TRECVID 2023 - A series of evaluation tracks in video understanding

George Awad {gawad@nist.gov}
National Institute of Standards and Technology, USA

Keith Curtis {keith.curtis@tus.ie}
Technological University of the Shannon, Ireland

Asad A. Butt {asad.butt@nist.gov}
Johns Hopkins University;
Information Access Division, National Institute of Standards and Technology, USA

Jonathan Fiscus {jfiscus@nist.gov} Afzal Godil {godil@nist.gov}
Yooyoung Lee {yooyoung@nist.gov} Eliot Godard {eliot.godard@gmail.com}
Information Access Division, National Institute of Standards and Technology, USA

Lukas Diduch {lukas.diduch@nist.gov}
Dakota-consulting, USA

Deepak Gupta {deepak.gupta@nih.gov} Dina Demner Fushman {ddemner@mail.nih.gov}
National Library of Medicine, National Institutes of Health, USA

Yvette Graham {graham.yvette@gmail.com}
ADAPT Centre, Trinity College Dublin, Ireland

Georges Quénot {Georges.Quenot@imag.fr}
Laboratoire d'Informatique de Grenoble, France

April 22, 2024

1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) is a TREC-style video analysis and retrieval evaluation with the goal of promoting progress in research and development of content-based exploitation and retrieval of information from digital video via open, tasks-based evaluation supported by metrology.

Over the last two decades this effort has yielded a better understanding of how systems can effectively accomplish such processing and how one can reliably benchmark their performance. TRECVID has been funded by NIST (National Institute of Standards and Technology) and other US government agencies. In addition, many organizations and individuals worldwide contribute significant time and effort.

TRECVID 2023 planned the following five tasks. From which, four tasks (AVS, VTT, DVU, & ActEV) continued from previous years, while a pilot task (MedVidQA) was introduced. In total, 26 teams from various research organizations worldwide signed up to join the evaluation campaign this year, where 16 teams (Table 1) completed one or more of the following five tasks, and 10 teams registered but did not submit any runs.

1. Ad-hoc Video Search (AVS)
2. Video to Text (VTT)
3. Deep Video Understanding (DVU)
4. Activities in Extended Video (ActEV)
5. Medical Video Question Answering (MedVidQA)

This year TRECVID continued the usage of the Vimeo Creative Commons collection dataset (V3C1 and V3C2) [Rossetto et al., 2019] of about 2,300 hours in total and segmented into 1.5 million short video shots to support the Ad-hoc video search task. The dataset is drawn from the Vimeo video sharing website under the Creative Commons licenses and reflects a wide variety of content, style, and source devices determined only by the self-selected donors. The VTT task also adopted a subset of 2000 short videos from the Vimeo V3C3 dataset.

For the ActEV task, about 16 hours of the Multiview Extended Video with Activities (MEVA) dataset was used which was designed to be realistic, natural and challenging for video surveillance domains in terms of its resolution, background clutter, diversity in scenes, and human activity/event categories.

The same licensed movie dataset of about 15 hours acquired in 2022 from KinoLorberEdu[1] was applied to the DVU task. In addition, a set of 14 Creative Common (CC) movies (total duration of 17.5 hr) previously utilized between 2020 and 2022 ACM Multimedia DVU Grand Challenges including their movie-level and scene-level annotations are being utilized as development dataset for the DVU task. The movies have been collected from public websites such as Vimeo and the Internet Archive. In total, the 14 movies consist of 621 scenes, 1572 entities, 650 relationships, and 2491 interactions.

The AVS results were judged by NIST human assessors, while the VTT and DVU task ground-truth was created by NIST human assessors and scored automatically later using Machine Translation (MT) metrics and Direct Assessment (DA) by Amazon Mechanical Turk workers on sampled runs. The systems submitted for the ActEV task evaluations were scored by NIST using reference annotations created by Kitware, Inc.

This paper is an introduction to the tasks, data, evaluation framework, and performance measures used in the 2023 evaluation campaign. For detailed information about the approaches and results, the reader should see the various site reports and the results pages available at the workshop proceeding online page [TV23Pubs, 2023]. Finally, we would like to acknowledge that all work presented here has been cleared by RPO (Research Protection Office)[2]

Disclaimer: Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

2 Datasets

Many datasets have been adopted and used across the years since TRECVID started in 2001 and all available resources and datasets from previous years can be accessed from our website[3]. In the following sections we will give an overview of the main datasets used this year across the different tasks.

