TRECVID 2022: Video to Text Description

Asad Anwar Butt
NIST; Johns Hopkins University

George Awad
NIST; Georgetown University

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.

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.
- Mixed up things a little with addition of Flickr videos 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 10 sec selected for this task.
  - Videos distributed directly to participants.

Test Dataset

- Manual selection of videos.
  - Watched 7600+ videos.
  - Selected 2008 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.

Annotation – Observations

- Average sentence length for each assessor:

Annotator	Avg. Length	# Videos
1	24.84	2008
2	18.06	2008
3	21.93	2008
4	26.09	2008
5	21.50	2008

Avg. sentence length: 24.46 words

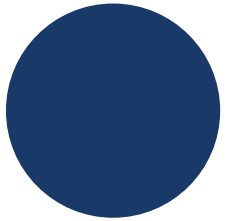
- Additional questions:

Please rate how difficult it was to describe the video.

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

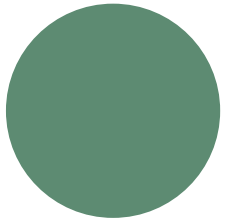
How likely is it that other assessors will write similar descriptions for the video?

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



Q1 Avg Score: 2.52 (Scale of 5)

Q2 Avg Score: 2.37 (Scale of 3)



Correlation between difficulty scores: -0.61

Participants

Teams	Organization
KSLAB	Nagaoka University of Technology
MLVC_HDU	Hangzhou Dianzi University
RUCAIM3-Tencent	Renmin University of China
VIDION	Carnegie Mellon University
WasedaMeiseiSsoftbank	Waseda University, Meisei University, SoftBank Corporation
ELT_01	Elyadata

- 6 teams participated with 24 runs

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:

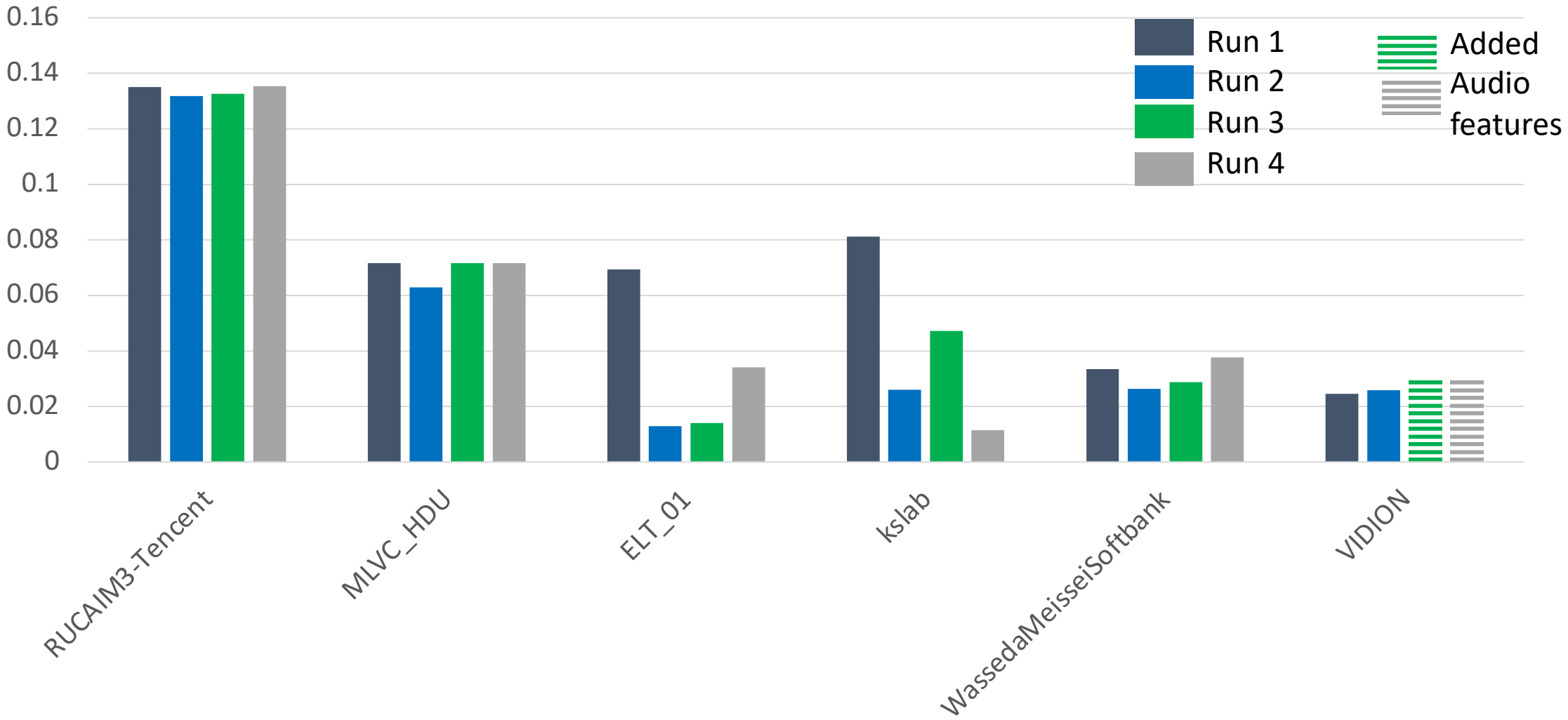
'V': Visual features only

'A': Both audio and visual features

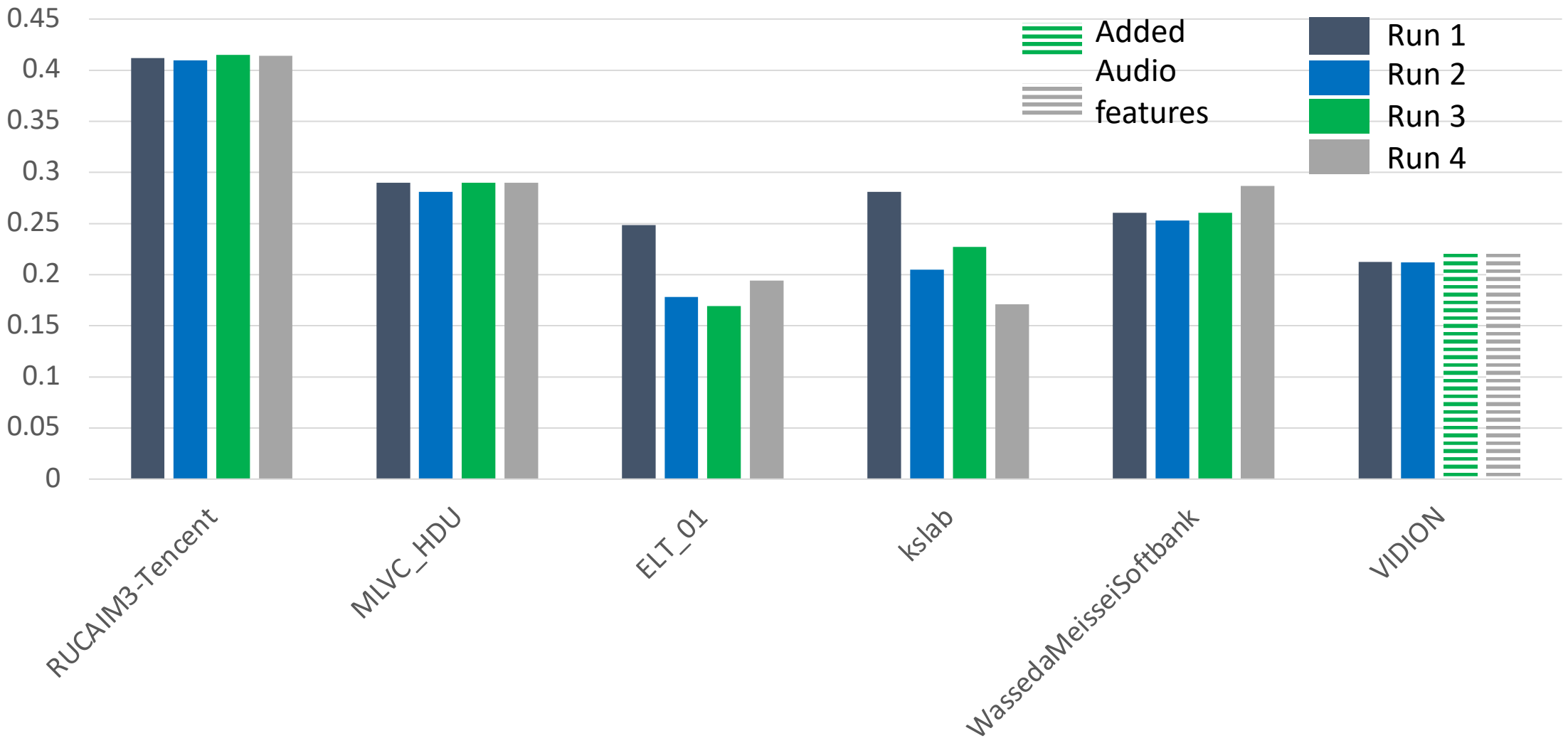
Submissions - Run Types

1 'VV' (Video Data/Visual Feats) • 12 runs	**2 'IV' (Image Data/Visual Feats)** • 5 runs
3 'BV' (I+V Data/Visual Feats) • 5 runs	**4 'VA' (Video Data/V+A Feats)** • 2 runs

