TRECVID 2023:
Video to Text Description

Asad Anwar Butt
NIST; Johns Hopkins University

George Awad
NIST

Yvette Graham
Trinity College Dublin

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

Goals and Motivation

- Measure how well an automatic system can describe a video in natural language.
- Measure how well an automatic system can match high-level textual descriptions to low-level computer vision features.
- Transfer successful image captioning technology to the video domain.
- Real world applications
  - Video summarization
  - Supporting search and browsing
  - Accessibility - video description to the blind
  - Video event prediction

System Task

Description Generation System

An orange Car is racing on the road.

Description Generation:
Automatically generate a text description for a given video.

Robustness Subtask:
Generate text descriptions after introducing noise in audio/visual channels.

Test Dataset

- VTT tasks from 2016 to 2019 used the Twitter Vines dataset.
  - Videos were ~6 sec long
  - Quality control issues
  - Links distributed instead of videos, leading to problem of removed links.
- Flickr videos are added in 2019.
- Dataset from 2020 onwards: V3C
  - The Vimeo Creative Commons Collection (V3C) is divided into 3 partitions.
  - Total duration: 3800+ hours.
  - V3C1 duration: 1000+ hours. Divided into more than 1 M segments. Only segments between 3 to 15 sec selected for this task.
  - Videos distributed directly to participants.

Test Dataset

NIST

- Manual selection of videos.
  - Watched 8000 videos.
  - Selected 2000 videos for annotation.
  - Subset of 300 videos were selected in 2021 to measure system progress over 3 years.
- Selection criteria mainly focused on diversity in videos.
- The V3C dataset removes some previous concerns:
  - Videos with multiple, unrelated segments that are not coherent.
  - Offensive videos.

Annotation Process

- A total of 5 assessors annotated the videos.
- Each video was annotated 5 times.
- Assessors were provided with training & annotation guidelines by NIST.
- For each video, assessors were asked to combine 4 facets if applicable:
  - Who is the video showing (objects, persons, animals, …etc) ?
  - What are the objects and beings doing (actions, states, events, …etc)?
  - Where (locale, site, place, geographic, ...etc) ?
  - When (time of day, season, ...etc) ?
- Their work was monitored, and feedback provided.
- NIST personnel were available for any questions or confusion.
- Our annotation process differentiates our dataset from other datasets.
  - Human annotators are hired & trained in-house (no crowd workers)
  - Annotators tend to provide more details

Annotation – Observations

NIST

- Average sentence length for each assessor:

Annotator	Avg. Length	# Videos
1	20.64	2000
2	20.48	2000
3	28.86	2000
4	29.38	2000
5	23.43	2000

Avg. sentence length: 24.56 words

- Additional questions:

Please rate how difficult it was to describe the video.

Very Easy (1) Easy (2) Medium (3) Hard (4) Very Hard (5)

How likely is it that other assessors will write similar descriptions for the video?

Not Likely (1) Somewhat Likely (2) Very Likely (3)

Q1 Avg Score: 2.22 (Scale of 5)

Q2 Avg Score: 2.52 (Scale of 3)

Correlation between difficulty scores: -0.53

Participants

Teams	Organization
KSLAB	Nagaoka University of Technology
MLVC_HDU	Hangzhou Dianzi University
RUC_AIM3	Renmin University of China
WasedaMeiseiSoftbank	Waseda University, Meisei University, SoftBank Corporation
BUPT_MCPRL	Beijing University of Posts and Telecommunications

- 5 teams participated with 25 runs
- 2 teams joined the robustness sub-task

Runs & Metrics

- Up to 4 runs per team
- Metrics used for evaluation:
  - CIDEr (Consensus-based Image Description Evaluation)
  - SPICE (Semantic Propositional Image Caption Evaluation)
  - METEOR (Metric for Evaluation of Translation with Explicit Ordering)
  - BLEU (BiLingual Evaluation Understudy)
  - STS (Semantic Textual Similarity)
  - DA (Direct Assessment), which is a crowdsourced rating of captions using Amazon Mechanical Turk (AMT)

Run Types

Training Data Types:

- '**I**': Only image captioning datasets
- '**V**': Only video captioning datasets
- '**B**': Both image and video captioning datasets

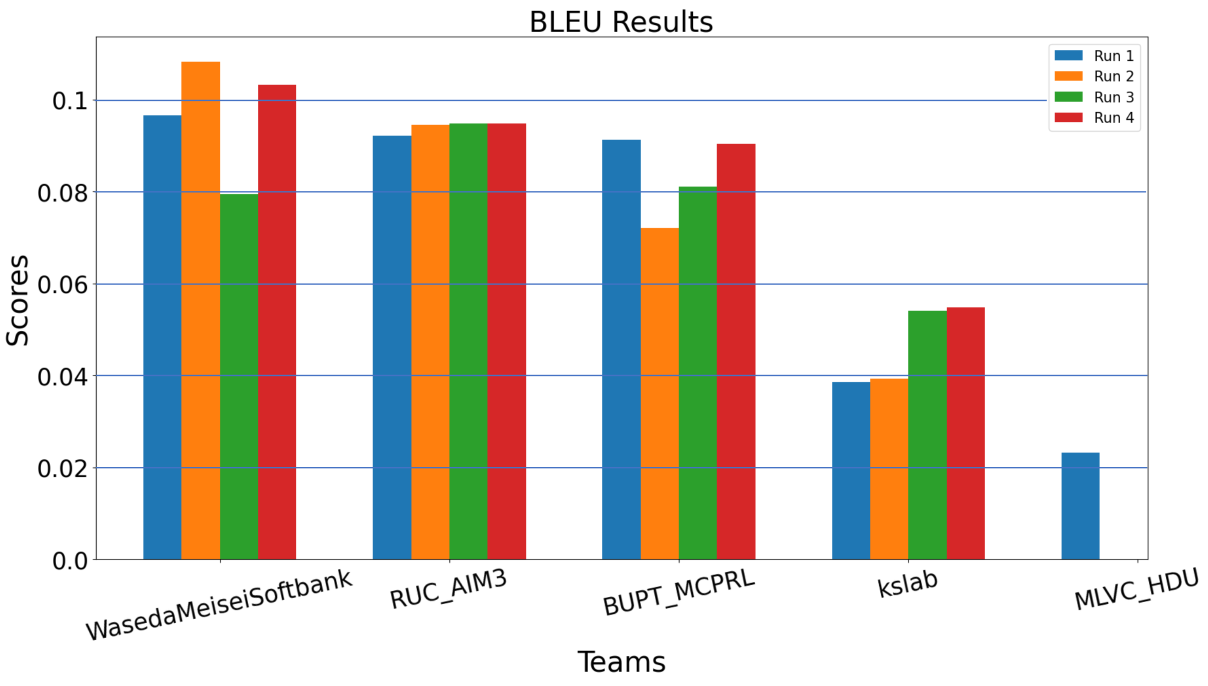
Features Used:

- '**V**': Visual features only
- '**A**': Both audio and visual features

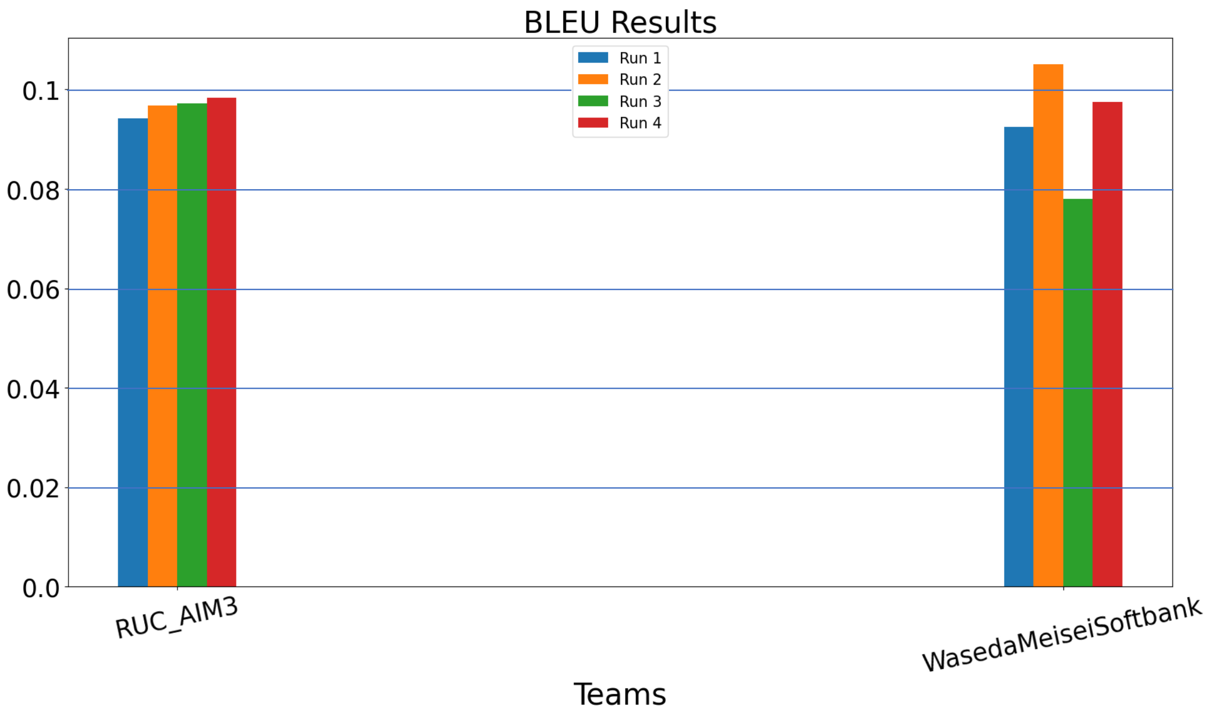
Submissions - Run Types

1 ‘VV’ (Video Data/Visual Feats) • 5 runs	**2 ‘IV’ (Image Data/Visual Feats)** • 2 runs
3 ‘BV’ (I+V Data/Visual Feats) • 16 runs	**4 ‘IA’ (Image Data/V+A Feats)** • 2 runs

