
0.0659	0.0607	0.0793	0.0953	0.0688	0.0672	0.0603	0.0362	0.0396	0.0513	0.0864	0.0815	0.0268	0.0135



https://www-nlpir.nist.gov/projects/tvpubs/tv22.slides/kindai_ogu_osaka.avs.slides.pdf

Proposed Methods



Full framework system

Late-fusion: fusion results on various models on 2022 query

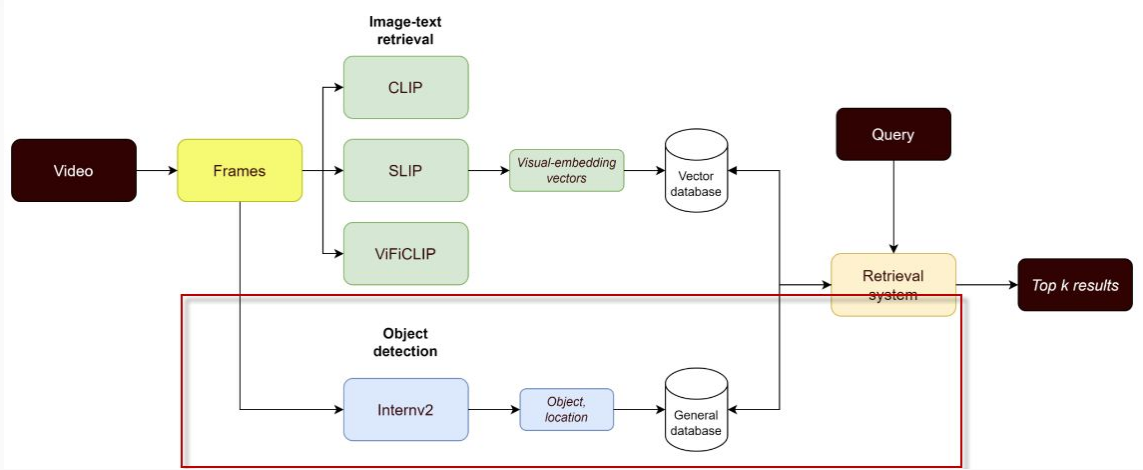
Selected text-retrieval model(s)				
CLIP (B/32)	x			
CLIP (L/14)	x	x	x	x
CLIP (L/14) DataComp	x	x	x	
CLIP (RN50x16)	x			x
CLIP (RN50x4)		x	x	x
BLIP (B/16)	x	x	x	x
CLIP-bnl	x	x	x	x
CLIP-finetuned	x	x	x	x
XCLIP		x	x	x
ViFi-CLIP	x	x		x
Fusing result (xinfAP)	0.1560	0.1547	0.1519	0.149

Table 2: Using the CombMNZ fusion method, the xinfAP scores on the Trecvid 2022 groundtruths are generated by combining the outputs of multiple text-retrieval models, with the selected models denoted by x.

Selected text-retrieval model(s)				
CLIP (L/14)	x	x	x	x
CLIP (L/14) DataComp	x			
CLIP (H/14) Laion2B	x			
CLIP (RN50x16)	x	x	x	x
CLIP (RN101)	x	x	x	x
SLIP (S/16)		x		x
BLIP (B/16)	x	x	x	x
CLIP-bnl	x	x	x	x
CLIP-finetuned	x	x	x	x
XCLIP	x	x	x	
ViFi-CLIP	x	x	x	x
Fusing result (xinfAP)	0.1705	0.1626	0.1624	0.1622

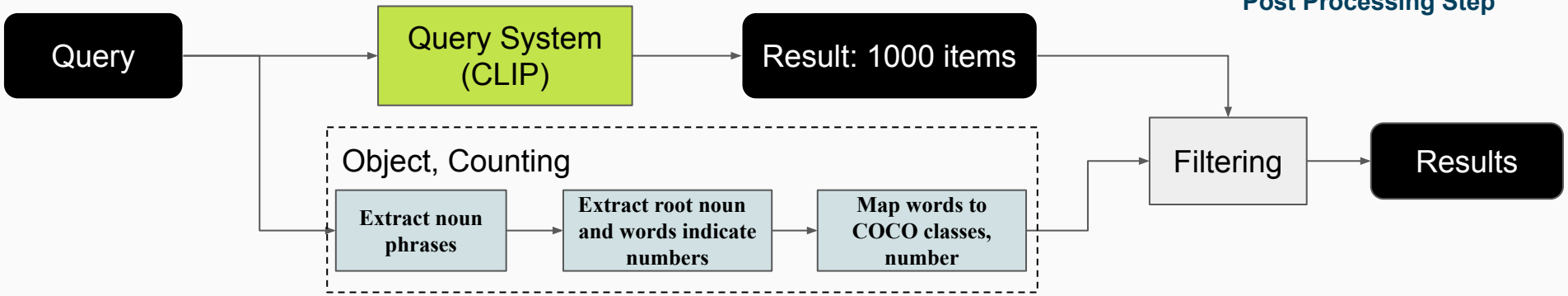
Table 3: Using the PosFuse fusion method, the xinfAP scores on the Trecvid 2022 groundtruths are generated by combining the outputs of multiple text-retrieval models, with the selected models denoted by x.

Enhancing Precision with Reranking



Full framework system

Post Processing Step



Submission

Fusion result using CombMNZ

Fusing result	CLIP B/32	CLIP L/14	CLIP L/14 DataComp	CLIP RN50x16	CLIP RN50x4	CLIP RN101	SLIP base (1)	BLIP (2)	CLIP-bnl (3)	CLIP-finetuned (3)	XCLIP (4)	ViFi-CLIP (5)
0.156		✓	✓	✓				✓	✓	✓		✓

⇒ Run 1: Fully automatic (F)

Fusion result using PosFuse

Fusing result	CLIP-L/14	CLIP-L/14 DataComp	CLIP-H/14 Laion2B	CLIP-RN 50x16	CLIP-RN101	SLIP base (1)	BLIP (2)	CLIP-bnl (3)	CLIP-finetuned (3)	XCLIP (4)	ViFi-CLIP (5)
0.1705	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓

⇒ Run 2: Fully automatic (F)

Submission

Fusion using CombMNZ + Object reranking

Fusing result	CLIP B/32	CLIP L/14	CLIP L/14 DataComp	CLIP RN50x16	CLIP RN50x4	CLIP RN101	SLIP base (1)	BLIP (2)	CLIP-bnl (3)	CLIP-finetuned (3)	XCLIP (4)	ViFi-CLIP (5)
0.1601		✓	✓	✓				✓	✓	✓		✓