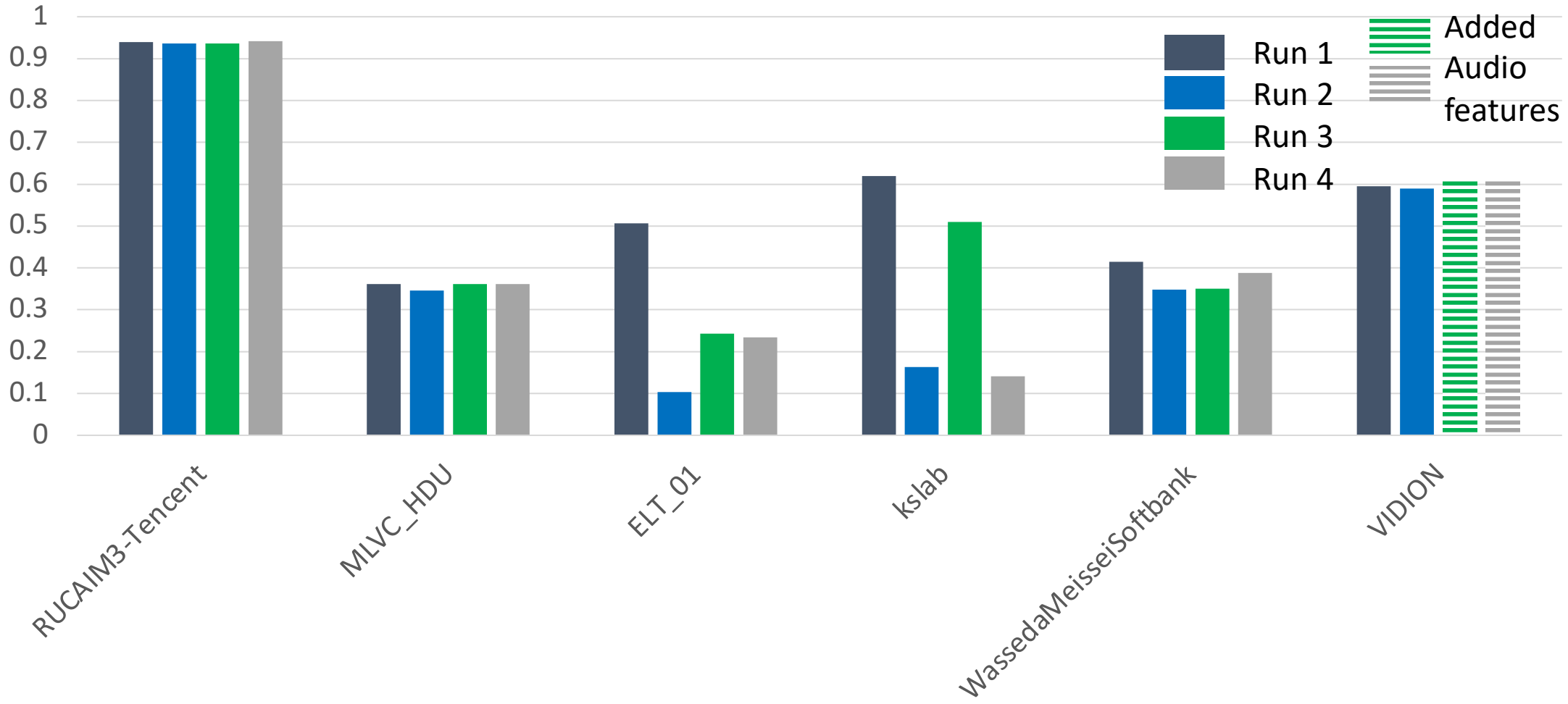
BLEU Results



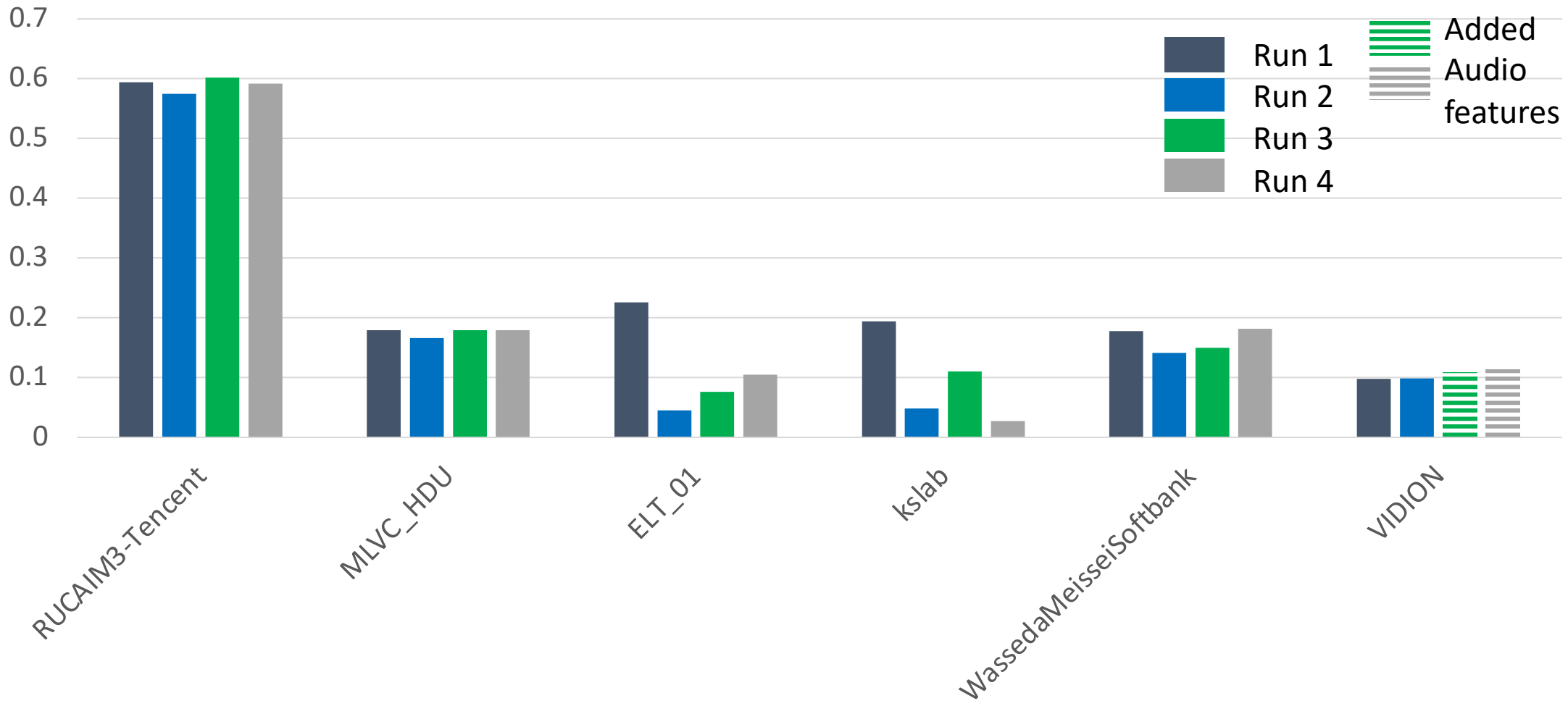
METEOR Results



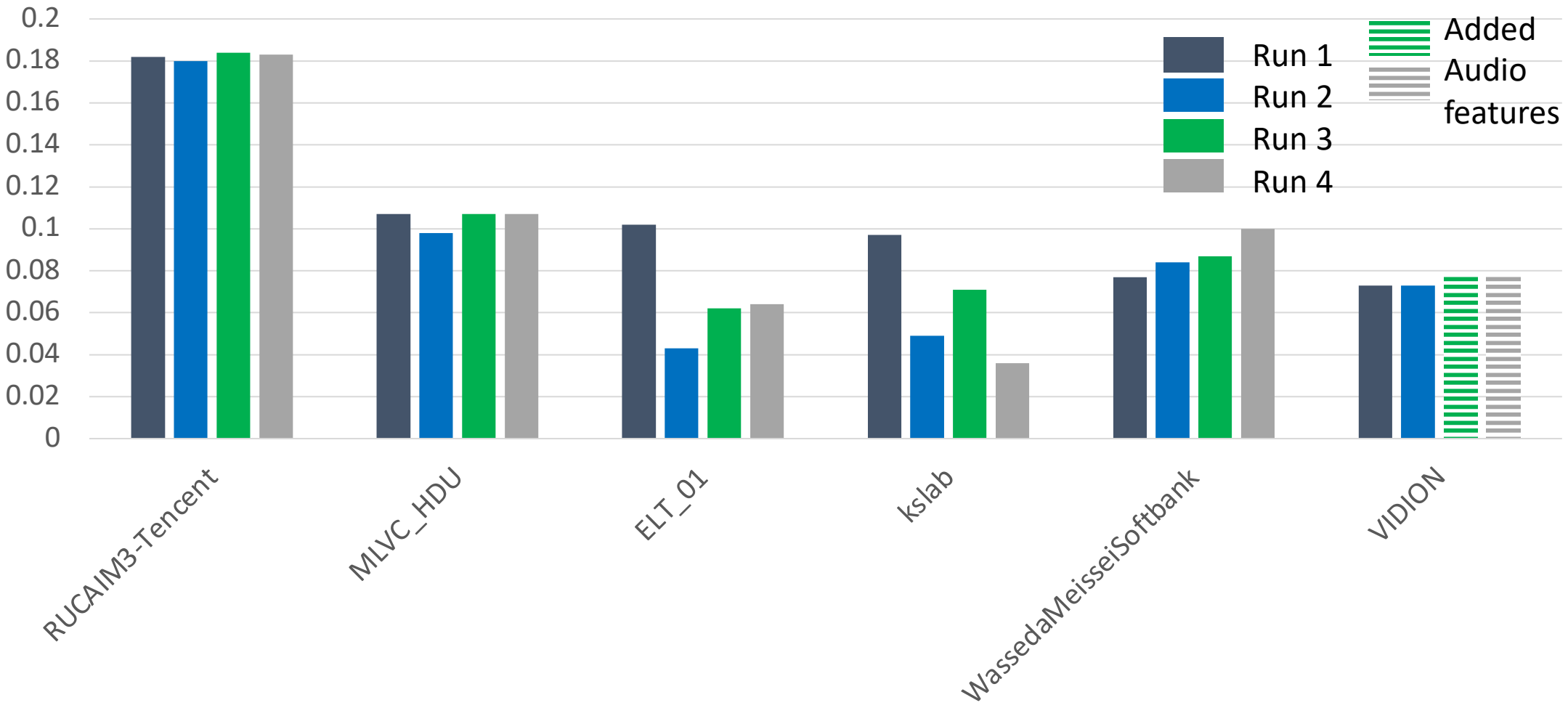
CIDER Results



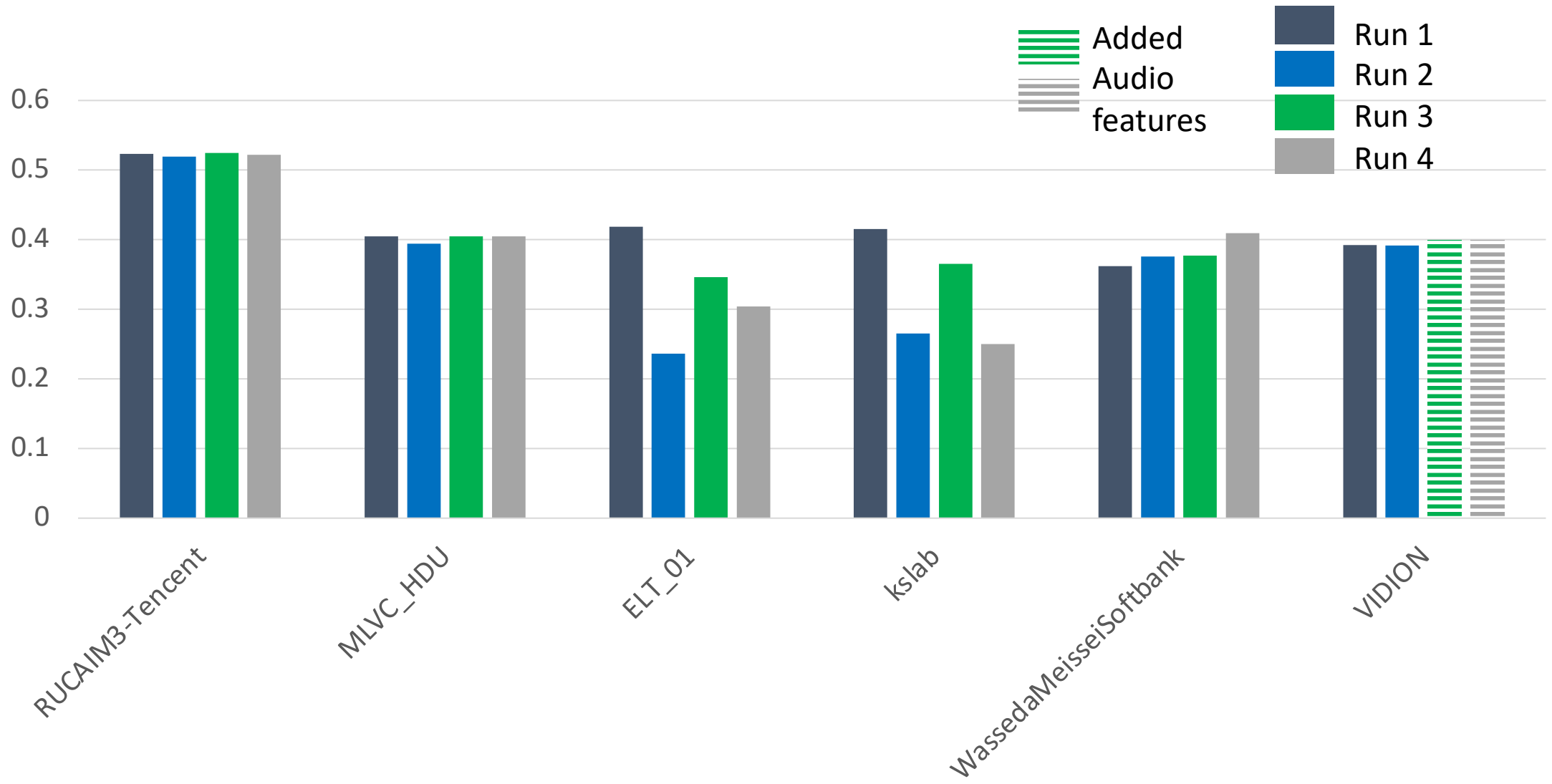
CIDER-D Results



SPICE Results



STS Results

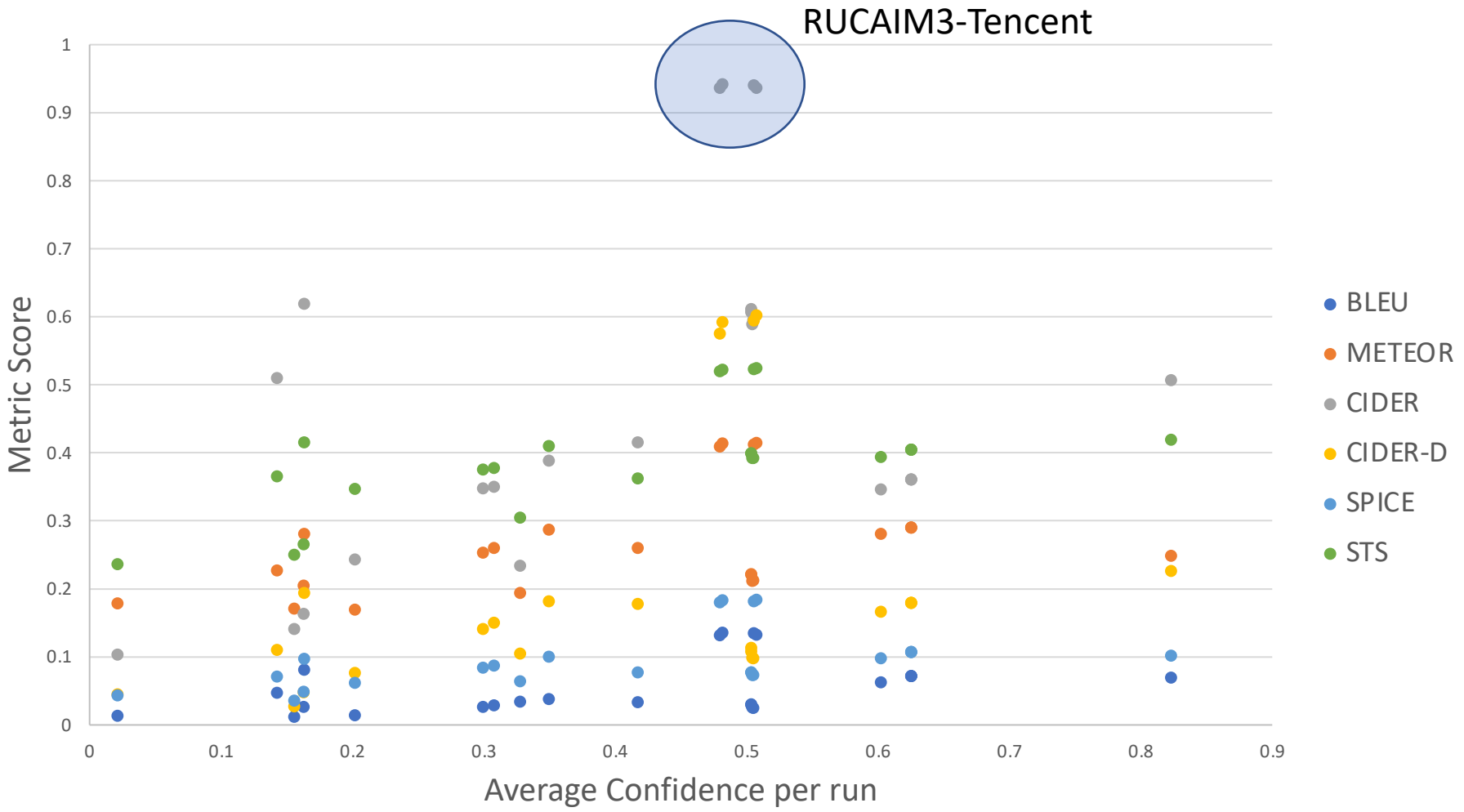