Results

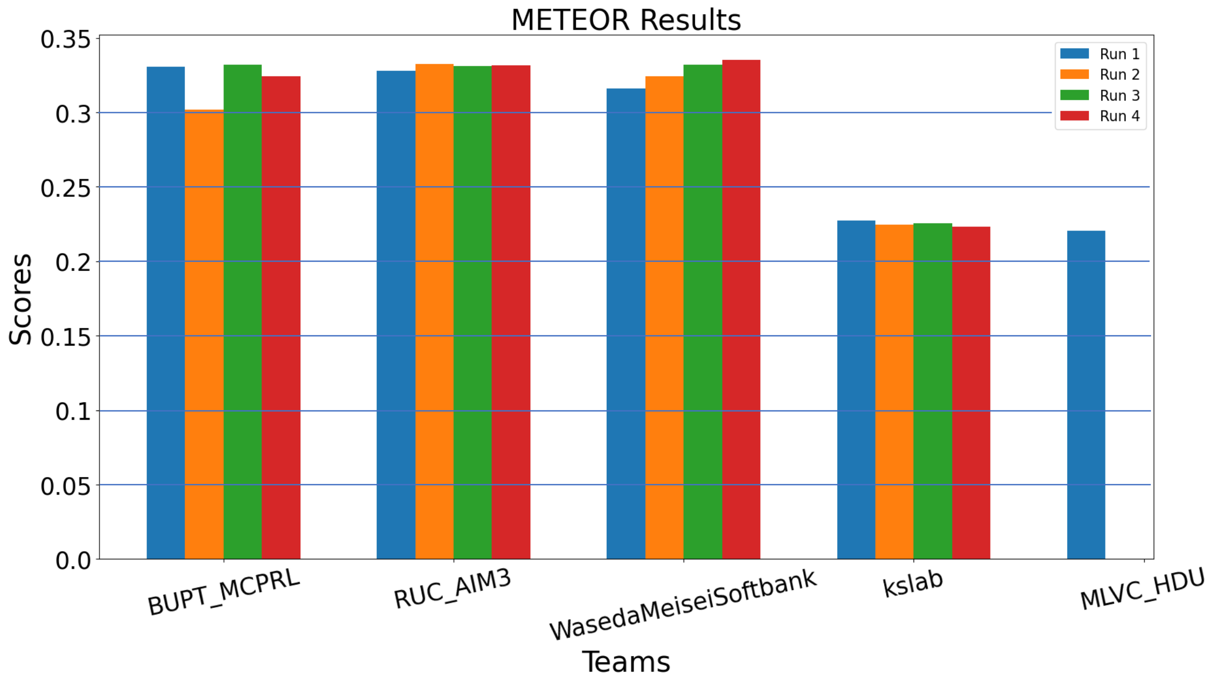


Main Task

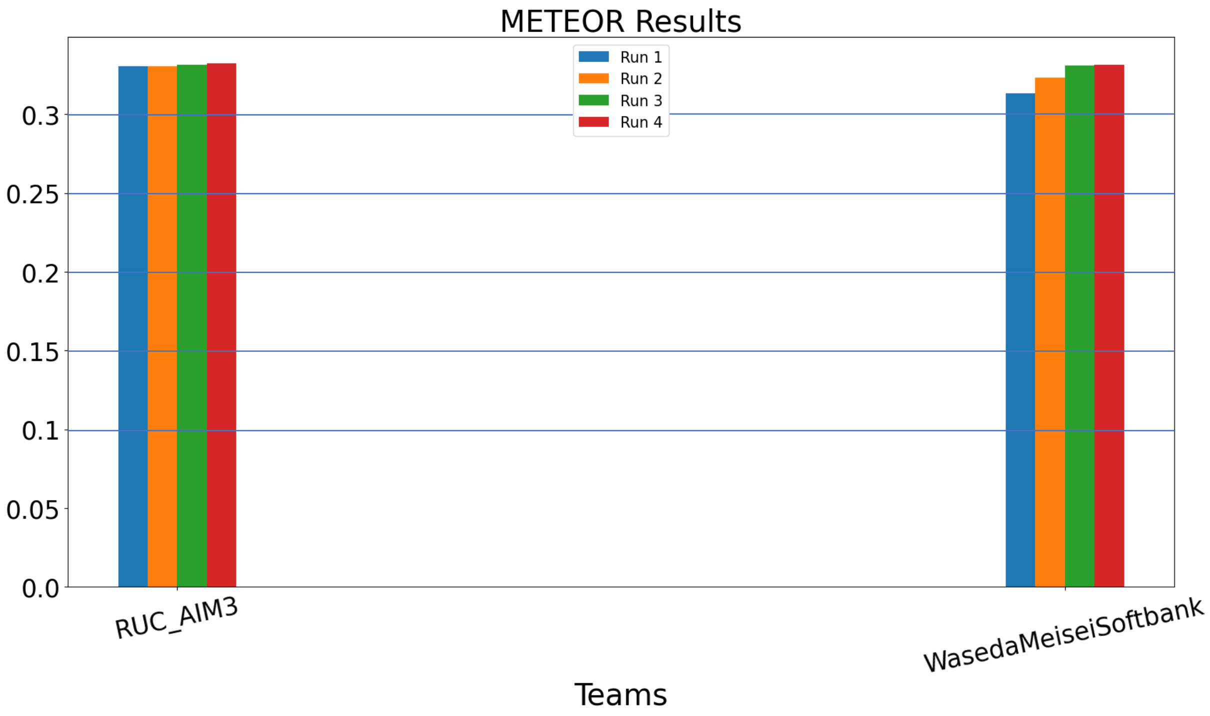


Robustness Task

Results

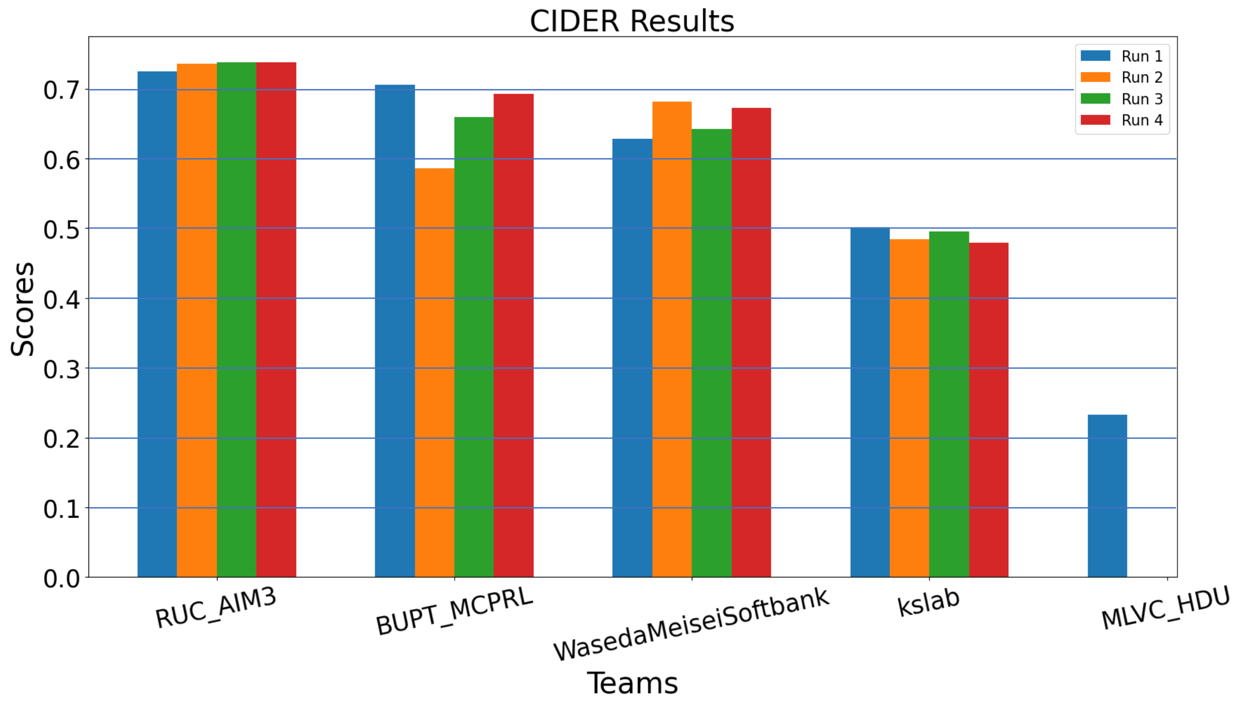


Main Task

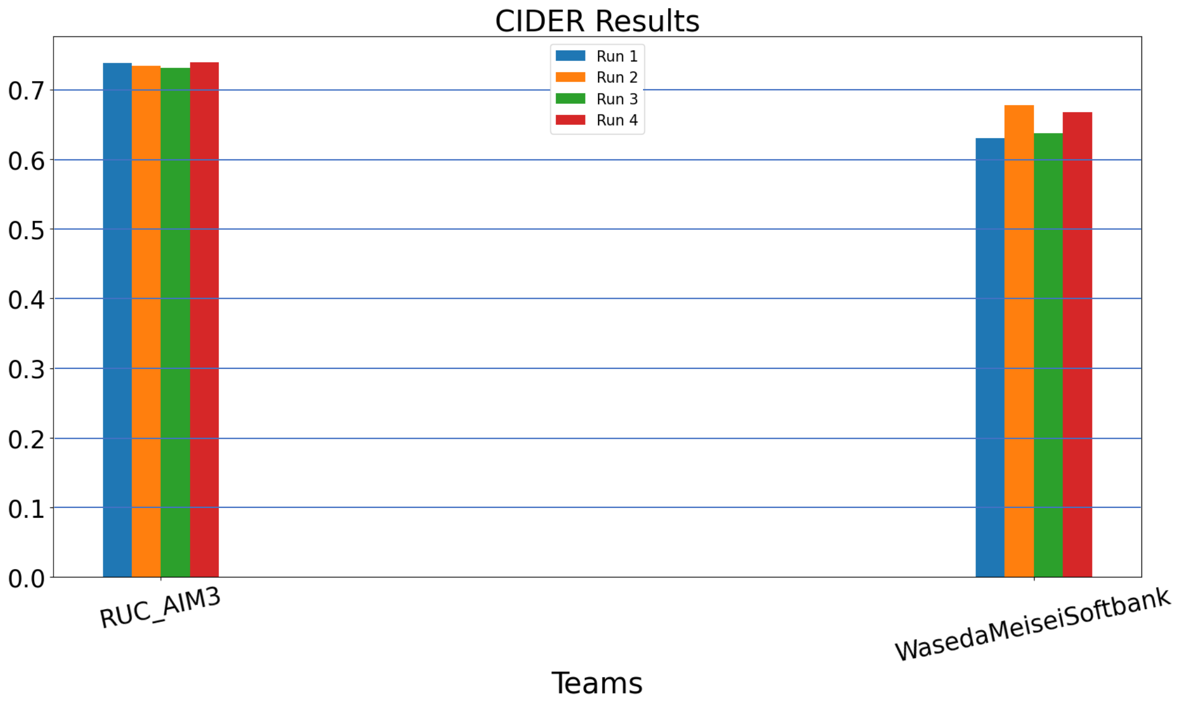


Robustness Task

Results

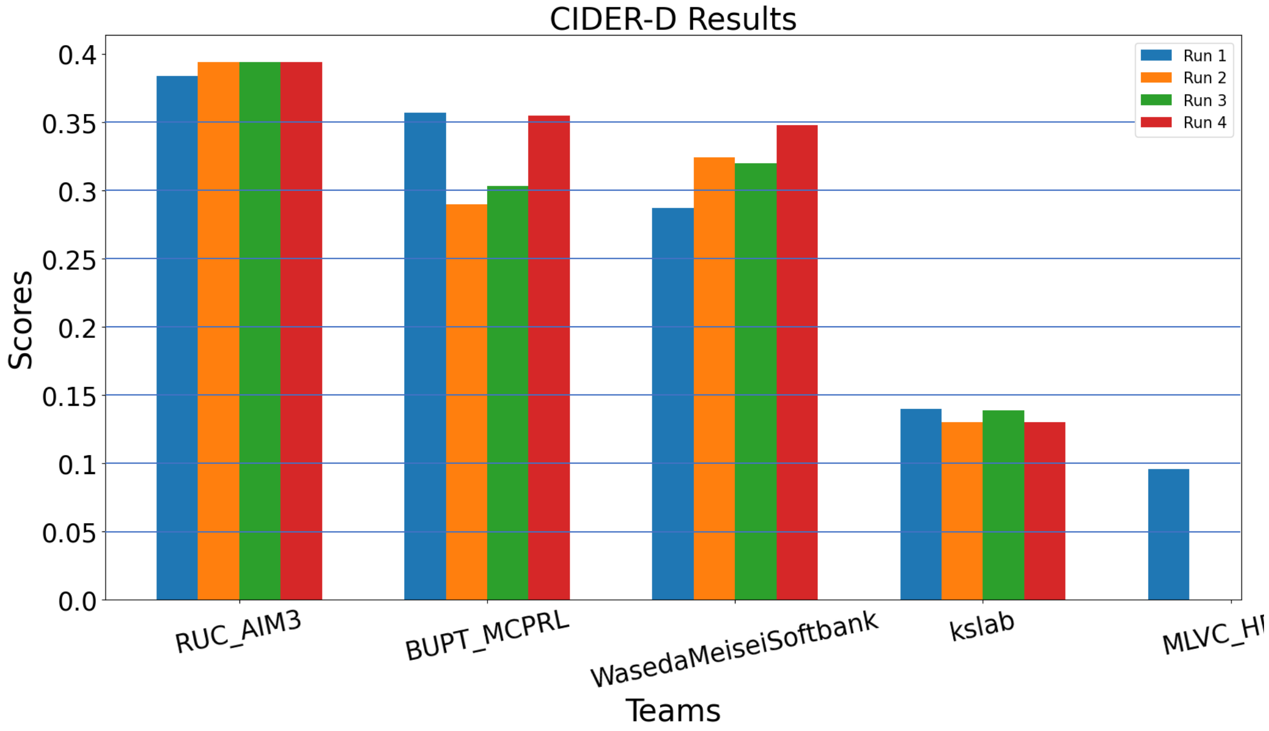


Main Task

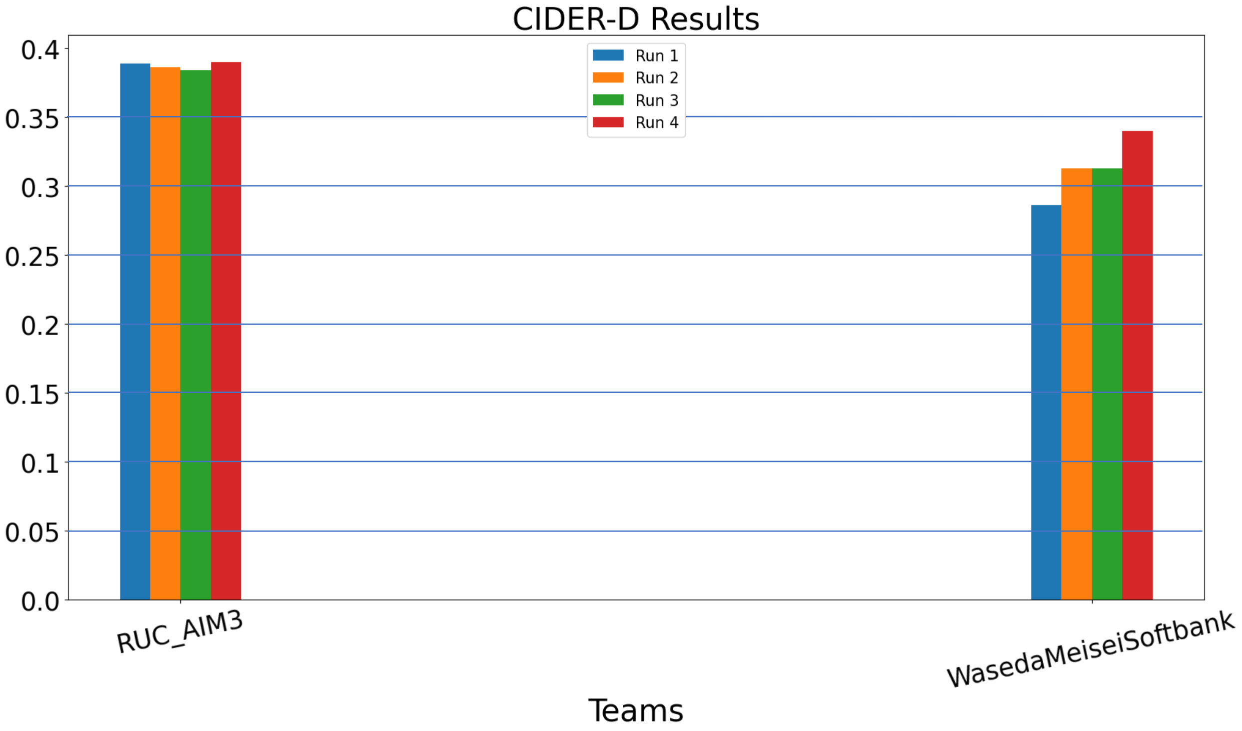


Robustness Task

Results

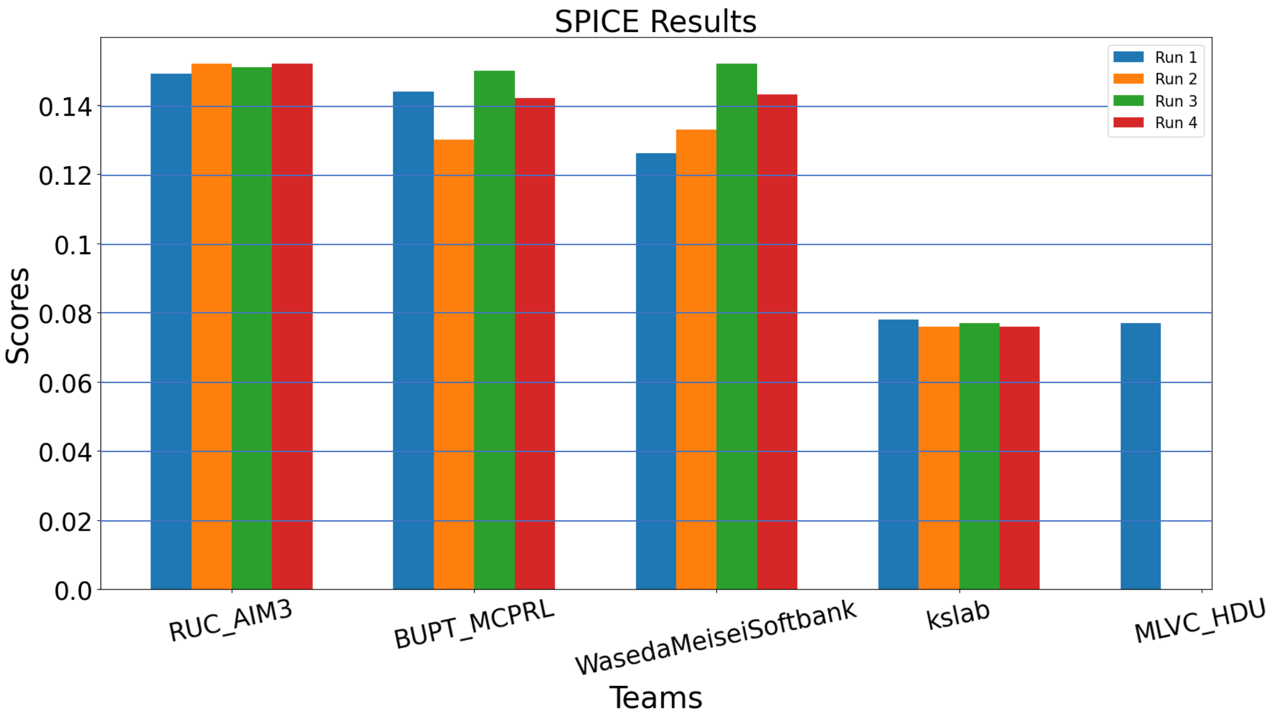


Main Task