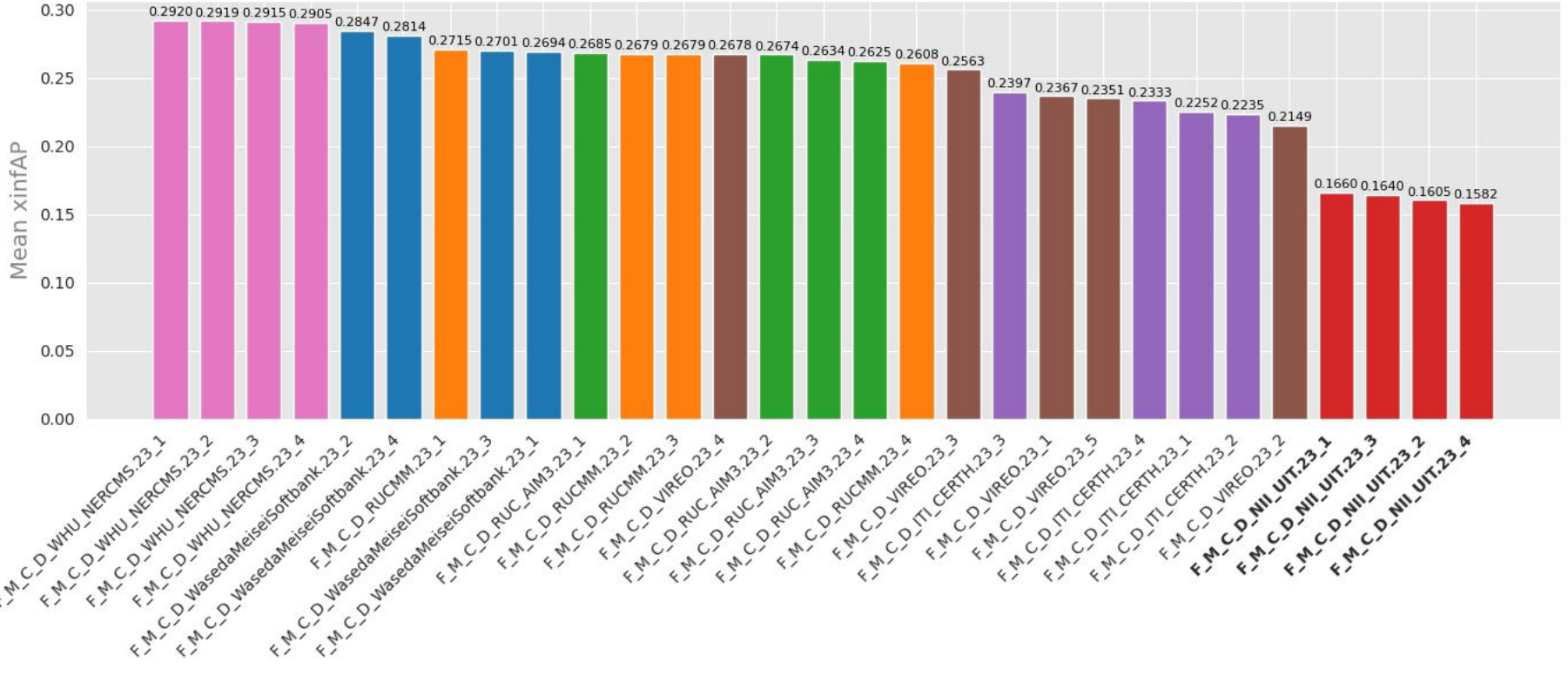
⇒ Run 3: Fully automatic (F)

Fusion using PosFuse + Object reranking

Fusing result	CLIP-L/14	CLIP-L/14 DataComp	CLIP-H/14 Laion2B	CLIP-RN 50x16	CLIP-RN101	SLIP base (1)	BLIP (2)	CLIP-bnl (3)	CLIP-finetuned (3)	XCLIP (4)	ViFi-CLIP (5)
0.1755	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓

⇒ Run 4: Fully automatic (F)