2.1 DVU Movies Training Dataset

The dataset consisted of two types of movie data with a total of 19 movies (23 hr) to support the Deep Video Understanding (DVU) task: The first is a set of 14 Creative Common (CC) movies (total duration of 17.5 hr) previously utilized in 2020 - 2022 ACM Multimedia DVU Grand Challenges including their movie-level and scene-level annotations. The movies have been collected from public websites such as Vimeo and the Internet Archive. In total, the 14 movies consist of 621 scenes, 1572 entities, 650 relationships, and 2491 interactions. The second is a set of 5 licensed movies from KinolorberEdu platform that have been used as testing data in 2022.

[1] https://www.kinolorberedu.com/

[2] under RPO number: #ITL-17-0025

[3] https://trecvid.nist.gov/past.data.table.html

Table 1: Participants and tasks

Task					Location	TeamID	Participants
MD	AV	DV	VT	AH			
$--$	$--$	$--$	VT	$--$	SAm	$camilouchile$	Uchile
MD	$--$	$--$	$--$	$--$	NAm	$UMass\_BioNLP$	UMass Amherst
$--$	AV	$--$	$--$	$--$	NAm	$suvosree$	Florida Atlantic University
MD	$--$	$--$	$--$	$--$	$Eur + Asia + Aus$	$Delphi$	City University of Hong Kong; University of Oxford; Australian National University
$--$	$**$	$--$	$**$	$--$	$Asia$	$MLVC\_HDU$	Hangzhou Dianzi University
MD	AV	$**$	VT	$**$	$Asia$	NII_UIT	National Institute of Informatics; University of Information Technology VNU-HCM, Vietnam (HCM-UIT)
$--$	$--$	$--$	$--$	$**$	$Asia$	$VIREO$	Singapore Management University City University of Hong Kong
$--$	$--$	DV	$--$	$--$	NAm	CMU_DVU	Carnegie Mellon University
$--$	AV	$--$	$--$	$--$	Eur	$HSMW$	University of Applied Sciences
$**$	$--$	$--$	$--$	$--$	NAm	$UMBCVQA$	University of Maryland Baltimore County
$**$	$--$	$--$	$--$	$--$	NAm	$UNCWAI$	University of North Carolina
$**$	$--$	$--$	$--$	$--$	$Asia$	$VPAI$	Hunan University; National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy Sciences
$**$	$--$	$--$	$--$	$--$	$Eur + Asia$	$doshisha\_uzl$	Doshisha University Institute of Medical Informatics, University of Lubeck DFKI
MD	$--$	$--$	$--$	$--$	Eur	$upvqa$	University Politehnica of Bucharest
$--$	AV	$--$	$--$	$**$	Eur	ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas
$--$	$**$	$--$	$**$	$--$	$Asia$	$BUPT\_MCPRL$	Beijing University of Posts and Telecommunications
$--$	$**$	$--$	$--$	$--$	$Asia$	FDU_AWS	Fudan University Amazon Web Service
$**$	$--$	$--$	$--$	$--$	$Asia$	MI_TJU	Tianjin University
MD	AV	DV	$--$	AH	$Asia$	PKU_WICT	Peking University
$--$	$--$	$--$	$--$	$**$	$Asia$	$RUCMM$	Renmin University of China
$--$	$--$	$--$	$**$	$**$	$Asia$	RUC_AIM3	Renmin University of China
$--$	$--$	$--$	$--$	AH	$Asia$	$TJUMMG$	Tianjin University
$--$	$--$	DV	VT	AH	$Asia$	VRR	Zhongyuan University of Technology
$--$	$--$	$**$	$--$	$**$	$Asia$	WHU_NERCMS	Wuhan University
$--$	$--$	$--$	$**$	$--$	$Asia$	$kslab$	Nagaoka University of Technology
$--$	$**$	$--$	$**$	$**$	$Asia$	$WasedaMeiseiSoftbank$	Waseda University, Meisei University, SoftBank Corporation

Task legend. DV:Deep Video Understanding; VT:Video to Text; AV:Activities in Extended videos; AH:Ad-hoc search; MD: Medical Video Question Answering; $--$:no run planned; $**$:submitted run(s)

2.2 Kinolorberedu Testing Dataset

A set of 5 movies licensed from Kino Lorber Edu (https://www.kinolorberedu.com/) is made available to support the deep video understanding task. All movies are in English with a duration between 1.5 - 2 hrs each. Participants were able to download the whole original movies, scenes boundary reference, and a few image examples of key characters and locations.