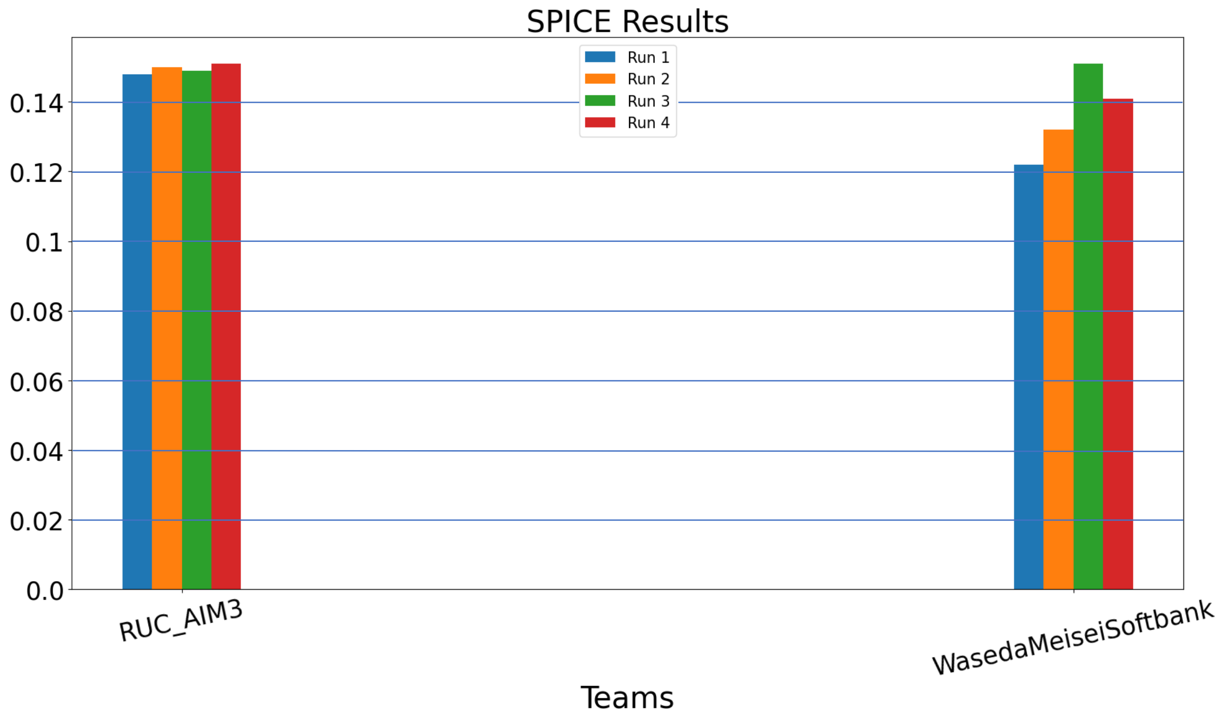


Robustness Task

Results

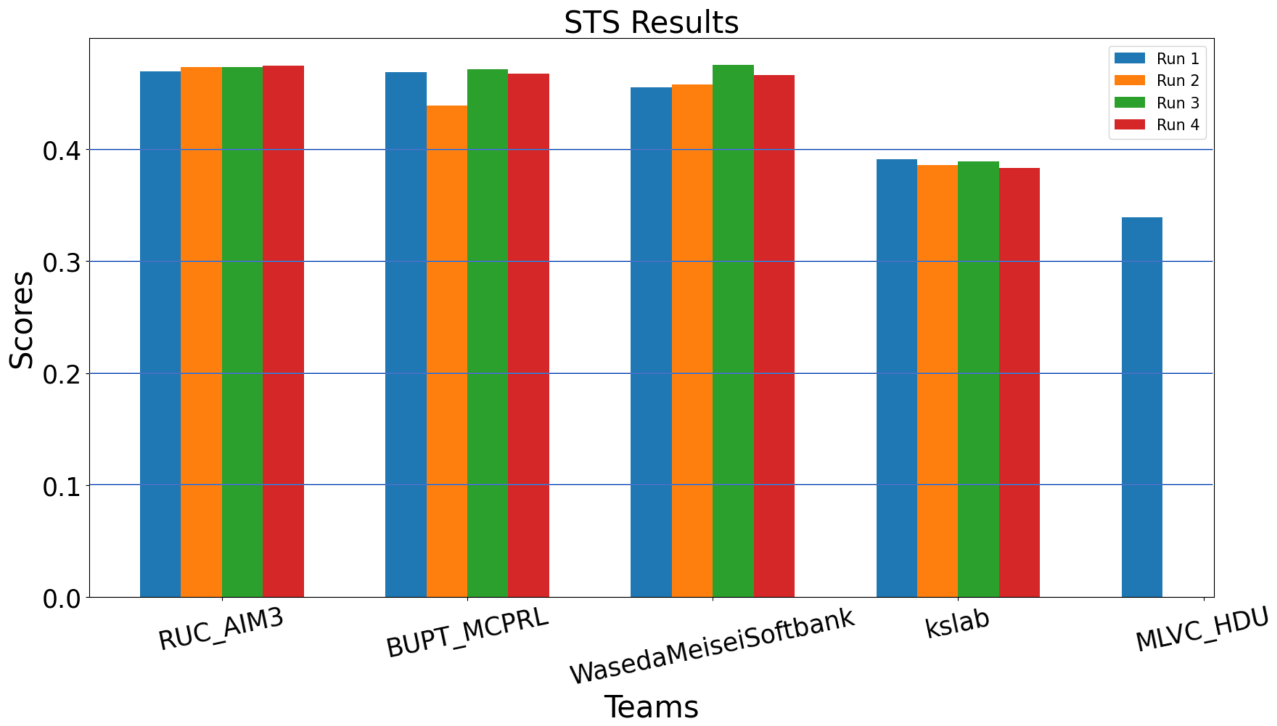


Main Task

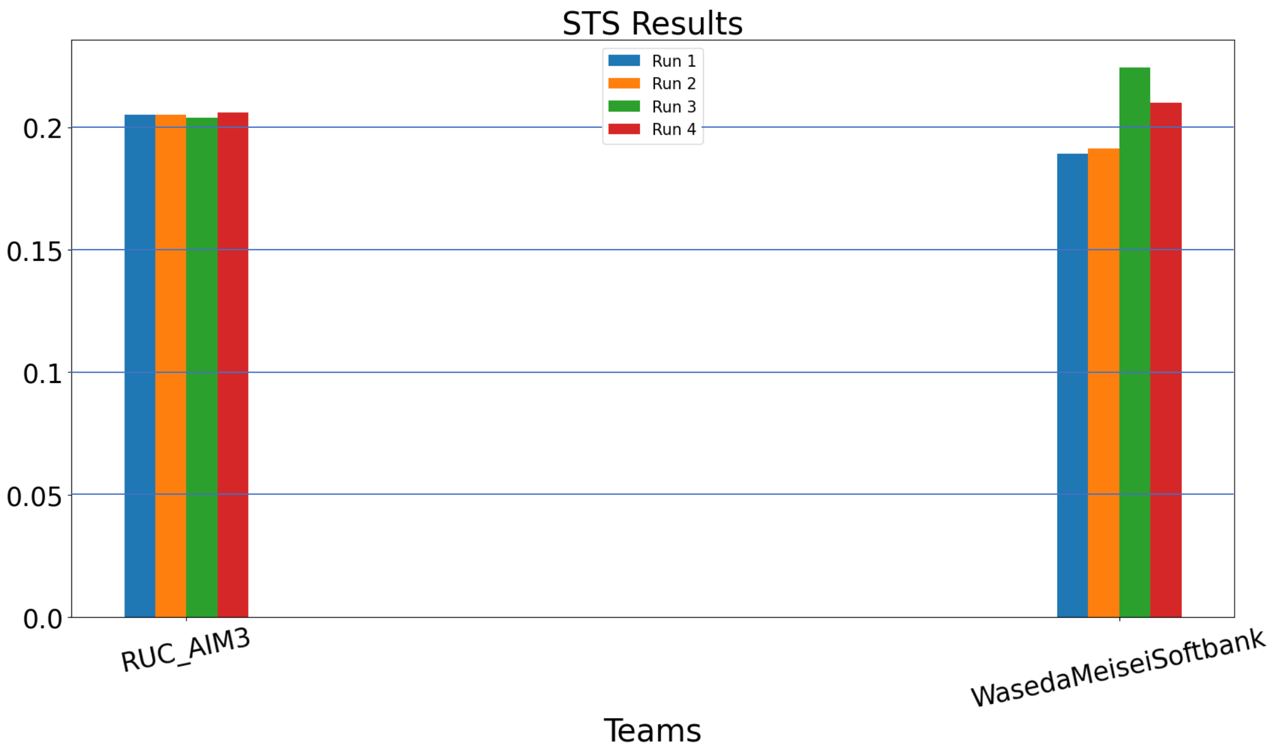


Robustness Task

Results



Main Task



Robustness Task

Correlation of Automated Metrics – Main Task

	CIDER	*CIDER-D*	*SPICE*	*METEOR*	*BLEU*	*STS*
CIDER	1	0.931	0.877	0.886	0.912	0.959
CIDER-D		1	0.971	0.966	0.916	0.959
SPICE			1	0.989	0.876	0.963
METEOR				1	0.923	0.971
BLEU					1	0.925
STS						1