Submission results: Automatic

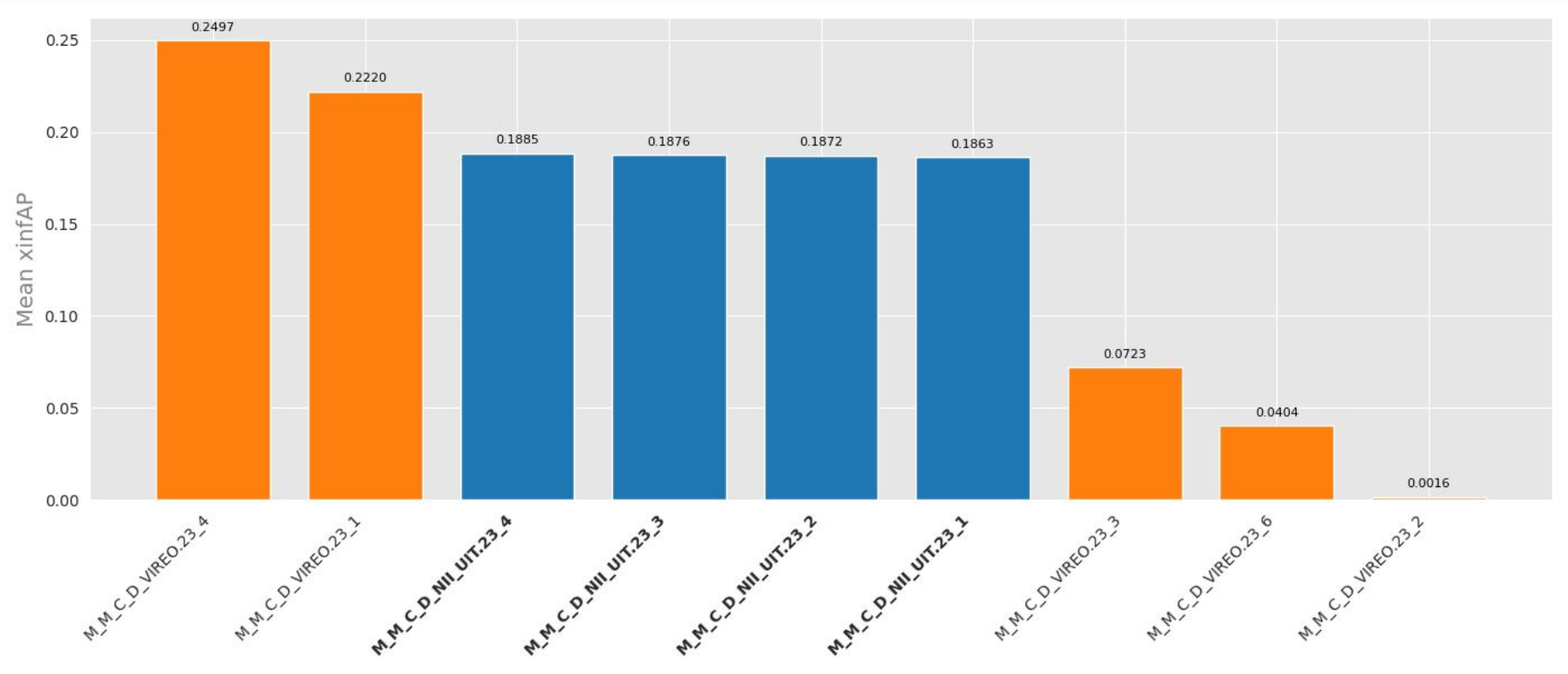


Query refined

Query id	Original query	Final
731	A man is seen with a baby	a baby and a man
732	A woman with red hair	A woman with red hair
733	A golf course	A golf course
734	A recording studio	A recording studio
735	A toy vehicle	A toy vehicle
736	A person opens a door and enters a location	a man entering an opened door
737	A woman wearing (dark framed) glasses	A woman wearing (dark framed) glasses
738	A police officer wearing a helmet	A police officer wearing a helmet
739	Two or more persons are seen in front of a chain link fence	Many people in front of a chain link fence
740	A heavy man indoors	A overweight man indoors
741	A red or blue scarf around someone's neck	a person wearing red or blue scarf
742	A child climbs an object outdoors	A child climbs an object outdoors
743	A man is talking in a small window located in the lower corner of the screen	a man is talking nearby a window which is in the bottom of the frame
744	A person taking picture using a cell phone camera	A person taking picture using a smartphone
745	A person wearing gloves while biking	A person wearing gloves while riding a bicycle
746	A man riding a scooter	A man riding a scooter
747	At least two persons are working on their laptops together in the same room indoors.	Many people are working with their laptop together in a room
748	A man carrying a bag on one of his shoulders (excluding backbags)	A man with a bag on one shoulder
749	A person wearing any kind of face or head mask	A person wearing face mask or head mask
750	A man with an earring in his left ear	A man with an earring in his left ear

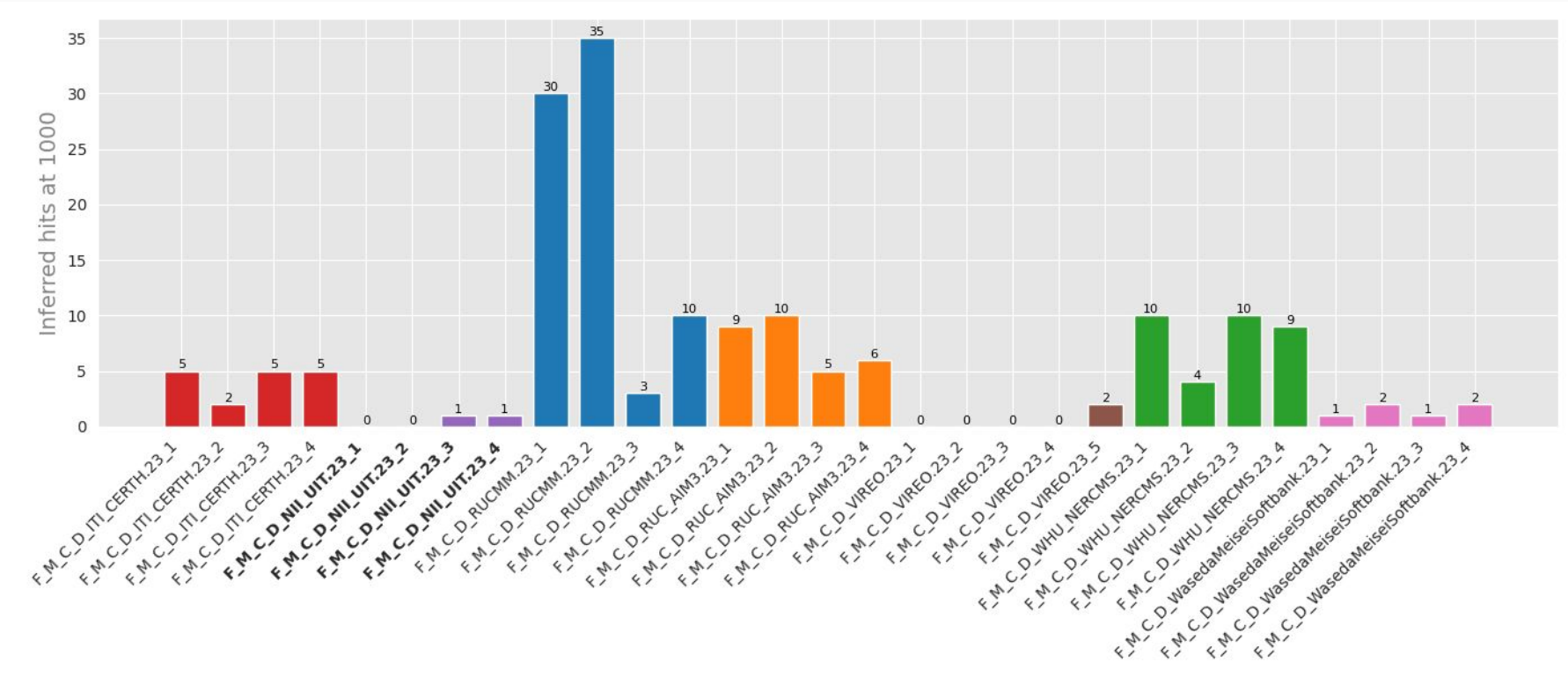
Table 4: This table show the original Trecvid 2023 queries and their respectively manually refined queries by our team.

Submission results: Manually



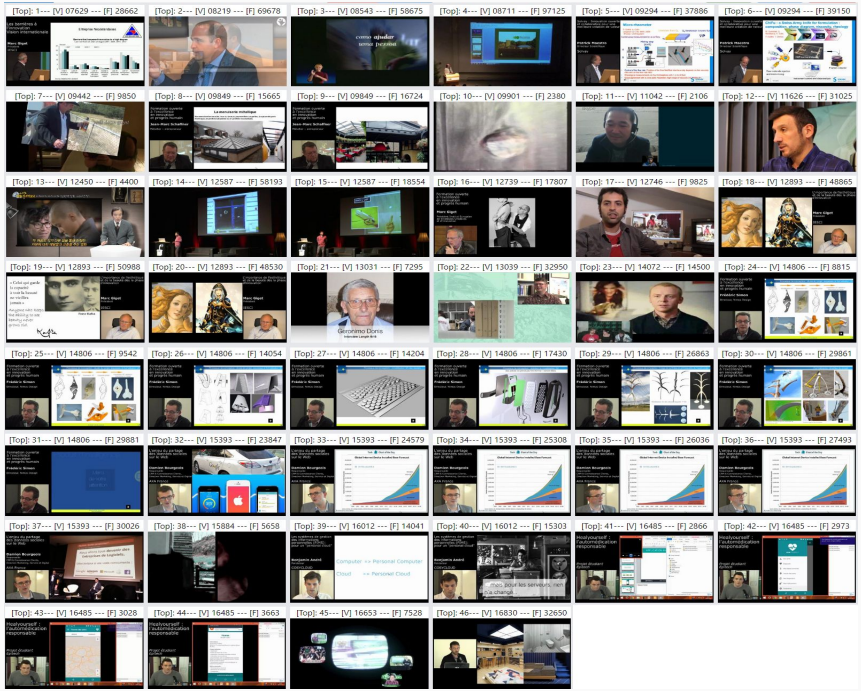
Analysis (query 743) - most team fail

Query 743: A man is talking in a small window located in the lower corner of the screen