Added
Audio
features
Run 1
Run 2
Run 3
Run 4
0.6
0.5
0.4
0.3
0.2
0.1
0
RUCAIM3-Tencent
MLVC_HDU
ELT_01
kslab
WassedaMeisseiSoftbank
VIDION

Correlation of Run Scores – Automated Metrics

	BLEU	*METEOR*	*CIDER*	*CIDER-D*	*SPICE*	*STS*
BLEU	1.00	0.95	0.80	0.94	0.96	0.86
METEOR	0.95	1.00	0.80	0.96	0.98	0.89
CIDER	0.80	0.80	1.00	0.85	0.85	0.91
CIDER-D	0.94	0.96	0.85	1.00	0.98	0.87
SPICE	0.96	0.98	0.85	0.98	1.00	0.94
STS	0.86	0.89	0.91	0.87	0.94	1.00

Confidence Scores

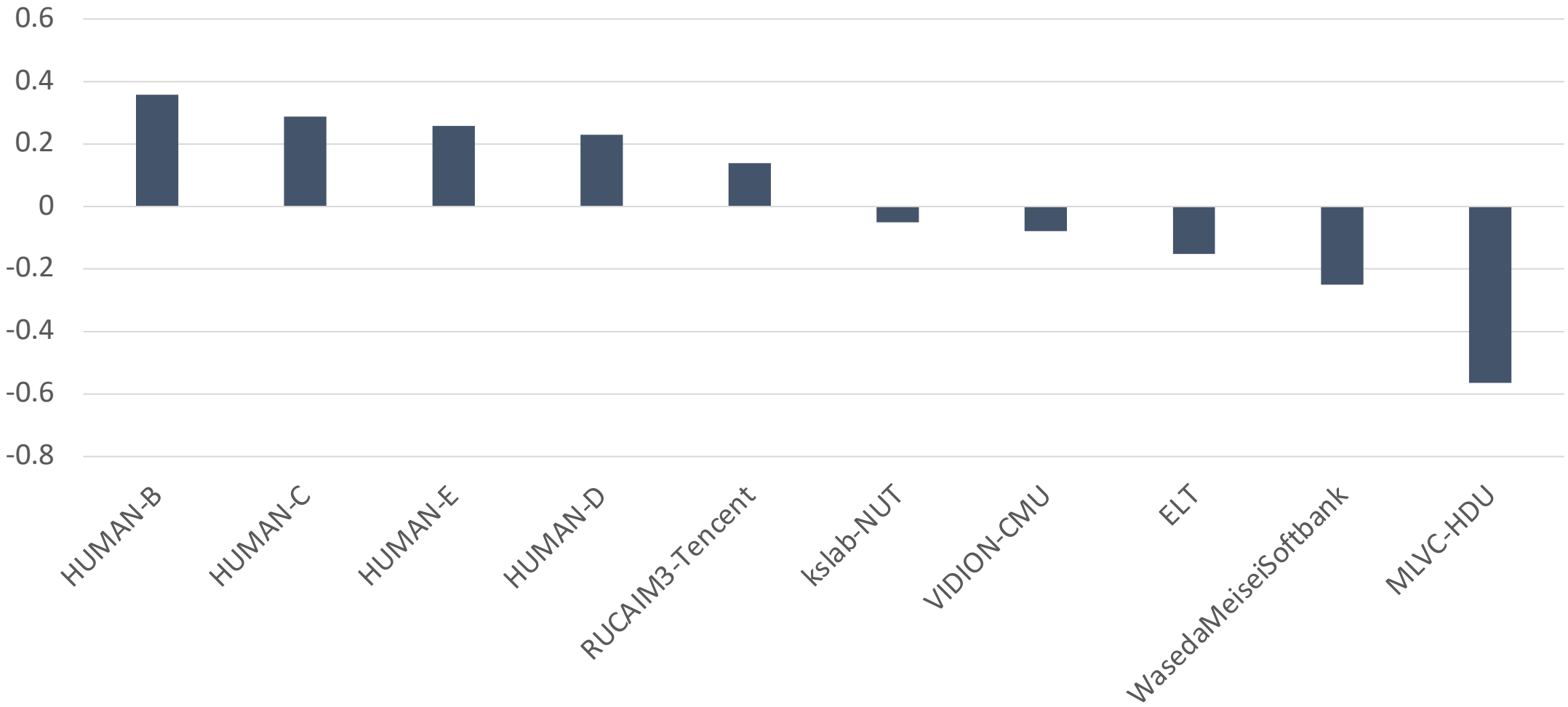
- Teams were asked to provide confidence scores for the generated sentences.