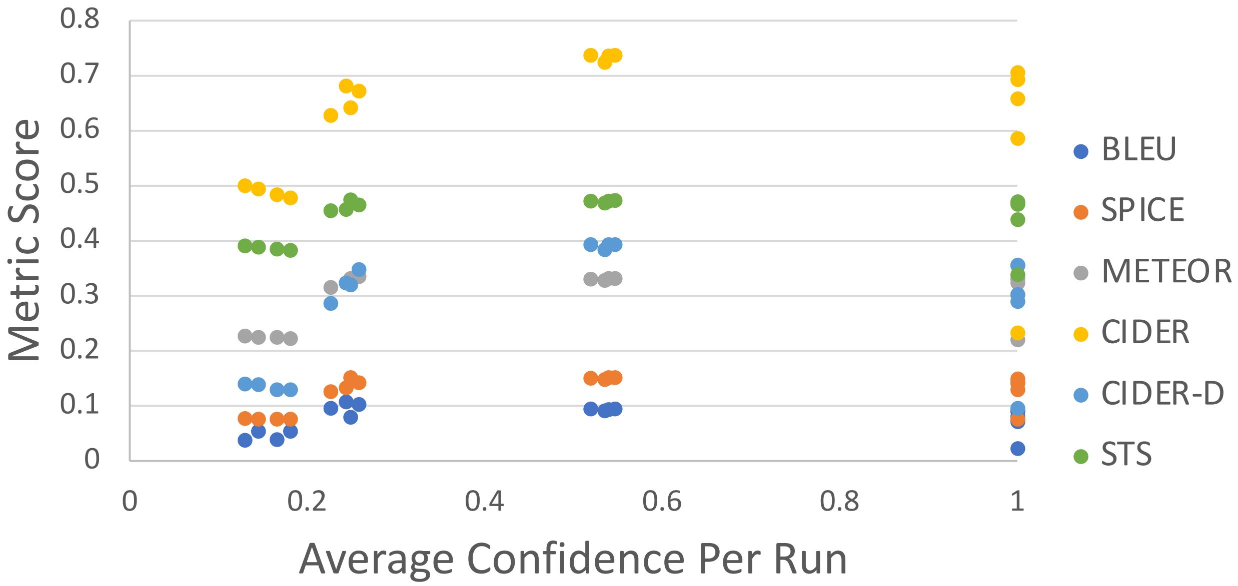
Correlation of Automated Metrics – Robustness Task

NIST

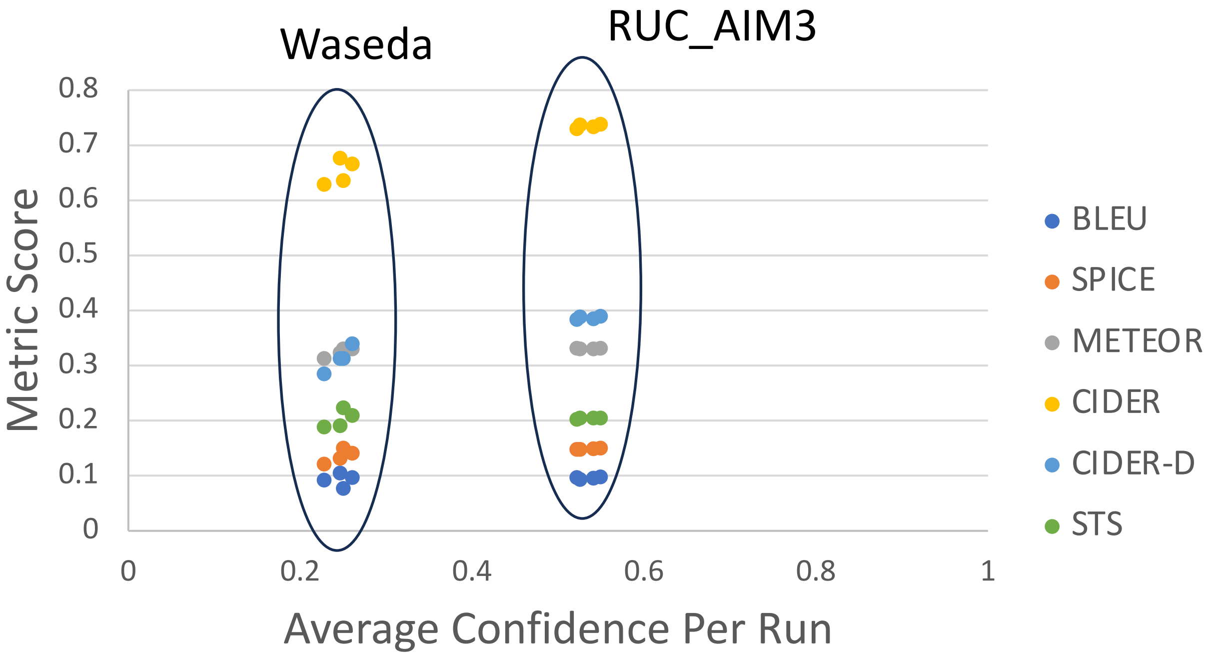
	CIDER	*CIDER-D*	*SPICE*	*METEOR*	*BLEU*	*STS*
CIDER	1	0.963	0.631	0.595	0.456	0.007
CIDER-D		1	0.771	0.74	0.277	0.236
SPICE			1	0.939	-0.256	0.769
METEOR				1	-0.08	0.745
BLEU					1	-0.693
STS						1

Confidence vs Score

- Teams were asked to provide confidence scores for the generated sentences.