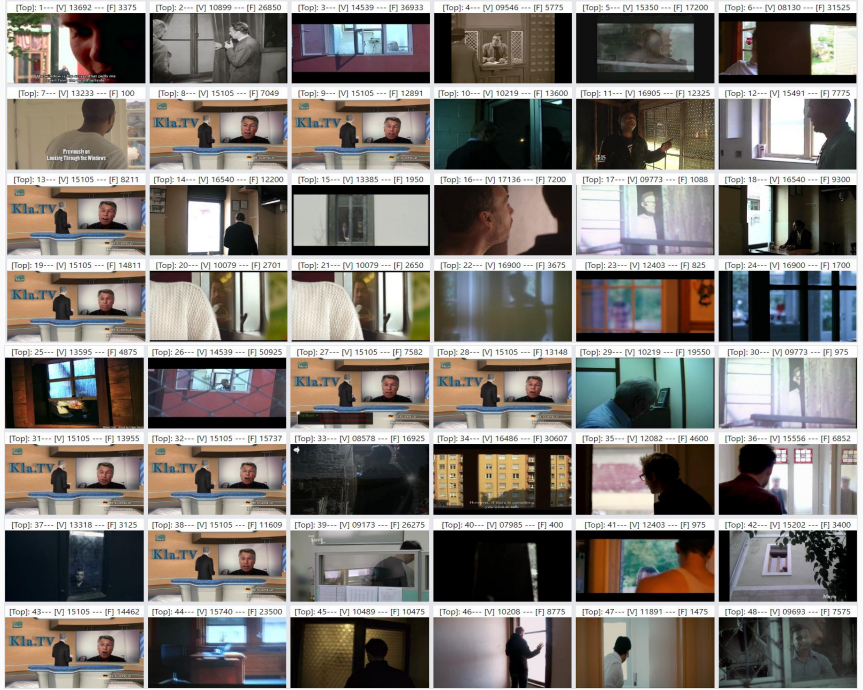


Analysis (query 743) - most team fail

Query 743: A man is talking in a small window located in the lower corner of the screen