Direct Assessment

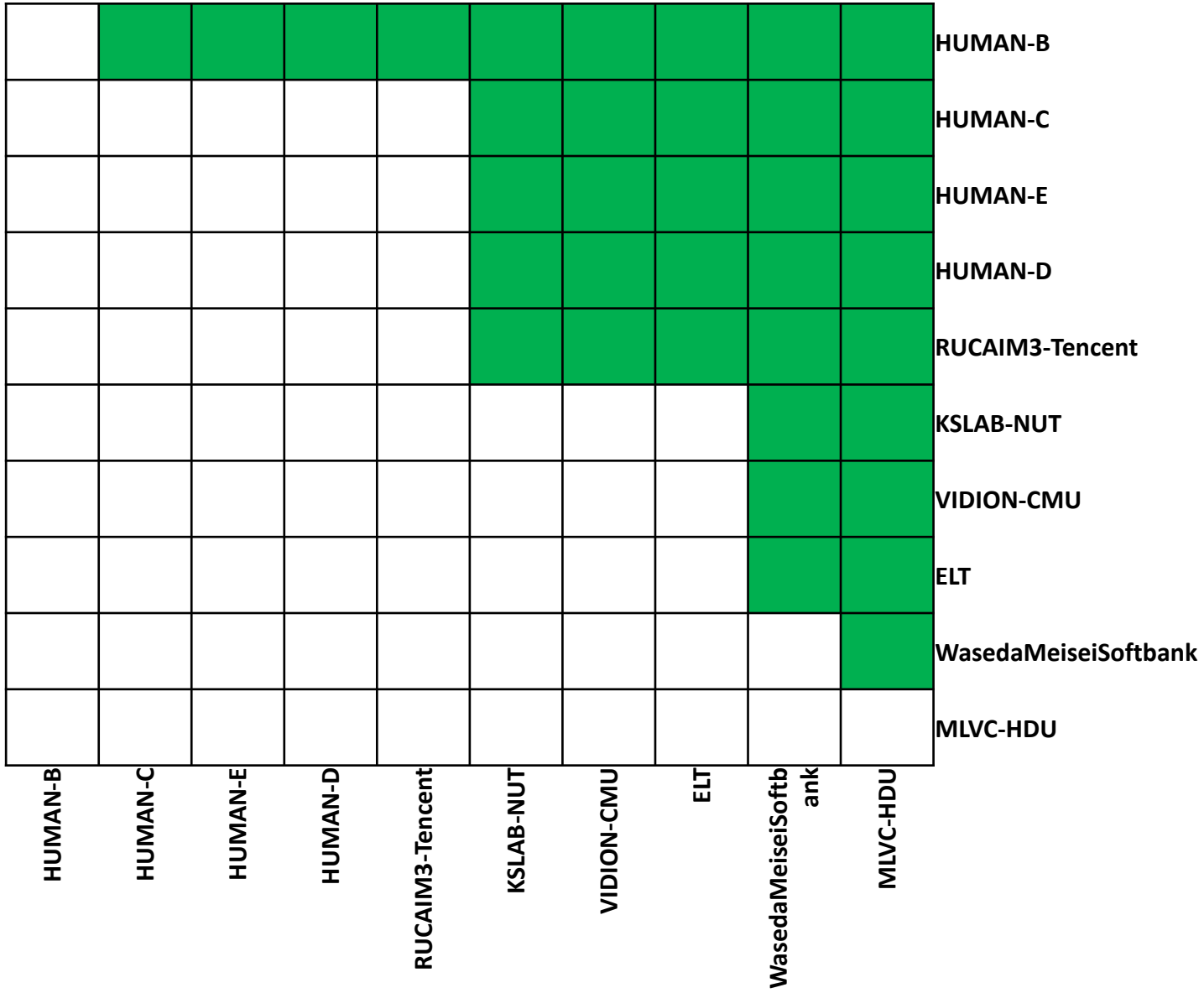
- DA uses crowdsourcing to evaluate how well a caption describes a video.
- Human evaluators rate captions on a scale of 0 to 100.
- DA conducted on only primary runs for each team.
- The DA score is reported as follows:
  - Raw score is the average score for each run over all videos. It ranges between 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 - Z


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
HUMAN-B
HUMAN-C
HUMAN-E
HUMAN-D
RUCAIM3-Tencent
kslab-NUT
VIDION-CMU
ELT
WasedaMeiseiSoftbank
MLVC-HDU

DA Result - Significance

- Green squares indicate a significant “win” for the row over the column.
- Amongst systems, RUCAIM3-Tencent leads the others.

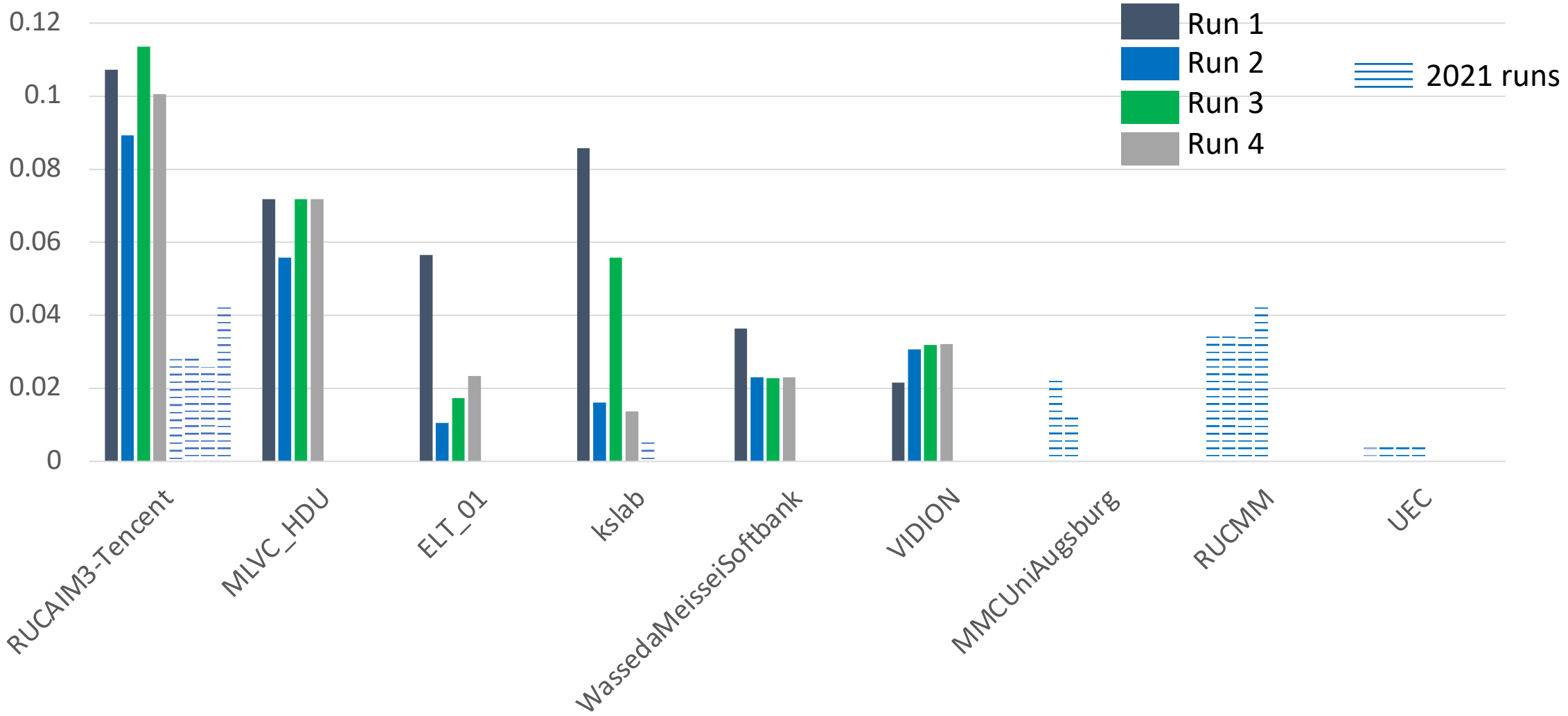


Correlation (DA , automatic metrics)