For the DVU robustness subtask, we created three additional datasets by adding noise to the main task testing dataset with, audio noise only, video noise only, and with both audio and video noise.

2.3 Vimeo Creative Commons Collection (V3C) Dataset

Two sub-collections (V3C1 and V3C2) [Rossetto et al., 2019] have been adopted to support the AVS task. Together, they are composed of about 17,000 Vimeo videos (2.9 TB, 2300 h) with Creative Commons licenses and a mean duration of 8 min. All videos have some metadata available such as title, keywords, and description in json files. They have been segmented into 2 508 113 short video segments according to the provided master shot boundary files. In addition, keyframes and thumbnails per video segment have been extracted and made available. V3C2 was used for testing, while V3C1 was available for development along with the previous Internet Archive datasets (IACC.1-3) of about 1800 h. In addition to the above, a third subset of short videos from the sub-collection V3C3 dataset was used to test the Video to Text systems.

2.4 MEVA Dataset

The ActEV Sequestered Data Leaderboard (SDL) competition is based on the Multiview Extended Video with Activities (MEVA) dataset ([Kitware, 2020] mevadata.org) which was collected and annotated specifically for the development and evaluation of public safety video activity detection capabilities at the Muscatatuck Urban Training Center by Kitware, Inc. for the IARPA DIVA (Deep Intermodal Video Analytics) program and the broader research community. This dataset contains time-synchronized multi-camera, continuous, long-duration video, often taken at significant stand-off ranges from the activities. Metadata and auxiliary data for the site were provided as is typical for public-safe scenarios where detailed knowledge of the site is available to systems. Provided data will include a map and 3D site model of the test area, approximate camera locations for the publicly released video data, and camera models for released sensor video. The dataset was collected with both EO (Electro-Optical) and IR (Infrared) sensors, with over 100 actors performing in various scripted and non-scripted activities in various scenarios. The activities included person and multi-person activities, person-object interaction activities, vehicle activities, and person-vehicle interaction activities.

The dataset was captured with off-the-shelf cameras. Both overlapping and non-overlapping views are in the data set. There are 25 EO cameras and 4 IR cameras. The IR cameras are paired with EO cameras with roughly the same location and orientation. The spatial resolution of the EO cameras is 1920x1080 or 1920x1072 and the IR cameras is 352x240. All the video cameras have a frame rate of 30 frames/second, have a fixed orientation except one, and all are synchronized with the GPS time signal. The number of indoor cameras is 11 and the number of outdoor cameras is 18. Figure 1 shows different image montages of randomly selected videos[4]

Test Data

The TRECVID'23 ActEV Self-Reported Leaderboard (SRL) test dataset is a 16-hour collection of videos with 20 activities, which only consists of Electro-Optics (EO) camera modalities from public cameras. The TRECVID'23 ActEV SRL test dataset is the same as the one used for TRECVID'22 ActEV SRL, CVPR ActivityNet 2022 ActEV SRL and the WACV'22 ActEV SRL challenges.

Training and Development Data

In December 2019, the public MEVA dataset was released with 328 hours of ground-camera data and 4.2 hours of Unmanned Aerial Vehicle video. 160 hours of the ground camera video have been annotated by the same team that has annotated the ActEV test set. Additional annotations have been performed by the public and are also available in the annotation repository.

[4]CC BY-4.0 license

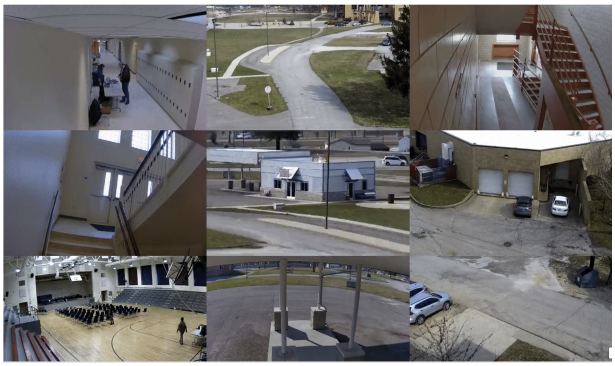

Figure 1: Montage of randomly selected video clips

2.5 TRECVID-VTT

This dataset contains short videos that are between 3 seconds and 10 seconds long. The video sources are from Twitter Vine, Flickr, and V3C2. The dataset is being updated annually and in total, there are 12,870 videos with captions. Each video has between 2 and 5 captions, which have been written by dedicated annotators. The collection includes 6475 URLs from Twitter Vine and 6,395 video files in webm format with Creative Commons License. Those 6,395 videos have been extracted from Flickr and the V3C2 dataset.