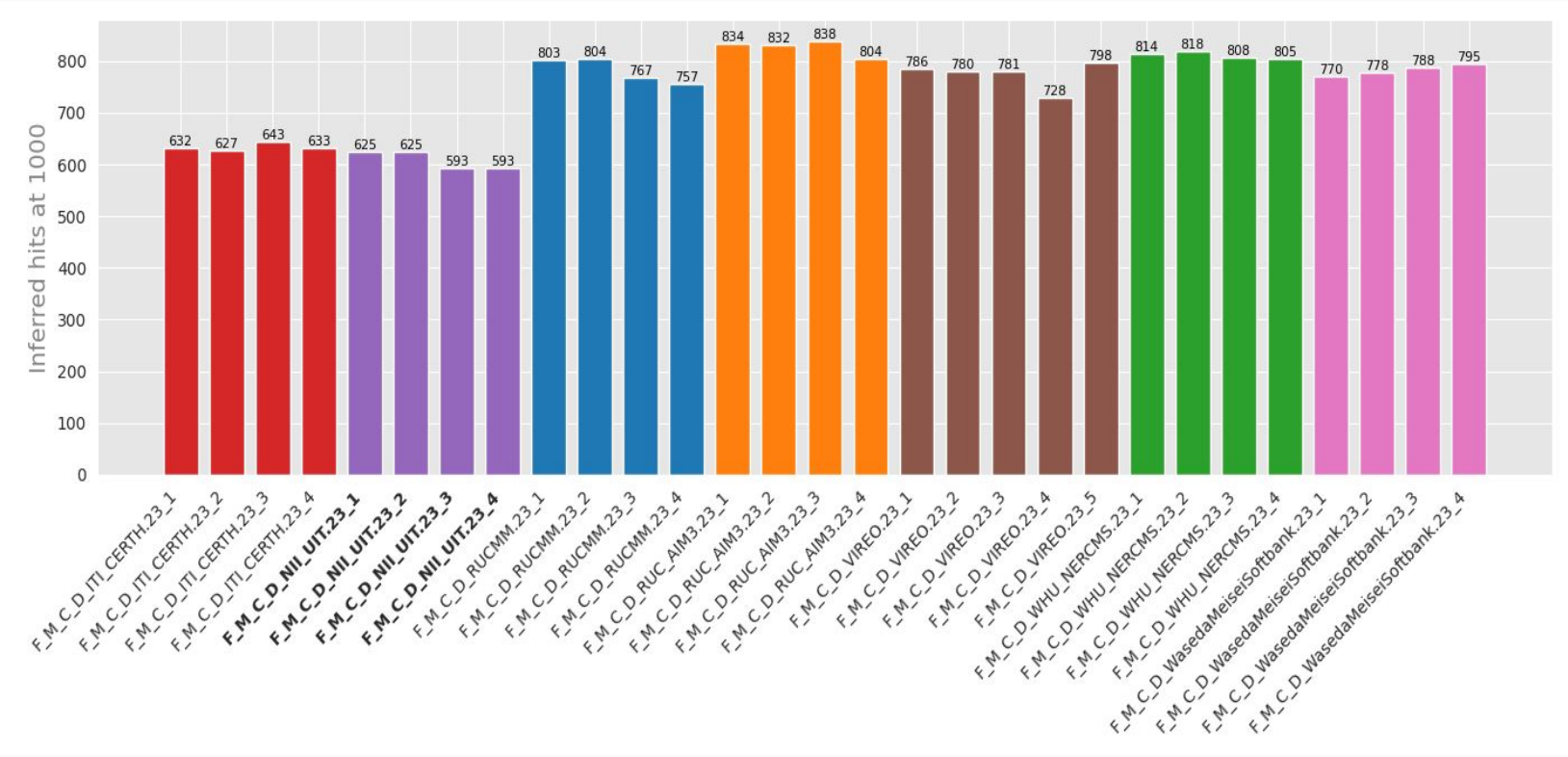
Ground Truth of query 743



(Ours) Submission on query 743 - Run 1

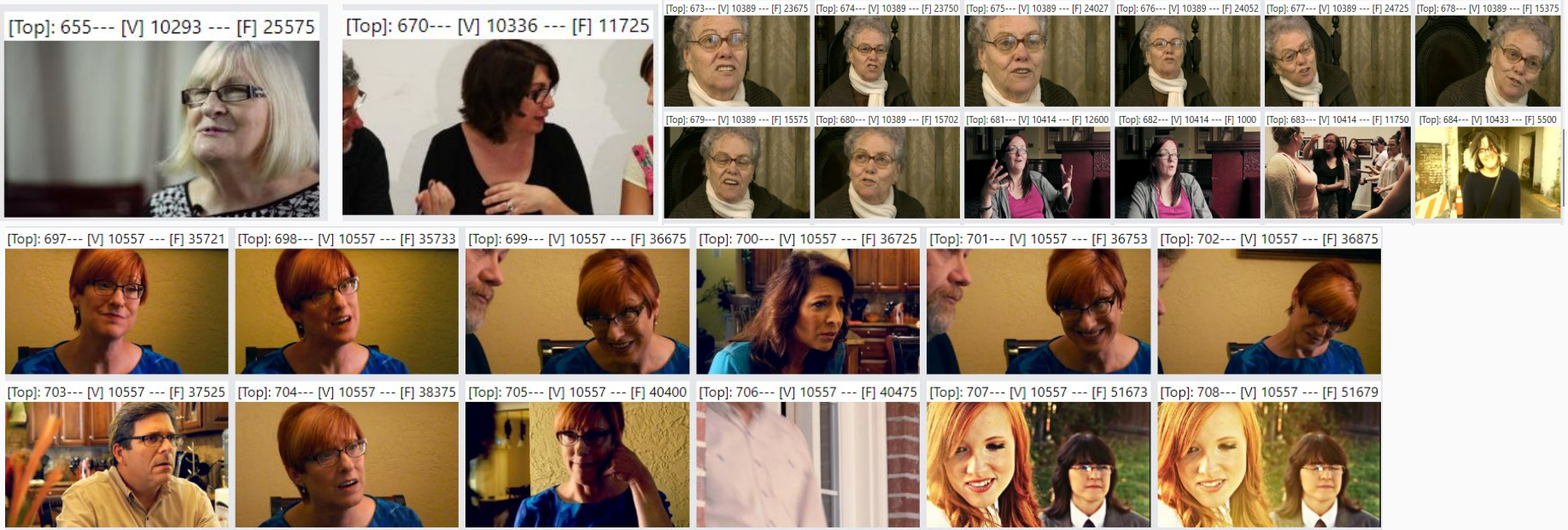
Analysis (query 737)

Query 737: A woman wearing (dark framed) glasses



Analysis (query 737)

Query 737: A woman wearing (dark framed) glasses

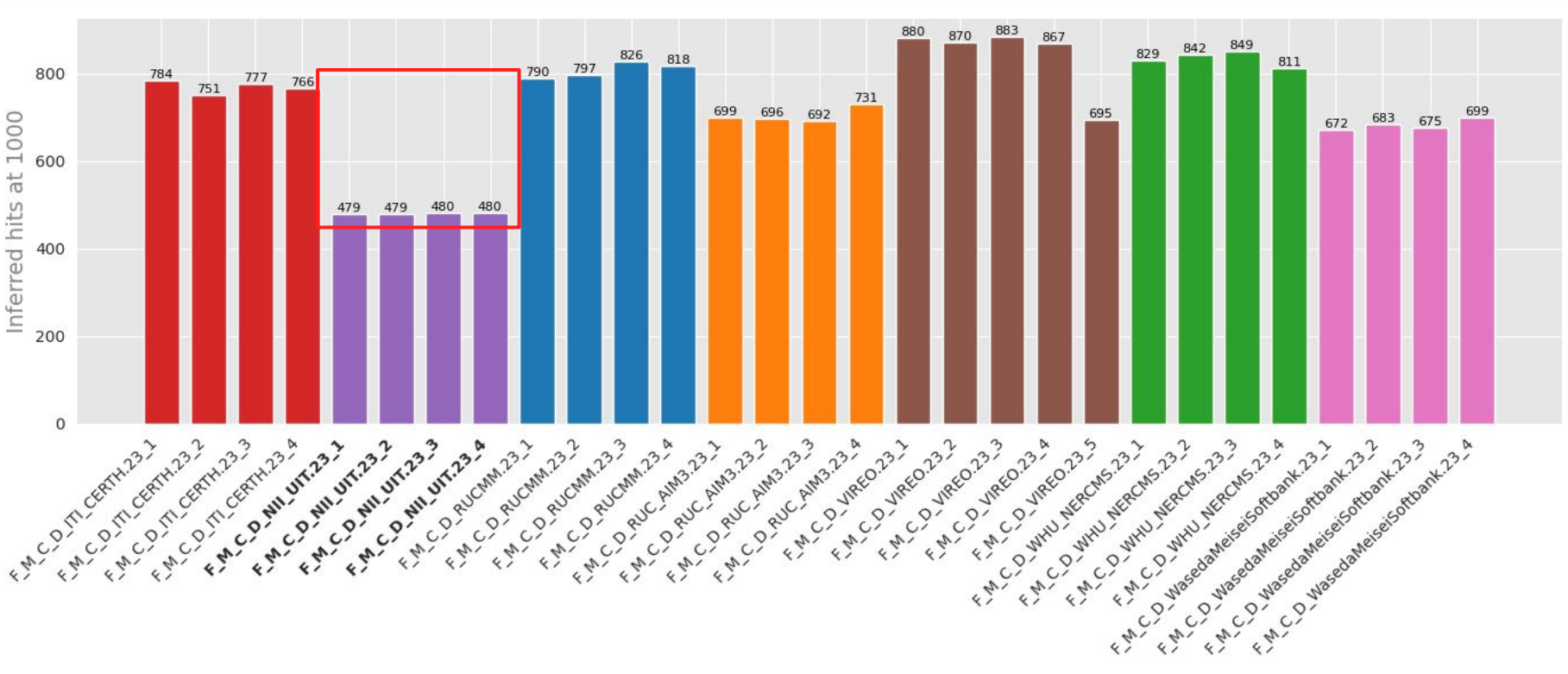


Ground Truth of query 737

Miss in our system → Reason: maybe the dark colour of the glasses was too hard to catch