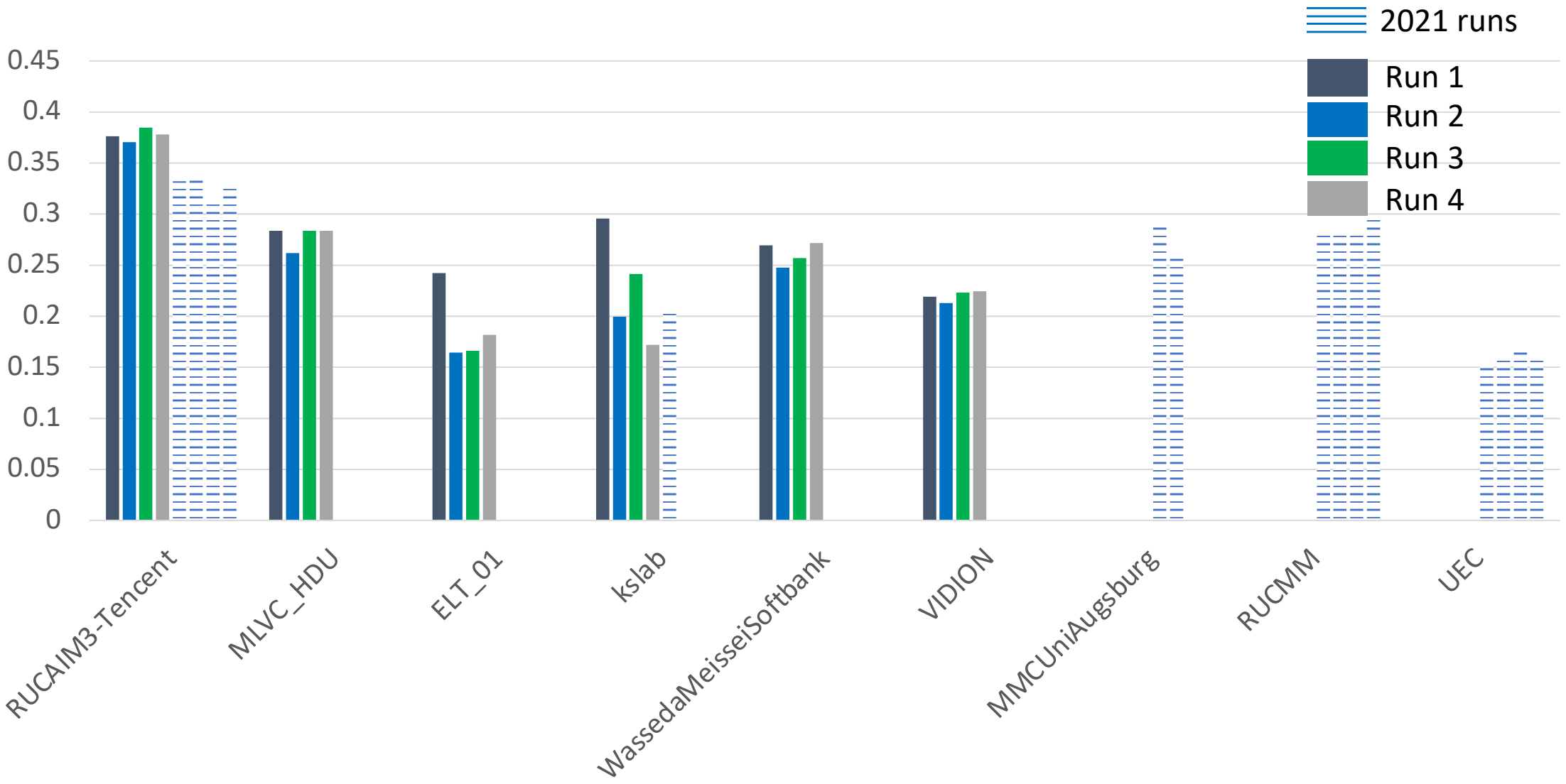
	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
DA	0.44	0.41	0.89	0.59	0.47	0.62