Main Task

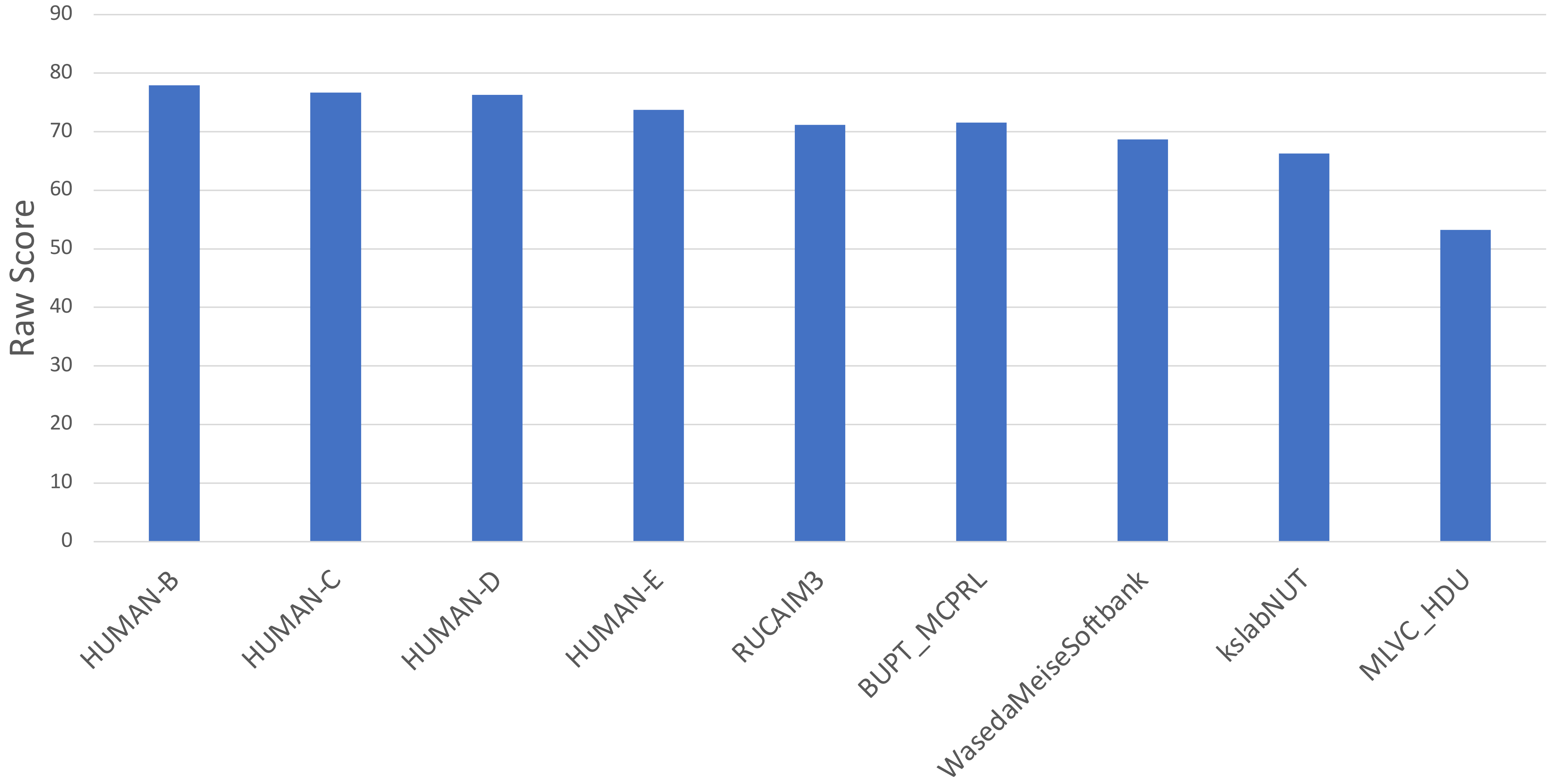


Robustness Task

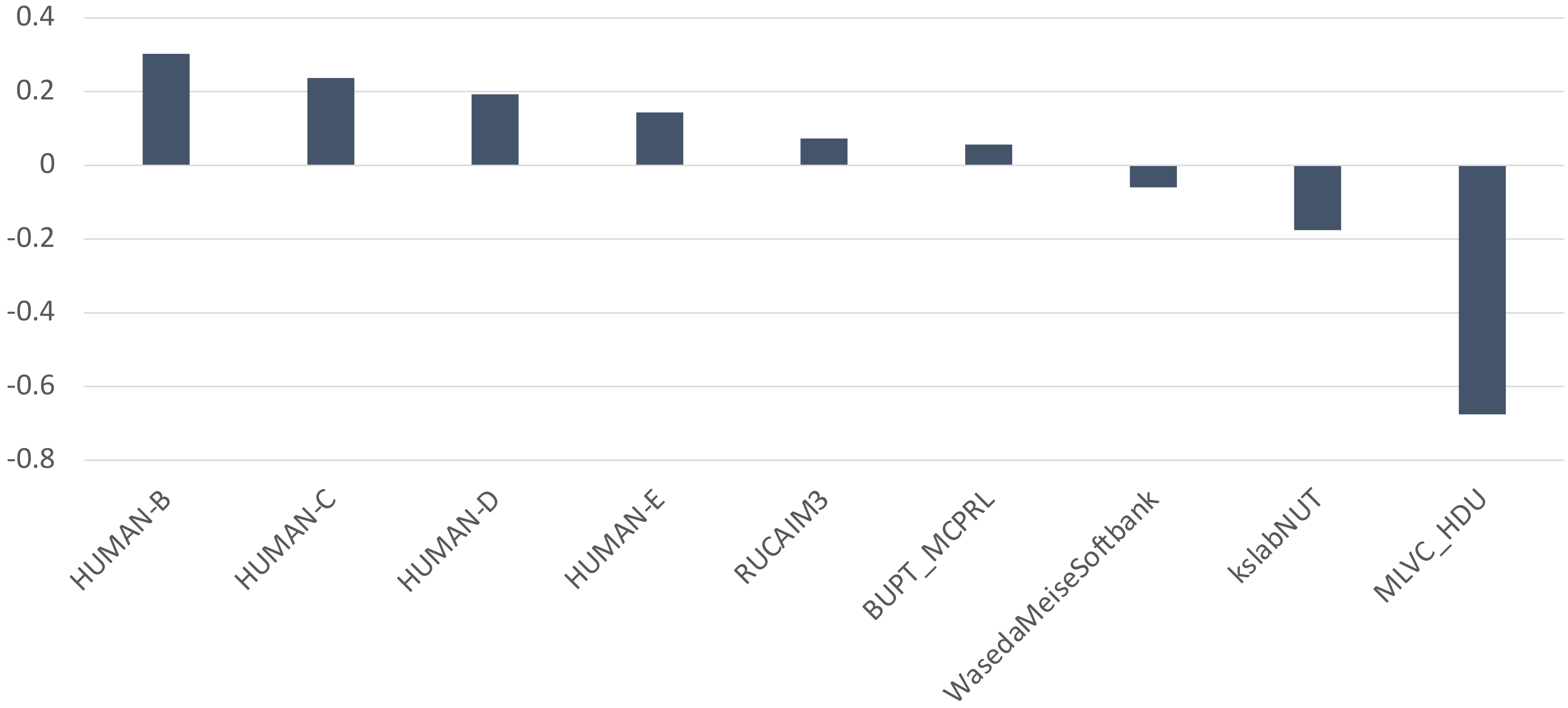
Direct Assessment

- DA uses crowdsourcing to evaluate how well a caption describes a video.
- Human evaluators rate captions on a scale of 0 to 100.
- DA conducted on only primary runs for each team.
- The DA score is reported as follows:
  - Raw score is the average score for each run over all videos. It ranges between 0 and 100.
  - Z score is standardized per individual AMT worker's mean and standard deviation score. The average Z score is then reported for each run.

DA Results - Raw



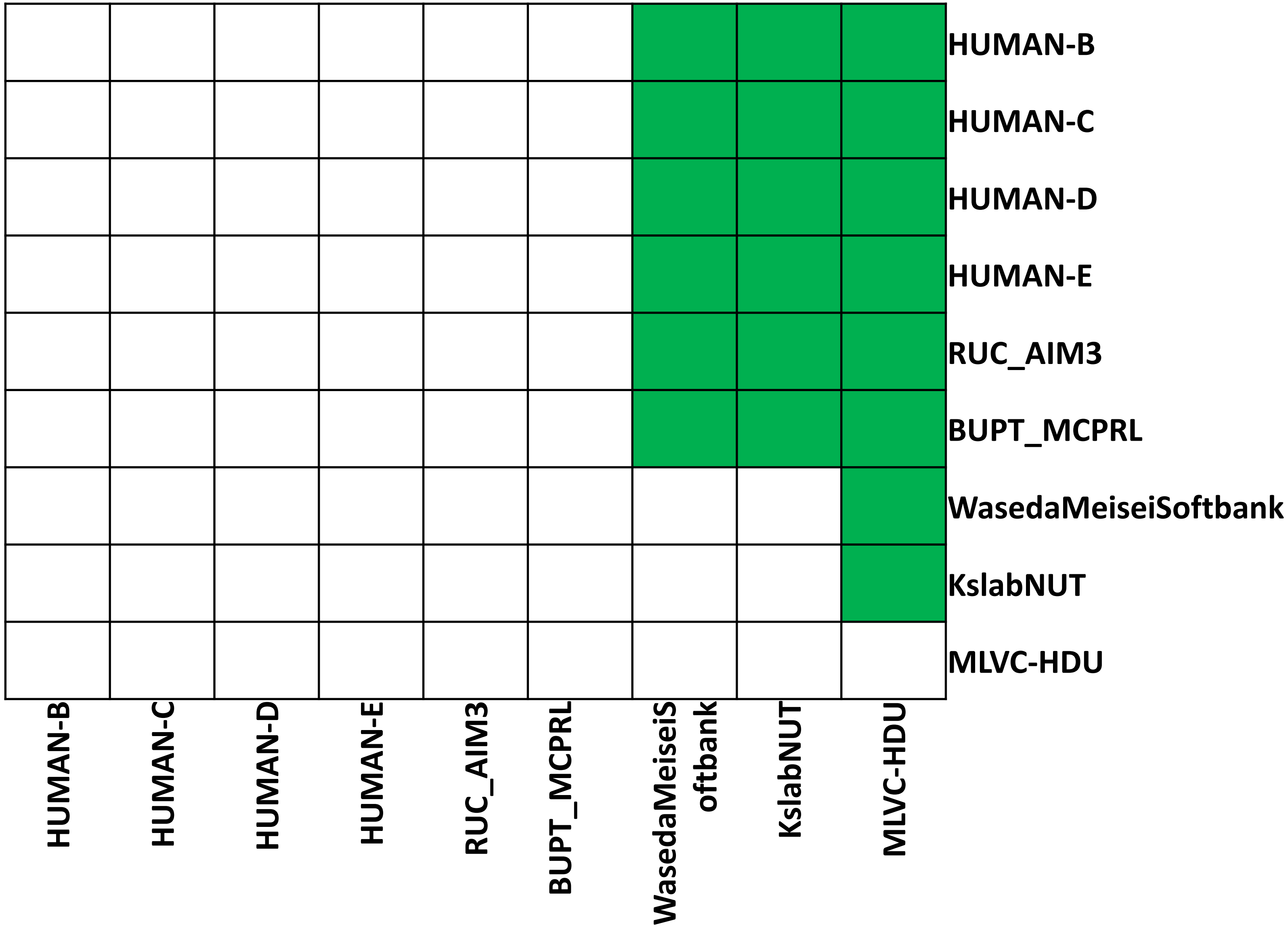
DA Results - Z



0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
HUMAN-B
HUMAN-C
HUMAN-D
HUMAN-E
RUCAIM3
BUPT_MCPRL
WasedaMeiseSoftbank
kslabNUT
MLVC_HDU

DA Result - Significance

- Green squares indicate a significant “win” for the row over the column.
- Amongst systems, RUC_AIM3 and BUPT_MCPRL leads the others.