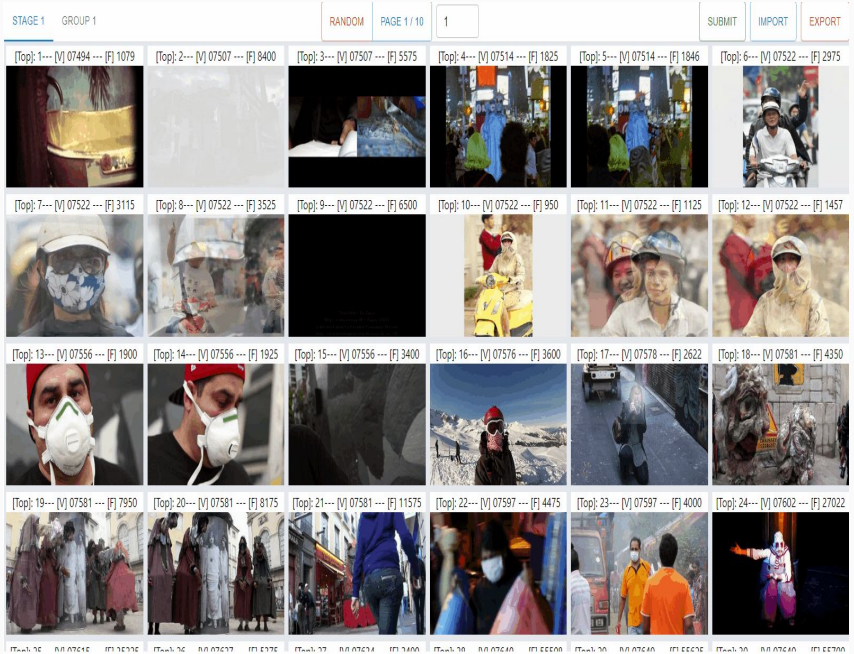
Analysis (query 749) - lowest score compare to others

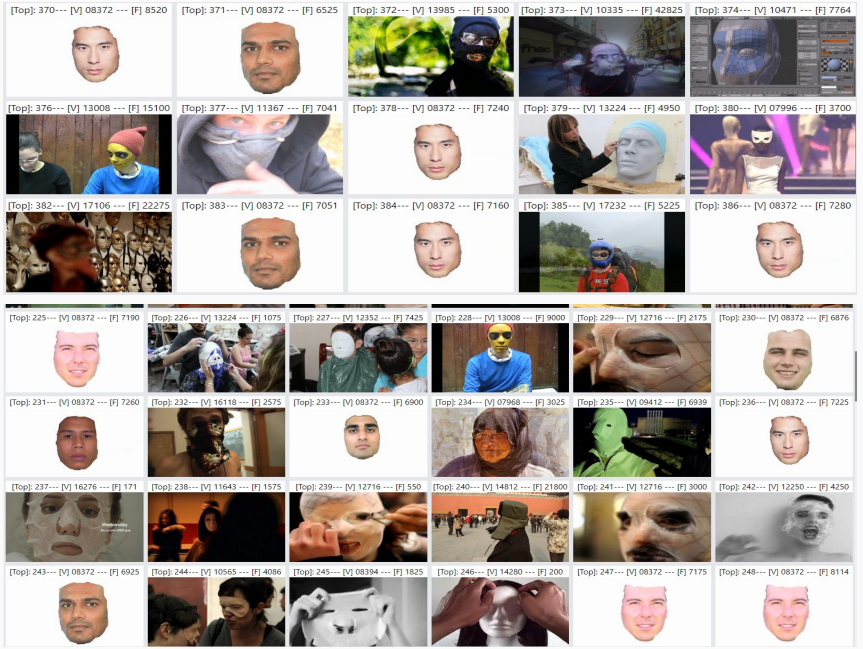
Query 749: A person wearing any kind of face or head mask



Analysis (query 749)

Query 749: A person wearing any kind of face or head mask



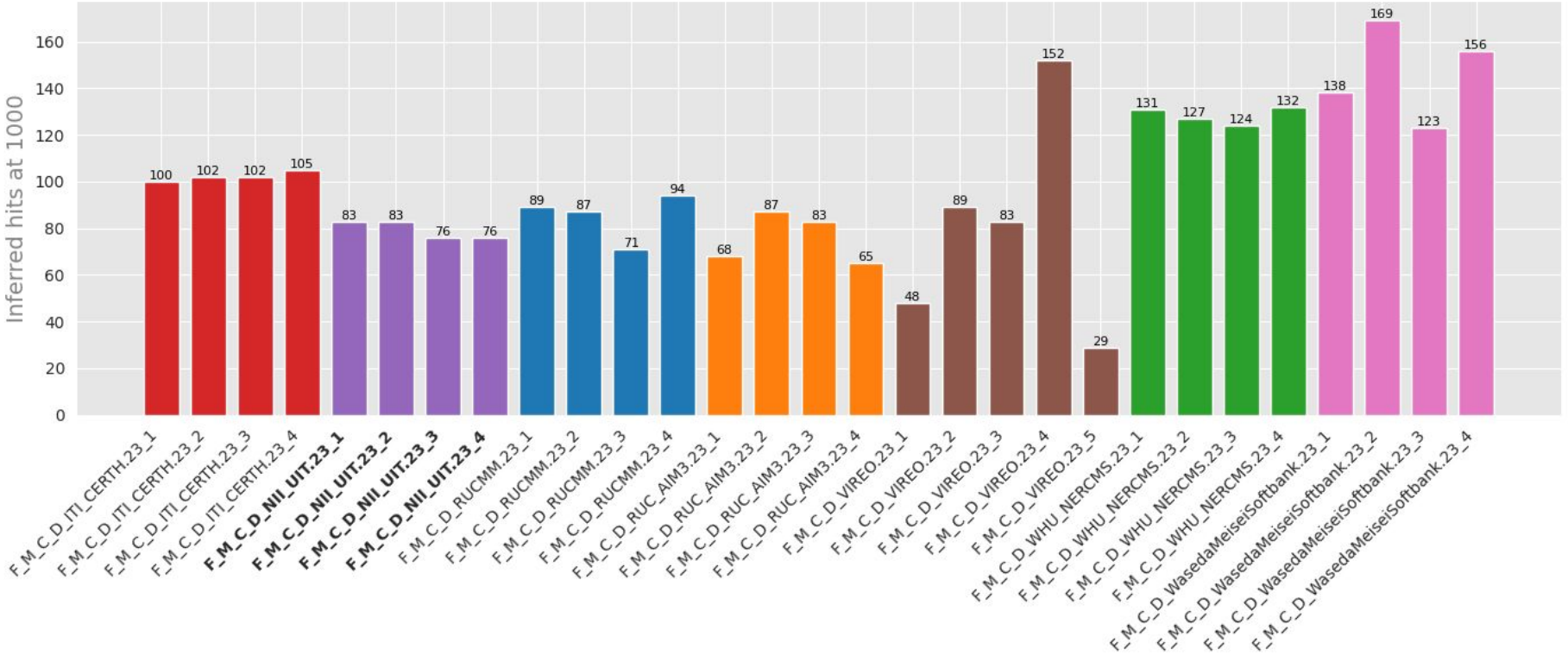


Our system misunderstand the "or" (it actually face mask or head mask)