**Based only on the primary run by each team

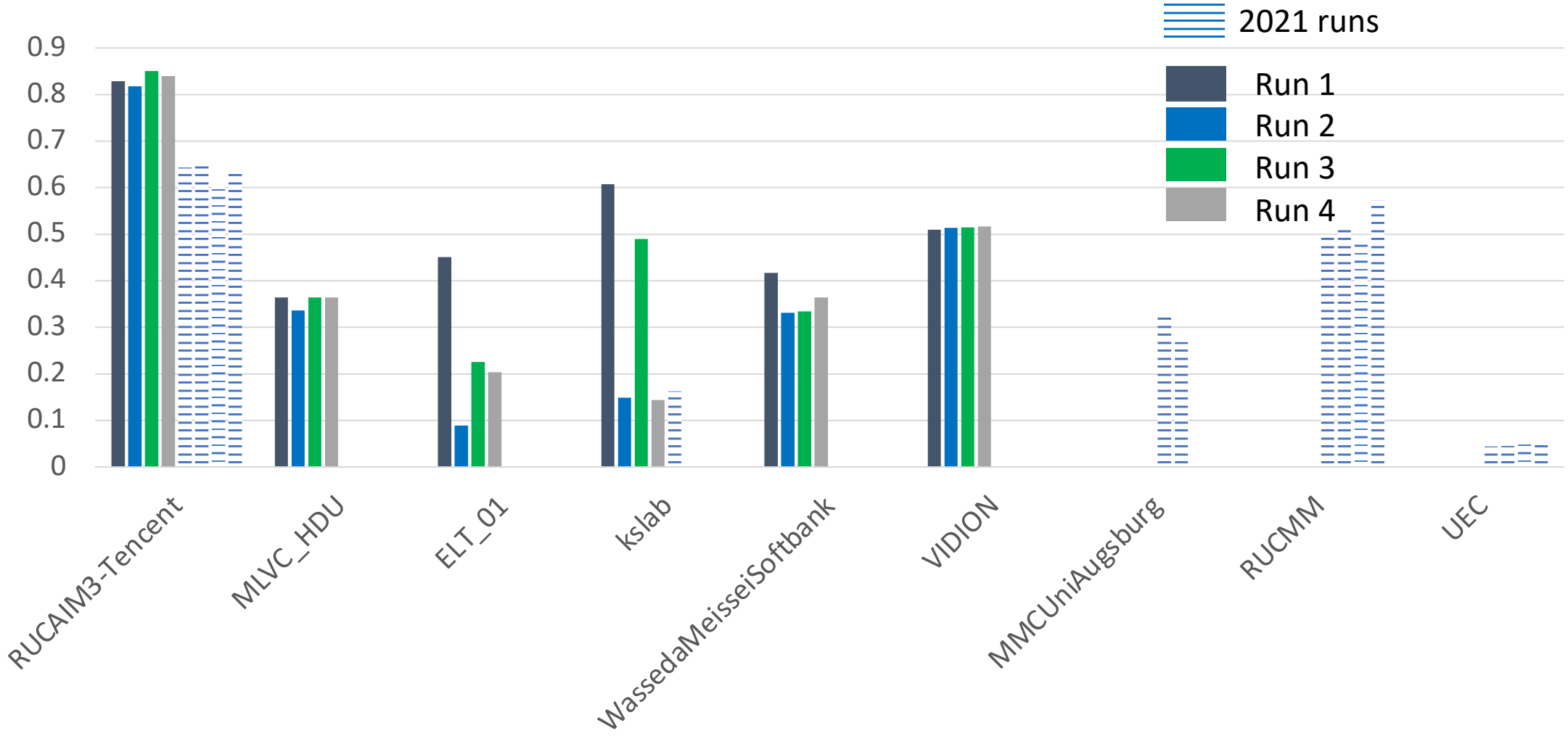
Progress subtask - BLEU Results



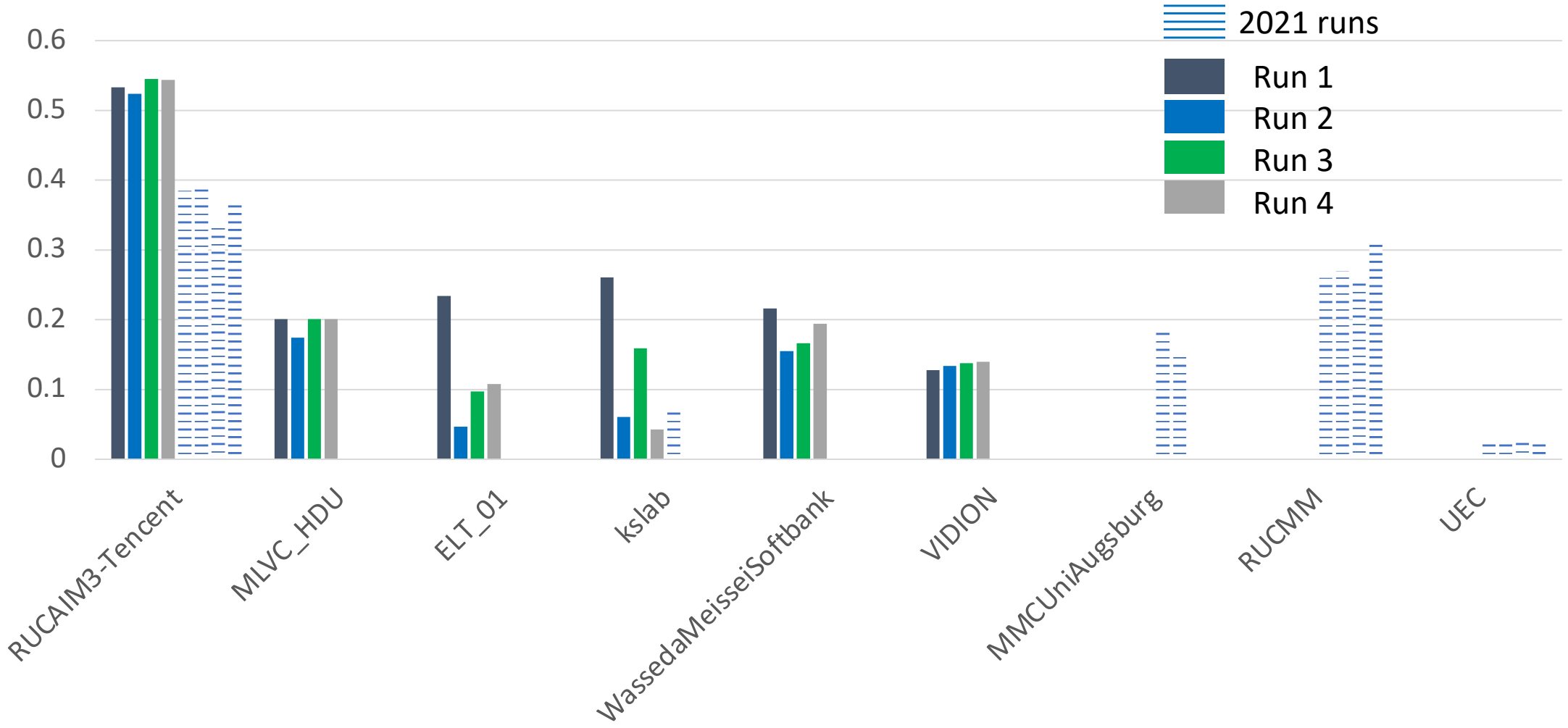
Progress subtask - METEOR Results



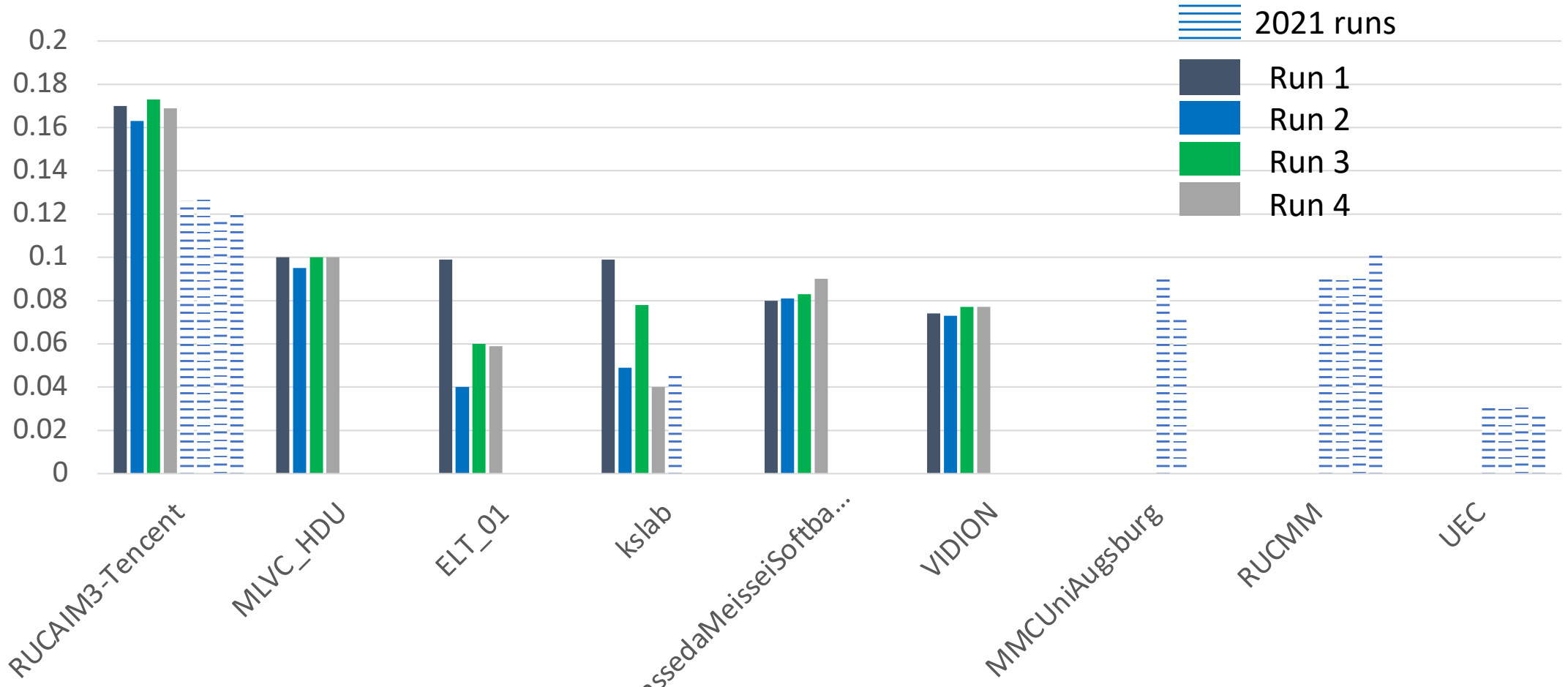
Progress subtask - CIDER Results



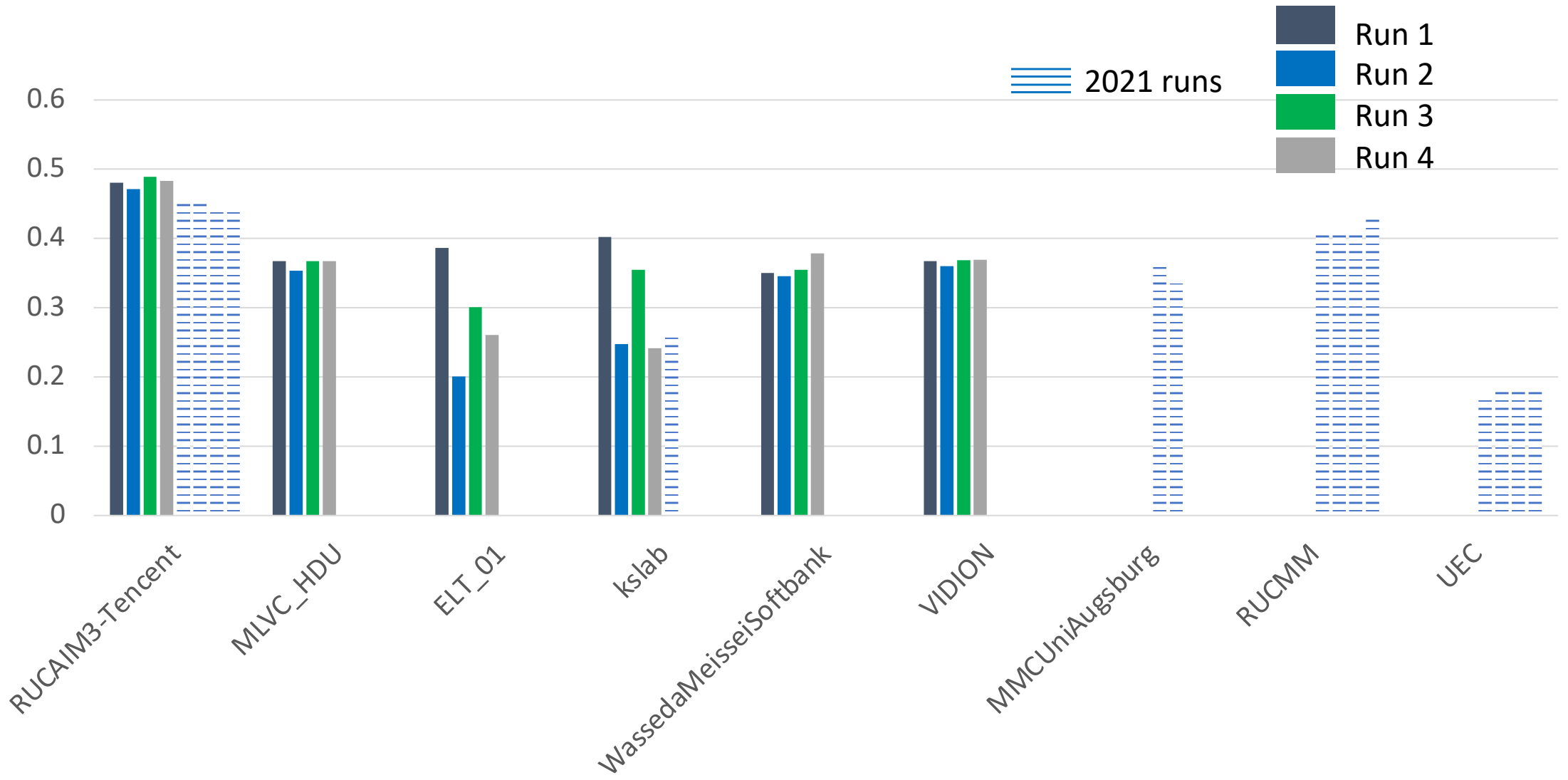
Progress subtask - CIDER-D Results



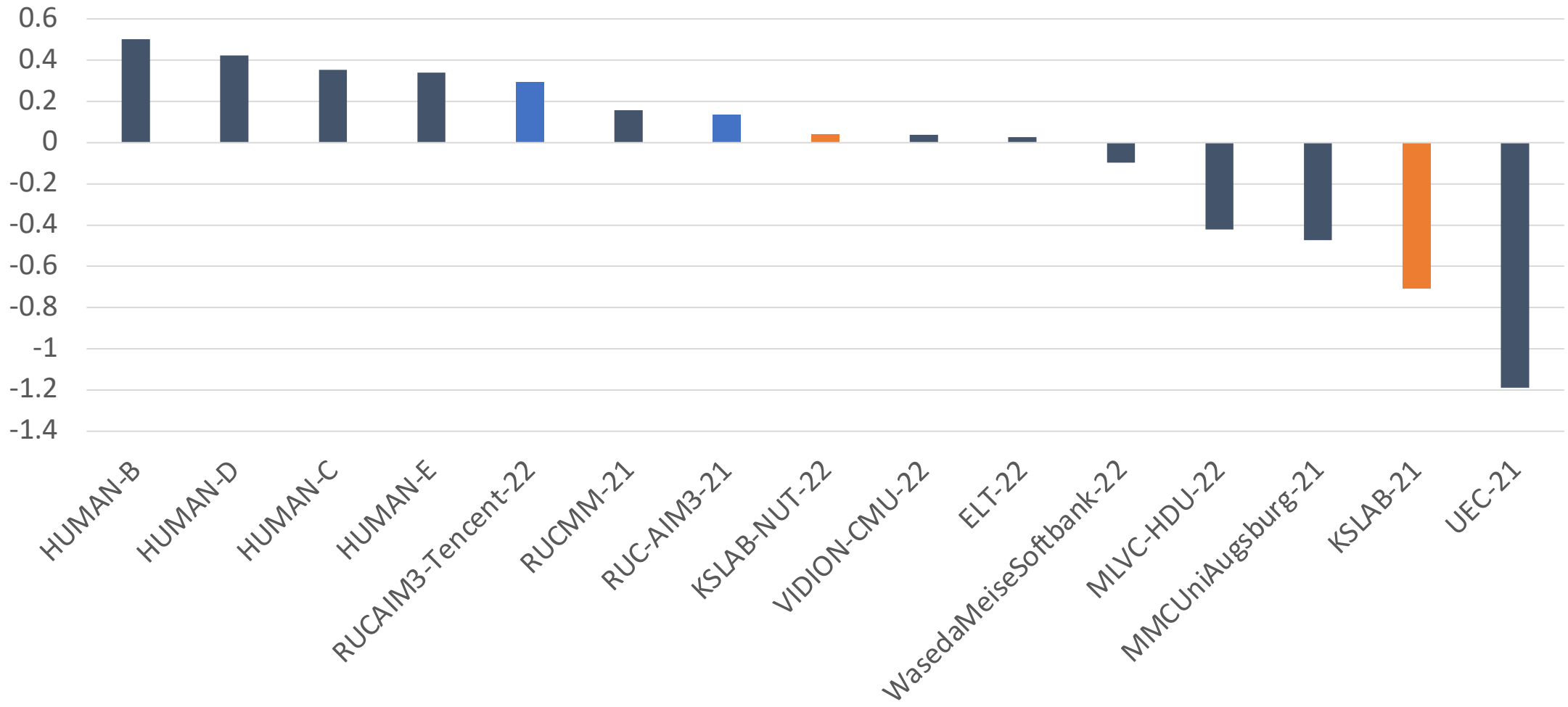
Progress subtask - SPICE Results



Progress subtask - STS Results



Progress subtask - DA Results - Z



Examples (GT vs Submissions)

GT:

1- A person using a small black tool to unscrew small screws on a device with white taped handlebars in a room with white brick walls.
2- A man is tightening something indoors.
3- A man is showing how to tighten the screws of some bolts in a bar wrapped in white rubber.
4- A person's hands are screwing in two screws into a black square object attached to a white curved tube inside a room.
5- A man uses a black instrument to screw something onto a white bar.

Submissions:

1- A pair of hands attaching a camera to a bike strap.
2- a man in a blue shirt is using a machine
3- a man is working on a bicycle handlebar in a room
4- a man working on a bike
5- an older woman is working on something in her home studio
6- a man is using a tool to make adjustments on a bicycle.

Examples (GT vs Submissions)

GT:
1- A bride in white is holding a microphone in her hand and is talking to the guests around the table next to big windows on a sunny day.
2- A bride stands next to her groom talking to him with a microphone in her hand.
3- A bride is standing up and thanking someone while her groom is sitting down and looking at her and another lady is wiping her tears in an open place in the daytime.
4- In a reception hall with a large window showing palm trees outside in the daytime, a blonde bride is standing making a speech at her wedding reception to her groom who is seated next to her.
5- A young blonde bride dressed in white is holding a microphone, and standing next to a big window with a view of palm trees, is making a speech while the groom and other guests sitting at a long table are listening.

Submissions:
1- A woman holding a dog at a wedding.
2- A groom in a white dress with a flower in her hand and a woman in a white dress with a flower in front of her
3- a bride in a white wedding dress is standing in front of a group of people in a room with large windows
4- a bride and groom in front of a window
5- a man is on a stage and is giving a speech to a woman.
6- an african american bride and groom are standing by the window in their wedding day

Conclusion

- This was the first year using the V3C1 test data (following two years of V3C2).
- Participation in the task is stable.
- Few teams used audio features.
- 2nd year for the progress subtask. We hope more existing teams submit again next year to measure progress over 3 years.
- High correlation between all automatic metrics.
- In general, the metrics reported higher scores compared to 2021 (caution: different testing dataset, but same domain).

Thank you!