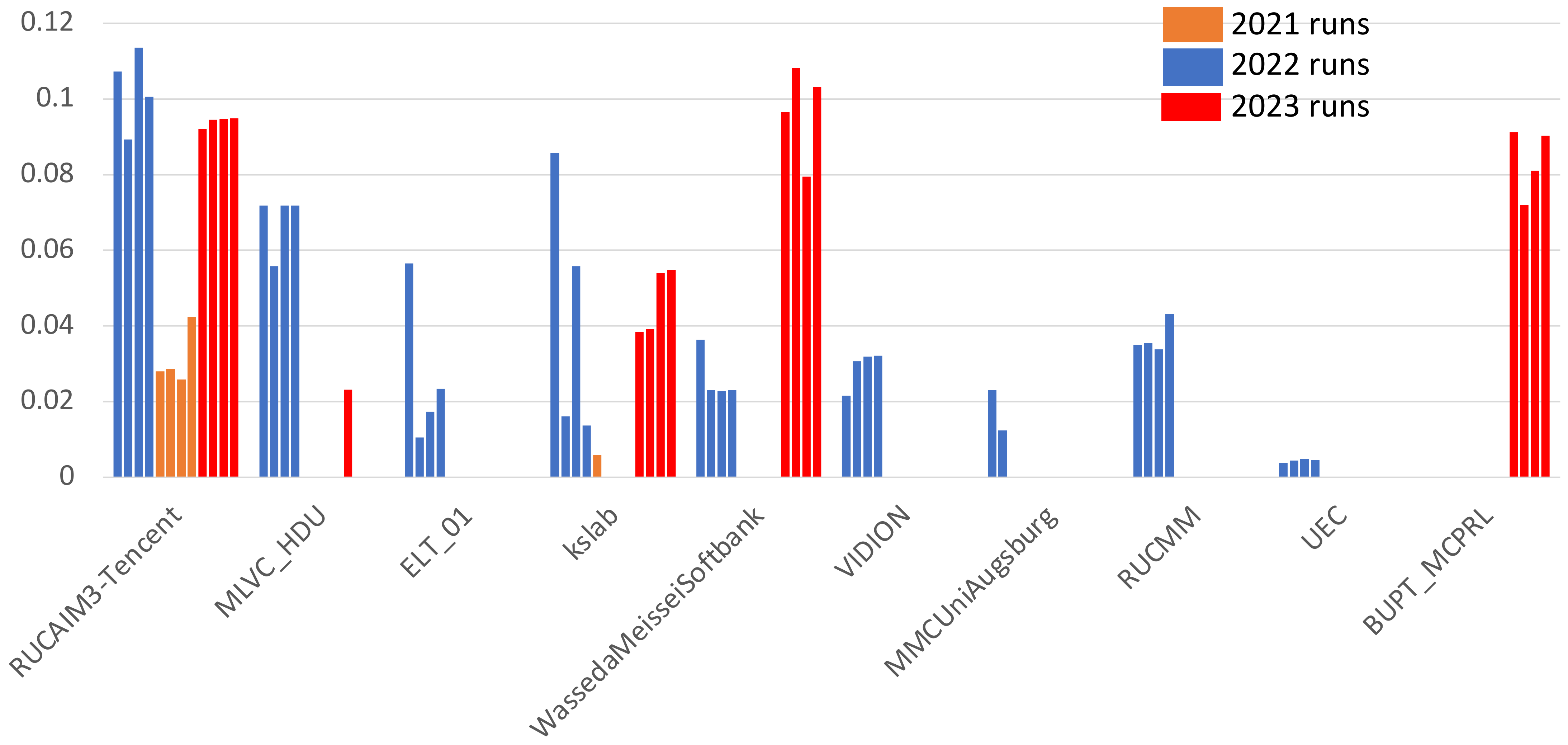


Correlation (DA , automatic metrics)

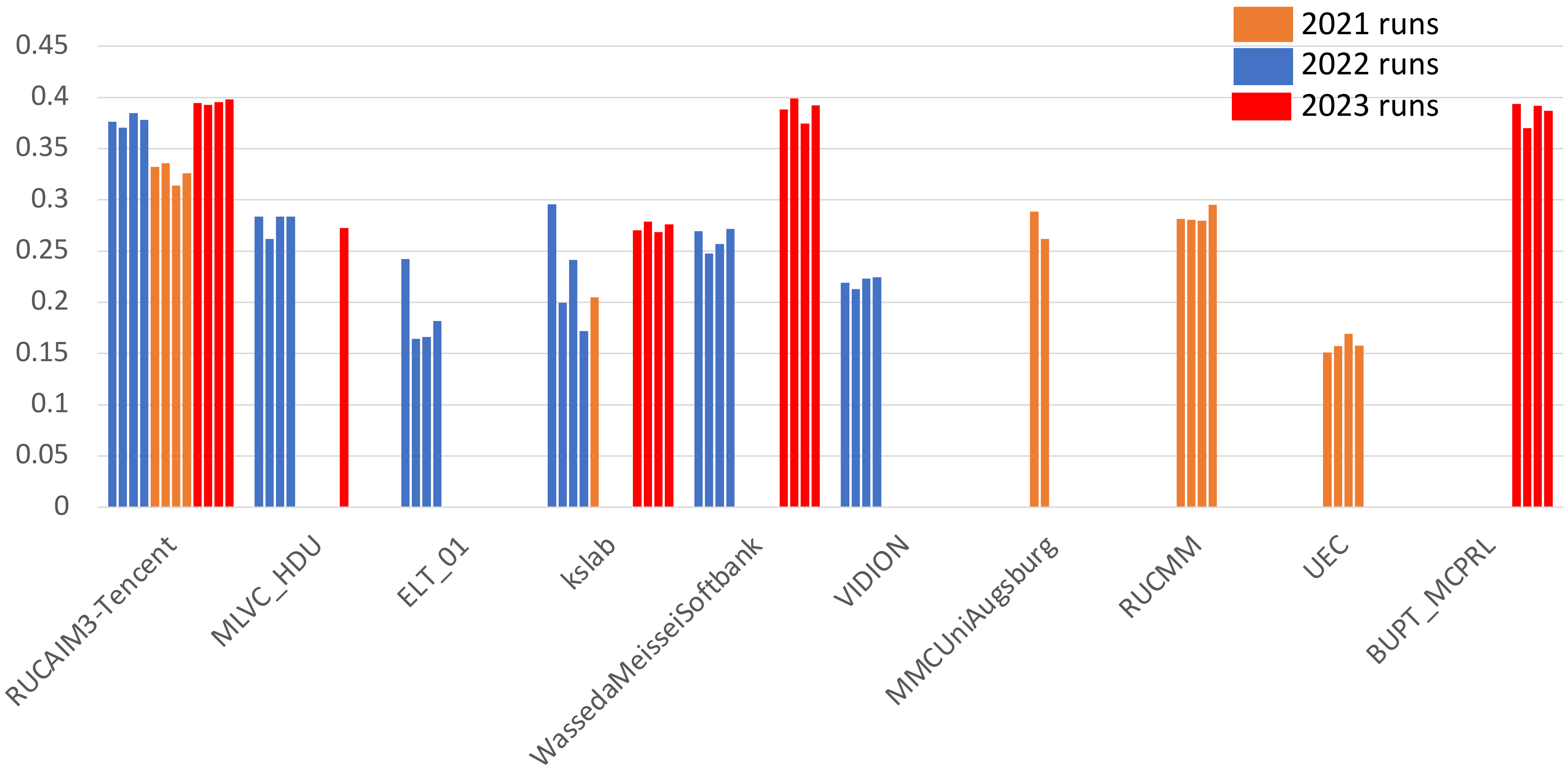
	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
DA	0.89	0.82	0.98	0.87	0.81	0.94