For robustness testing, we only created one new dataset by adding noise to the main task test data in both the audio and video channels. The main reason for creating only one dataset was to make it easier for teams to take part in the subtask.

3 Evaluated Tasks

3.1 Ad-hoc Video Search

The Ad-hoc Video Search (AVS) task aims to model the end user video search use case, who is looking for segments of video containing people, objects, activities, locations, etc., and combinations of the former. More focus on fine-grained descriptions was given to provided queries. The task was coordinated by NIST and by the Laboratoire d'Informatique de Grenoble.

The task for participants was defined as the following: given a standard set of master shot boundaries (about 1.4 million shots defined by starting time and ending time in the original whole videos) from the V3C2 test collection and a list of 30 ad-hoc textual queries (see Appendix A and B), participants were asked to return for each query, at most the top 1000 video clips from the master shot boundary reference set, ranked according to the highest probability of containing the target query. The presence of each query was assumed to be binary, i.e., it was either present or absent in the given standard video shot.

Judges at NIST followed several rules in evaluating system output. For example, if the query was true for some frame (sequence) within the shot, then it was true for the shot. In addition, query definitions such as "contains x" or words to that effect are short for "contains x to a degree sufficient for x to be recognizable as x by a human". This means among other things that unless explicitly stated, partial visibility or audibility may suffice. Lastly, the fact that a segment contains video of a physical object representing the query target, such as photos, paintings, models, or toy versions of the target (e.g. picture of Barack Obama vs Barack Obama himself), was NOT grounds for judging the query to be true for the segment. Containing video of the target within video (such as a television showing the target query) may be grounds for doing so. Three main submission types were accepted:

- Fully automatic runs (no human input in the loop): The system takes a query as input and produces results without any human intervention.
- Manually-assisted runs: where a human can formulate the initial query based on topic and query interface, not on knowledge of collection or search results. The system takes the formulated query as input and produces results without further human intervention.
- Relevance-Feedback: The system takes the official query as input and produces initial results, then a human judge can assess the top-30 results and input this information as feedback to the system to produce a final set of results. This feedback loop is strictly permitted for only up to 3 iterations.

In general, runs submitted were allowed to choose any of the following four training types:

- A - used only V3C1 training data
- D - used any other training data (except the testing dataset V3C2)
- E - used only training data collected automatically using only the official query textual description

- F - used only training data collected automatically using a query built manually from the given official query textual description

The training categories "E" and "F" are motivated by the idea of promoting the development of methods that permit the indexing of concepts in video clips using only data from the web or archives without the need for additional annotations. The training data could for instance consist of images or videos retrieved by a general-purpose search engine (e.g., Google) using only the query definition with only automatic processing of the returned images or videos.

The progress subtask objective is to measure system progress on a set of 20 fixed topics (Appendix B). As a result, 2022 systems were allowed to submit results for 20 common topics (not evaluated in 2022) that will be fixed for three years (2022-2024). This year NIST evaluated progress runs submitted in 2022 and 2023 so that teams can measure their progress against two years, while in 2024 they can measure their progress against three years. In general, the 20 fixed progress topics are divided equally into two sets of 10 topics to be evaluated in 2023 and 2024.

A Novelty run type was also allowed to be submitted within the main task. The goal of this run type is to encourage systems to submit novel and unique relevant shots not easily discovered by other runs. In other words, to find rare true positive shots. Finally, teams were allowed to submit an optional explainability parameter with each shot. This was formulated as a keyframe and bounding box to localize the region that supports the query evidence.

Dataset

The V3C2 dataset (drawn from a larger V3C video dataset [Rossetto et al., 2019]) was adopted as a testing dataset. It is composed of 9760 Vimeo videos (1.6 TB, 1300 h) with Creative Commons licenses and a mean duration of 8 min. All videos have some metadata available e.g., title, keywords, and description in json files. The dataset has been segmented into 1 425 454 short video segments according to the provided master shot boundary files. In addition, keyframes and thumbnails per video segment have been extracted and made available. For training and development, all previous V3C1 dataset (1000 h) and Internet Archive datasets (IACC.1-3) with about 1 800 h were made available with their ground truth and xml meta-data files. Throughout this report we do not differentiate between a clip and a shot and thus they may be used interchangeably.