(Ours) Submission on query 749 - Run 1

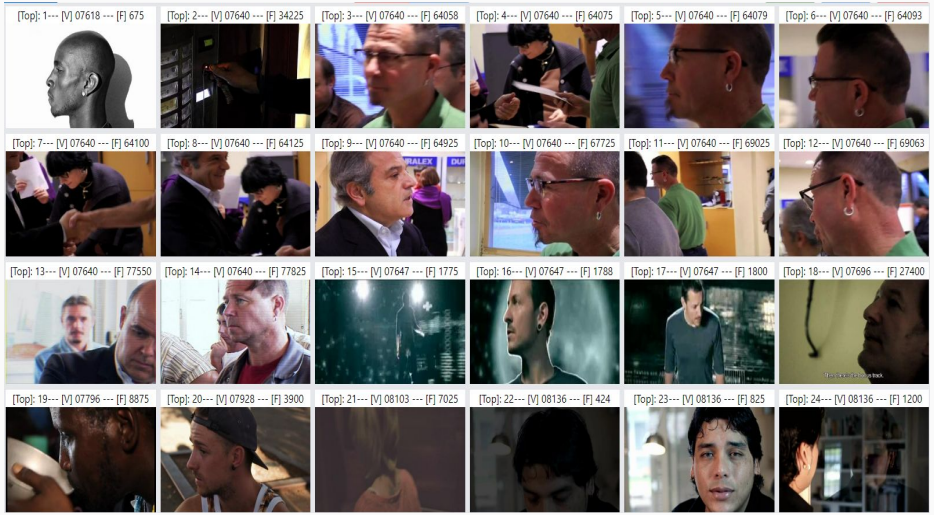
Analysis (query 750) - fail

query 750: A man with an earring in his **left ear**



Analysis (query 750) - fail

query 750: A man with an earring in his **left ear**

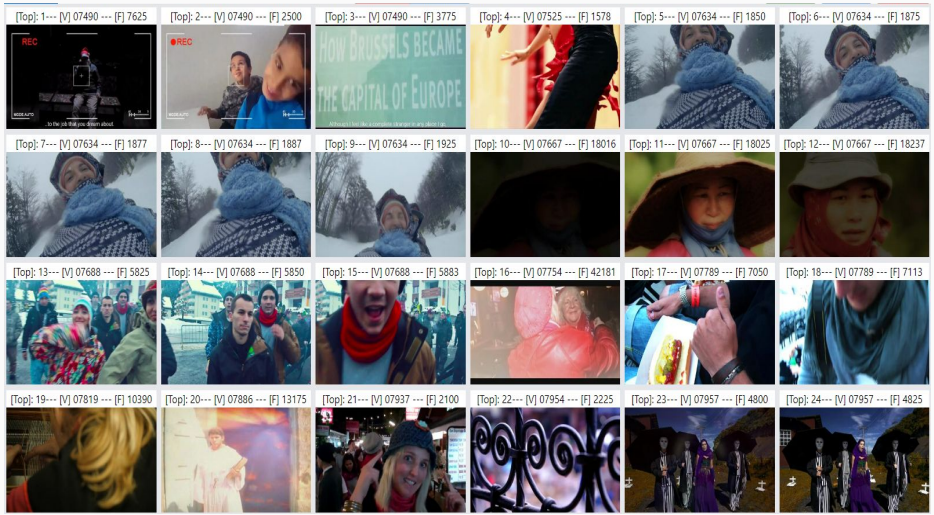


Ground Truth of query 750

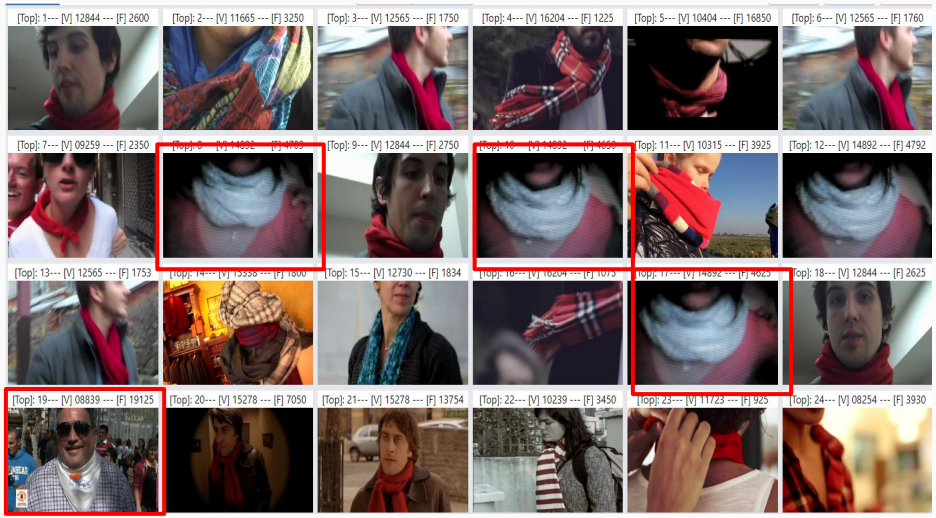
(Ours) Submission on query 750 - Run 1

Analysis (query 741) - fail

query 741: A red or blue scarf around someone's neck



Ground Truth of query 741



CERTAINLY NOT RED OR BLUE

(Ours) Submission on query 750 - Run 1

Impact of refined query: Analysis

	Query id	1731	1732	1733	1734	1735	1736	1737	1738	1739	1740	1741	1742	1743	1744	1745	1746	1747	1748	1749	1750	Sum of differences
Run Manually.1 - Run Automatic.1	hit@10	3	0	2	-1	0	0	-2	-1	0	6	-2	-1	0	-2	0	1	0	-1	-3	1	0
	hit@30	2	-4	-4	2	-2	3	0	0	-3	12	-2	-4	0	-4	2	3	6	0	-1	1	7
	hit@100	1	0	-1	-1	-7	5	0	10	-12	32	-1	-2	0	-2	2	9	10	8	14	-2	63
	hit@1000	2	-7	0	13	-13	23	-17	17	-22	155	35	-5	0	21	-35	26	104	-2	100	-4	391
Run Manually.2 - Run Automatic.2	hit@10	3	0	2	-1	0	0	-2	-1	1	7	-2	-1	0	-2	3	1	0	-1	-2	1	6
	hit@30	0	-4	-4	2	-2	1	0	0	0	12	-2	-2	0	-3	5	3	12	0	-2	1	17
	hit@100	-1	0	-1	-1	-7	8	2	10	2	31	-1	-3	0	2	4	10	29	9	14	-3	104
	hit@1000	2	-7	0	13	-13	23	-17	17	-22	155	35	-5	0	21	-35	26	104	-2	100	-4	391
Run Manually.3 - Run Automatic.3	hit@10	-1	0	-1	0	2	-1	1	-1	2	5	-1	3	0	1	0	0	3	2	-1	0	13
	hit@30	-3	4	1	-1	2	-2	-2	-3	-5	15	-1	6	0	5	1	0	10	4	-1	-1	29
	hit@100	2	6	-4	-5	-5	3	-8	-7	-16	27	4	4	0	11	5	0	0	-4	11	2	26
	hit@1000	4	-6	-7	-24	-13	13	42	-15	-26	96	39	3	-1	17	-18	-5	117	14	124	18	372
Run Manually.4 - Run Automatic.4	hit@10	-1	0	-1	0	2	-1	1	-1	3	5	-1	2	0	1	2	0	7	2	-1	0	19
	hit@30	-1	4	1	-1	2	0	0	-3	2	15	-1	6	0	4	3	0	15	4	-1	-1	48
	hit@100	4	6	-4	-5	-5	7	-4	-7	1	28	5	2	0	13	9	1	27	-2	14	2	92
	hit@1000	4	-6	-7	-24	-13	13	42	-15	-26	96	39	3	-1	17	-18	-5	117	14	124	18	372

Table 5: The provided table illustrates the variation in the number of hits at different cutoff levels (10, 30, 100, 1000). The column headings in the table represent abbreviations for each submission run, where "Manually" corresponds to *M_M_C_D*, and "Automatic" corresponds to *F_M_C_D*.