**Based only on the primary run by each team

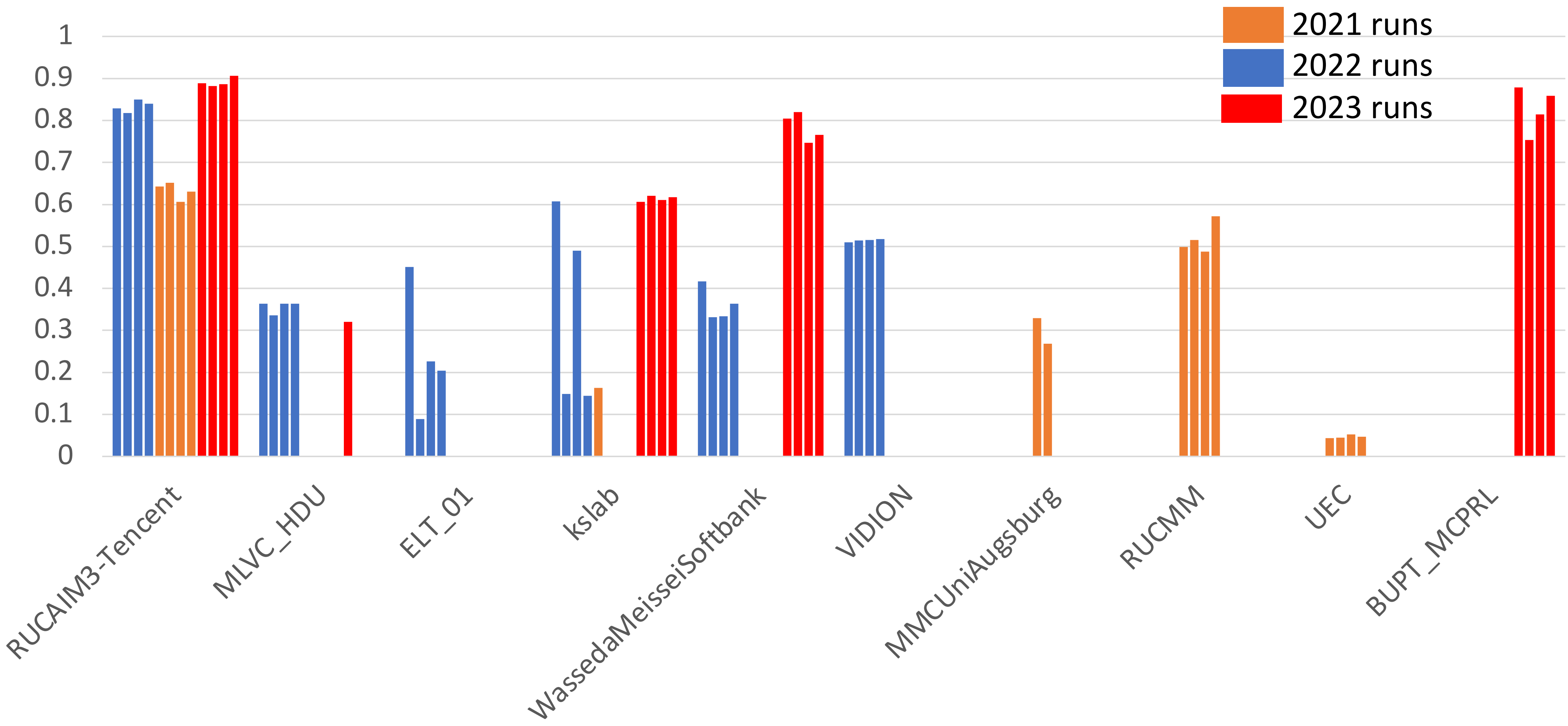
Progress subtask - BLEU Results



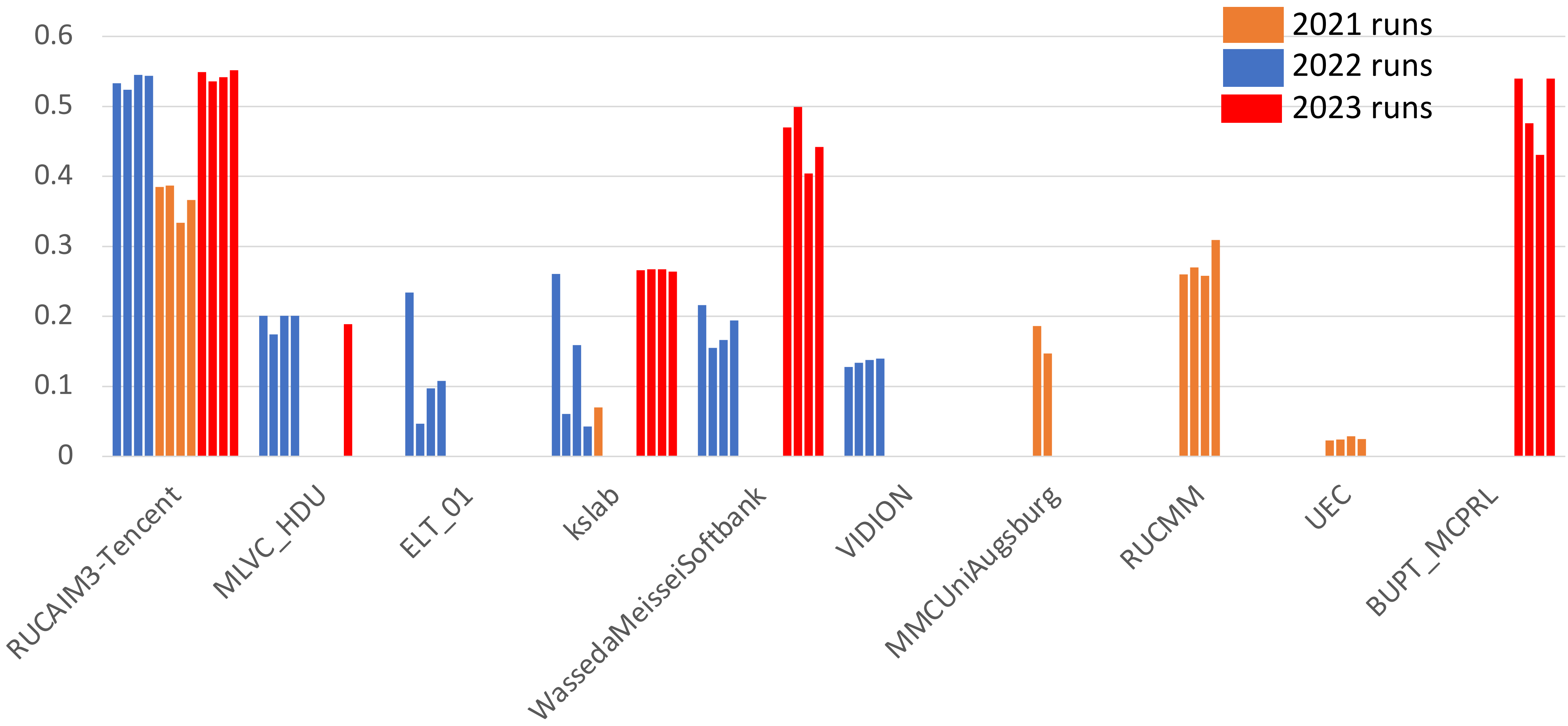
Progress subtask - METEOR Results



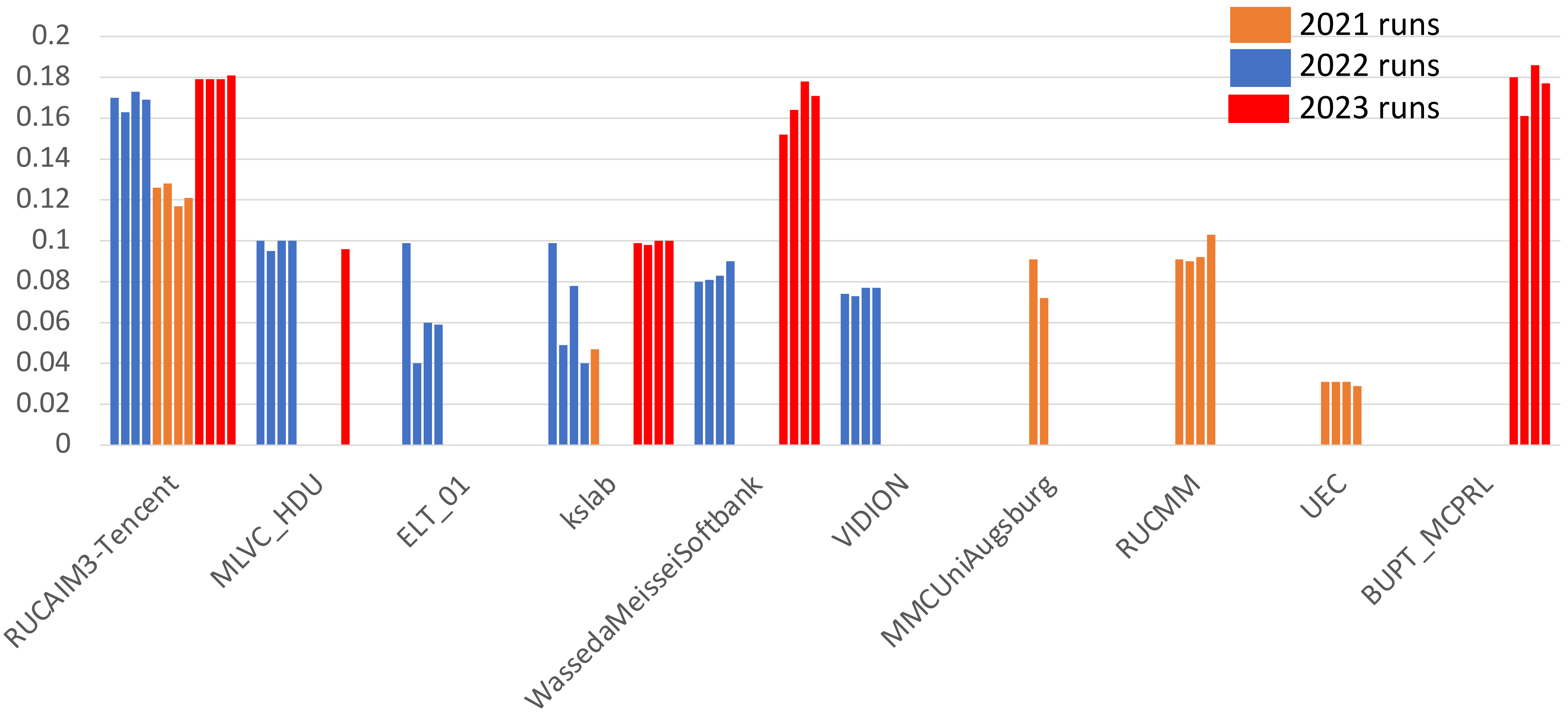
Progress subtask - CIDER Results



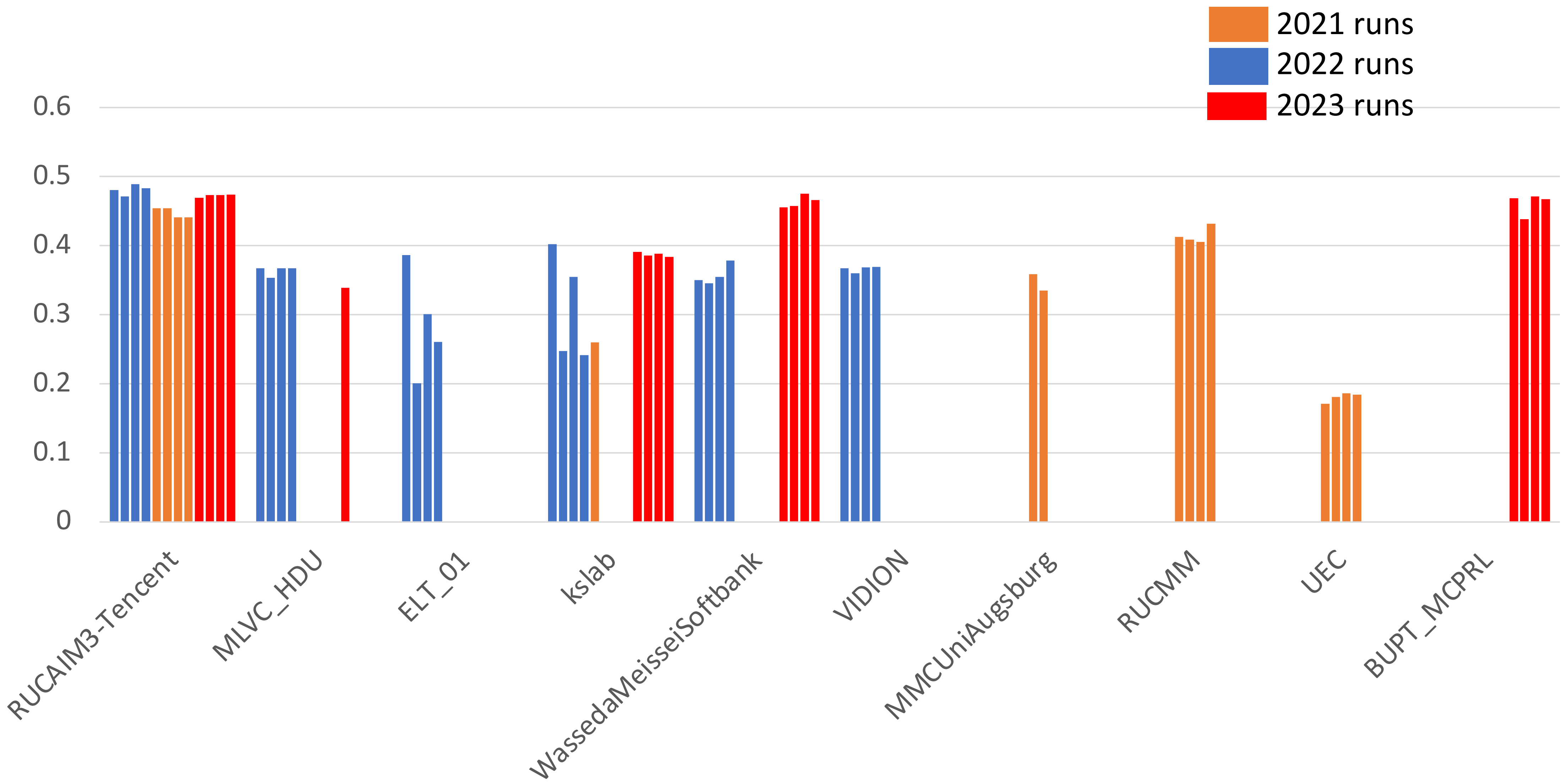
Progress subtask - CIDER-D Results



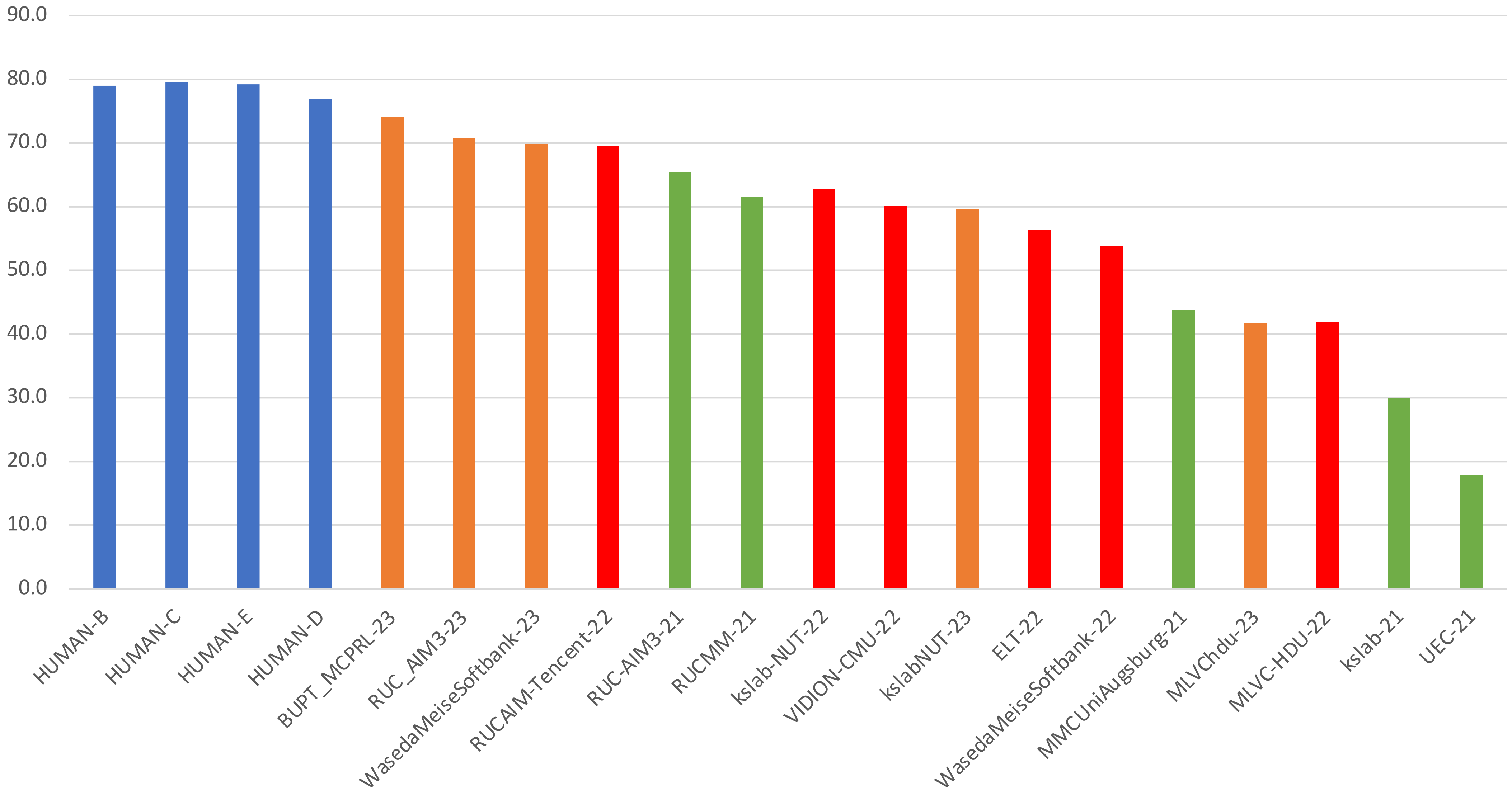
Progress subtask - SPICE Results



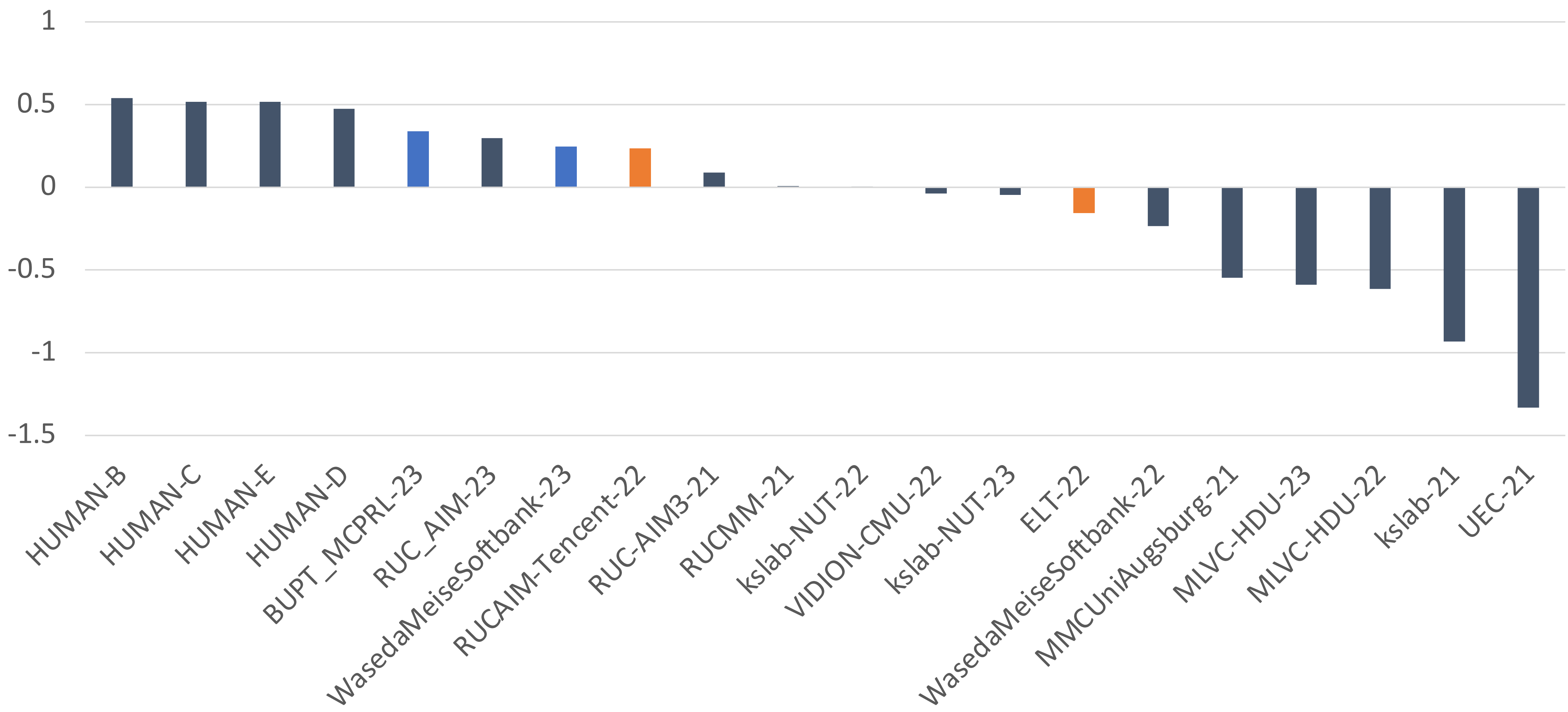
Progress subtask - STS Results



DA Results - Raw



Progress subtask - DA Results - Z



Progress subtask - DA Significant Diff. Results

1

- Human-B
- Human-C
- Human-E
- Human-D

2

- BUPT_MCPRL-23
- RUC_AIM3-23
- WasedaMeiseSoftbank-23
- RUCAIM_Tencent-22

3

- RUC-AIM3-21
- RUCMM-21
- Kslab-NUT-22
- VIDION-CMU-22
- Kslab-NUT-23
- ELT-22
- WasedaMeiseSoftbank-22

4

- MMCuniAugsburg-21
- MLVC-HDU-23
- MLVC-HDU-22

5

- Kslab-21

6

- UEC-21

Examples (GT vs Submissions)

GT:
1- Closeup video of a white male taking aim with a rifle.
2- A man's eye can be seen as he looks at something off camera, then raises a rifle with a scope mounted and aims at what he was looking at.
3- A middle aged man is readying himself to aim his gun toward something.
4- Closeup of the eyes of a white man raising a rifle to his eyes and taking aim.
5- A Caucasian man looks and then lifts his rifle to shoot.

Submissions:
1- a close up of a man’s eyes as he looks through a scope
2- A close up of a man looking into the camera
3- a person is making faces
4- a man with a mustache and mustache is talking to the camera in a room with green walls
5- A man is looking into the camera.

Examples (GT vs Submissions)

GT:
Camels are walking in the desert followed by a video of a vehicle wheel going down a road.
During the day a number of camels walk in the desert and then the video shows a car driving down a road in an arid climate.
On an open desert space, several camels can be seen walking across a paved road just before a vehicle approaches.
In a wide flat desert area a vehicle drives past wild dromedary camels, which move away from the road as the vehicle approaches.
A group of camels are walking in the desert followed by a left front wheel of a car coming into view.

Submissions:
- camels are walking in a desert on a sunny day
- A camel walking in the desert
- a camel is seen running on the road on a sunny day
- a group of camels are walking on a dirt road in the desert on a sunny day
- A camel is walking in a desert on a sunny day.

Conclusion

- This was the first year using the V3C3 test data (following two years of V3C2 and 1 year of V3C1).
- Participation in the task is stable.
- Few teams used audio features.
- 3rd year for the progress subtask (still needed?).
- High correlation between all automatic metrics.
- First year to pilot robustness sub-task

Thank you!

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