Evaluation

Each group was allowed to submit up to 4 prioritized runs per submission type and per task type (main or progress), and two additional if they were of training type "E" or "F" runs. In addition, one novelty run type was allowed to be submitted within the main task.

In fact, 7 groups submitted a total of 73 runs with 43 main runs and 30 progress runs. One team submitted a novelty run. The 43 main runs consisted of 29 fully automatic, 10 manually-assisted runs, and 4 relevance feedback runs. Progress runs consisted of 19 fully automatic and 11 manually-assisted runs. As the evaluation will also take into consideration progress runs submitted in 2022, there were 23 fully automatic and 5 manually-assisted runs considered for scoring.

To prepare the results from teams for human judgments, a workflow was adopted to pool results from runs submitted. For each query topic, a top pool was created using 100 % of clips at ranks 1 to 300 across all submissions after removing duplicates. A second pool was created using a sampling rate of 25 % of clips at ranks 301 to 1000, not already in the top pool, across all submissions and after removing duplicates. Using these two master pools, we divided the clips in them into small pool files with about 1000 clips in each file. Five human judges (assessors) were presented with the pools - one assessor per topic - and they judged each shot by watching the associated video and listening to the audio then voting if the clip contained the query topic or not. Once the assessor completed judging for a topic, a second round of confirmation judging was conducted to take into consideration close neighborhood shots with opposite judging decisions as well as clips submitted by at least 10 runs at ranks 1 to 200 that were voted as false positive by the assessor. This final step was done as a secondary check on the assessors' judging work and to give them an opportunity to fix any judgment mistakes.

In all, 130 390 clips were judged while 121 415 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 21 234 with 9152 hits at submission ranks from 1 to 100, 7396 hits at submission ranks 101 to 300, and 4686 hits at submission ranks between 301 to 1000. Table 2 presents information about the pooling and judging per topic.

Measures

Work at Northeastern University [Yilmaz and Aslam, 2006] has resulted in methods for estimating standard system performance measures using relatively small samples of the usual judgment sets so that larger numbers of features can be evaluated using the same amount of judging effort. Tests on past data showed the measure inferred average precision (infAP) to be a good estimator of average precision [Over et al., 2006]. This year mean extended inferred average precision (mean xinfAP) was used which permits sampling density to vary [Yilmaz et al., 2008]. This allowed the evaluation to be more sensitive to clips returned below the lowest rank (≈300) previously pooled and judged. It also allowed adjustment of the sampling density to be greater among the highest ranked items that contribute more average precision than those ranked lower. The *sample_eval* software [5], a tool implementing xinfAP, was used to calculate inferred recall, inferred precision, inferred average precision, etc., for each result, given the sampling plan and a submitted run. Since all runs provided results for all evaluated topics, runs can be compared in terms of the mean inferred average precision across all evaluated query topics.

Ad-hoc Results

All submissions were of training type "D", and no runs using category "E" or "F" were submitted. It is encouraging to see relevance-feedback runs, as it has been several years that any team submitted any R runs. Tables 3, 4, and 5 show the results of all fully automatic (F), manually-assisted (M), and relevance-feedback (R) runs respectively. In general, for fully automatic results, the top scores and median (0.263) are higher than 2022. The top team (WHU_NERCMS) 4 runs achieved the top 4 places, while other team runs are also within close performance. For manually-assisted runs, we had two participating teams (VIREO and NII_UIT). Overall, compared to automatic runs, manually-assisted runs performed lower (with a median score of 0.1875) and comparing the performance of common teams we can see that team NII_UIT team top M run performed better than their top F run, however team VIREO top M run performed lower than their top F run. Regarding relevance-feedback runs, they all came from one team (WHU_NERCMS) with an overall median score of 0.2985 and top score (0.299) exceeding the top automatic and manual runs.

Run ID (appended with priority)	Mean xInfAP
WHU_NERCMS.23_2	0.292
WHU_NERCMS.23_1	0.292
WHU_NERCMS.23_3	0.291
WHU_NERCMS.23_4	0.29
WasedaMeiseiSoftbank.23_2	0.285
WasedaMeiseiSoftbank.23_4	0.281
RUCMM.23_1	0.272
WasedaMeiseiSoftbank.23_3	0.27
WasedaMeiseiSoftbank.23_1	0.269
RUC_AIM3.23_1	0.269
VIREO.23_4	0.268
RUCMM.23_3	0.268
RUCMM.23_2	0.268
RUC_AIM3.23_2	0.267
RUC_AIM3.23_3	0.263
RUC_AIM3.23_4	0.262
RUCMM.23_4	0.261
VIREO.23_3	0.256
ITI_CERTH.23_3	0.24
VIREO.23_1	0.237
VIREO.23_5	0.235
ITI_CERTH.23_4	0.233
ITI_CERTH.23_1	0.225
ITI_CERTH.23_2	0.224
VIREO.23_2	0.215
NII_UIT.23_1	0.166
NII_UIT.23_3	0.164
NII_UIT.23_2	0.16
NII_UIT.23_4	0.158

Table 3: AVS: Sorted scores of 29 automatic runs across all 20 main queries. All runs used training type "D".