Impact of refined query: Query 747

	Query id	1731	1732	1733	1734	1735	1736	1737	1738	1739	1740	1741	1742	1743	1744	1745	1746	1747	1748	1749	1750	Sum of differences
Run Manually.1 - Run Automatic.1	hit@10	3	0	2	-1	0	0	-2	-1	0	6	-2	-1	0	-2	0	1	0	-1	-3	1	0
	hit@30	2	-4	-4	2	-2	3	0	0	-3	12	-2	-4	0	-4	2	3	6	0	-1	1	7
	hit@100	1	0	-1	-1	-7	5	0	10	-12	32	-1	-2	0	-2	2	9	10	8	14	-2	63
	hit@1000	2	-7	0	13	-13	23	-17	17	-22	155	35	-5	0	21	-35	20	104	-2	100	-4	391
Run Manually.2 - Run Automatic.2	hit@10	3	0	2	-1	0	0	-2	-1	1	7	-2	-1	0	-2	3	1	0	-1	-2	1	6
	hit@30	0	-4	-4	2	-2	1	0	0	0	12	-2	-2	0	-3	5	3	12	0	-2	1	17
	hit@100	-1	0	-1	-1	-7	8	2	10	2	31	-1	-3	0	2	4	10	29	9	14	-3	104
	hit@1000	2	-7	0	13	-13	23	-17	17	-22	155	35	-5	0	21	-35	20	104	-2	100	-4	391
Run Manually.3 - Run Automatic.3	hit@10	-1	0	-1	0	2	-1	1	-1	2	5	-1	3	0	1	0	0	3	2	-1	0	13
	hit@30	-3	4	1	-1	2	-2	-2	-3	-5	15	-1	6	0	5	1	0	10	4	-1	-1	29
	hit@100	2	6	-4	-5	-5	3	-8	-7	-16	27	4	4	0	11	5	0	0	-4	11	2	26
	hit@1000	4	-6	-7	-24	-13	13	42	-15	-26	96	39	3	-1	17	-18	-3	117	14	124	18	372
Run Manually.4 - Run Automatic.4	hit@10	-1	0	-1	0	2	-1	1	-1	3	5	-1	2	0	1	2	0	7	2	-1	0	19
	hit@30	-1	4	1	-1	2	0	0	-3	2	15	-1	6	0	4	3	0	15	4	-1	-1	48
	hit@100	4	6	-4	-5	-5	7	-4	-7	1	28	5	2	0	13	9	1	27	-2	14	2	92
	hit@1000	4	-6	-7	-24	-13	13	42	-15	-26	96	39	3	-1	17	-18	-3	117	14	124	18	372

Table 5: The provided table illustrates the variation in the number of hits at different cutoff levels (10, 30, 100, 1000). The column headings in the table represent abbreviations for each submission run, where "Manually" corresponds to *M_M_C_D*, and "Automatic" corresponds to *F_M_C_D*.

Original: "***At least two*** *people are working with their laptops together in a room*"

Refined: "***Many*** *people are working with their laptops together in a room*"

Impact of refined query: Query 740

	Query id	1731	1732	1733	1734	1735	1736	1737	1738	1739	1740	1741	1742	1743	1744	1745	1746	1747	1748	1749	1750	Sum of differences
Run Manually.1 - Run Automatic.1	hit@10	3	0	2	-1	0	0	-2	-1	0	6	-2	-1	0	-2	0	1	0	-1	-3	1	0
	hit@30	2	-4	-4	2	-2	3	0	0	-3	12	-2	-4	0	-4	2	3	6	0	-1	1	7
	hit@100	1	0	-1	-1	-7	5	0	10	-12	32	-1	-2	0	-2	2	9	10	8	14	-2	63
	hit@1000	2	-7	0	13	-13	23	-17	17	-22	155	35	-5	0	21	-35	26	104	-2	100	-4	391
Run Manually.2 - Run Automatic.2	hit@10	3	0	2	-1	0	0	-2	-1	1	7	-2	-1	0	-2	3	1	0	-1	-2	1	6
	hit@30	0	-4	-4	2	-2	1	0	0	0	12	-2	-2	0	-3	5	3	12	0	-2	1	17
	hit@100	-1	0	-1	-1	-7	8	2	10	2	31	-1	-3	0	2	4	10	29	9	14	-3	104
	hit@1000	2	-7	0	13	-13	23	-17	17	-22	155	35	-5	0	21	-35	26	104	-2	100	-4	391
Run Manually.3 - Run Automatic.3	hit@10	-1	0	-1	0	2	-1	1	-1	2	5	-1	3	0	1	0	0	3	2	-1	0	13
	hit@30	-3	4	1	-1	2	-2	-2	-3	-5	15	-1	6	0	5	1	0	10	4	-1	-1	29
	hit@100	2	6	-4	-5	-5	3	-8	-7	-16	27	4	4	0	11	5	0	0	-4	11	2	26
	hit@1000	4	-6	-7	-24	-13	13	42	-15	-26	96	39	3	-1	17	-18	-5	117	14	124	18	372
Run Manually.4 - Run Automatic.4	hit@10	-1	0	-1	0	2	-1	1	-1	3	5	-1	2	0	1	2	0	7	2	-1	0	19
	hit@30	-1	4	1	-1	2	0	0	-3	2	15	-1	6	0	4	3	0	15	4	-1	-1	48
	hit@100	4	6	-4	-5	-5	7	-4	-7	1	28	5	2	0	13	9	1	27	-2	14	2	92
	hit@1000	4	-6	-7	-24	-13	13	42	-15	-26	96	39	3	-1	17	-18	-5	117	14	124	18	372

Table 5: The provided table illustrates the variation in the number of hits at different cutoff levels (10, 30, 100, 1000). The column headings in the table represent abbreviations for each submission run, where "Manually" corresponds to *M_M_C_D*, and "Automatic" corresponds to *F_M_C_D*.

Original: "*A **heavy** man indoors*"

Refined: "*A **overweight** man indoors*"

Thanks for listening!