To test if there were significant differences between the runs submitted, we applied a randomization test [Manly, 1997] on the top 10 runs for each run type category using a significance threshold of $p<0.05$.

For automatic runs, the analysis showed there is no statistical difference between Waseda team runs 2 and 4 and between runs 1 and 3, and there is a statistical difference between runs 2 & 4 and 1 & 3. Team WHU_NERCMS run 1 is better than team RUC_AIM3 run 1, and top 4 runs of WHU_NERCMS are not significantly better from each other. With respect to manually-assisted runs, the test indicated

[5] http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/

Table 2: Ad-hoc search pooling and judging statistics

Topic number	Total submitted	Unique submitted	total that were unique %	Number judged	unique that were judged %	Number relevant	judged that were relevant %
1681	57992	52854	91.14	5313	10.05	441	8.30
1683	57947	49205	84.91	4104	8.34	1107	26.97
1685	57975	48981	84.49	3854	7.87	1090	28.28
1687	57993	53584	92.40	9229	17.22	381	4.13
1689	57985	53223	91.79	6665	12.52	446	6.69
1691	57867	50865	87.90	4079	8.02	460	11.28
1693	57954	52938	91.34	3664	6.92	971	26.50
1695	57957	52927	91.32	4380	8.28	766	17.49
1697	57924	52615	90.83	2736	5.20	310	11.33
1699	57988	53281	91.88	6168	11.58	144	2.33
1731	43000	42096	97.90	3683	8.75	1100	29.87
1732	43000	41545	96.62	3149	7.58	1005	31.91
1733	43000	38144	88.71	2600	6.82	298	11.46
1734	43000	38433	89.38	2874	7.48	1133	39.42
1735	43000	38341	89.17	4825	12.58	732	15.17
1736	43000	42538	98.93	3948	9.28	326	8.26
1737	43000	42230	98.21	3940	9.33	1790	45.43
1738	43000	42116	97.94	3161	7.51	369	11.67
1739	43000	40442	94.05	3691	9.13	249	6.75
1740	43000	42276	98.32	6185	14.63	1507	24.37
1741	43000	40426	94.01	4149	10.26	475	11.45
1742	43000	42364	98.52	4192	9.90	110	2.62
1743	43000	42587	99.04	5819	13.66	46	0.79
1744	43000	42337	98.46	4112	9.71	345	8.39
1745	43000	42193	98.12	4161	9.86	2496	59.99
1746	43000	42188	98.11	4053	9.61	249	6.14
1747	43000	40428	94.02	3264	8.07	267	8.18
1748	43000	42322	98.42	3978	9.40	439	11.04
1749	43000	42312	98.40	3177	7.51	1826	57.48
1750	43000	42219	98.18	5237	12.40	356	6.80

that there is no statistical difference between VIREO runs 1, 4, and 5. And also, no statistical difference between NII_UIT runs 1,2,3 and 4. Finally for R runs it was indicated that there is no difference between the top 2 runs, while run 1 is better than run 4.

Figure 2 shows for each topic the number of relevant and unique shots submitted by all teams combined (blue color). On the other hand, the orange bars show the total non-unique true shots submitted by at least 2 or more teams. The chart is sorted by the number of unique hits. One-third of all hits are unique.

The four topics: 1745, 1740, 1749, and 1737 achieved the most unique hits while also reporting a high number of hits overall, while the three topics: 1697, 1742, and 1733 reported the lowest unique hits. In general, topics that reported a high number of hits consisted of both unique and non-unique hits, while topics that reported a low number of hits mainly only consisted of non-unique hits, representing the difficulty of the query. While it is hard to draw conclusions about why hits vary by topic, there seems to be a correlation between the relative easiness of the query and its components (e.g. more actions/activities in combination with objects or conditions (spatial or temporal) are harder and are being detected less). We should also note here that high/low hits per topic don't necessarily mean

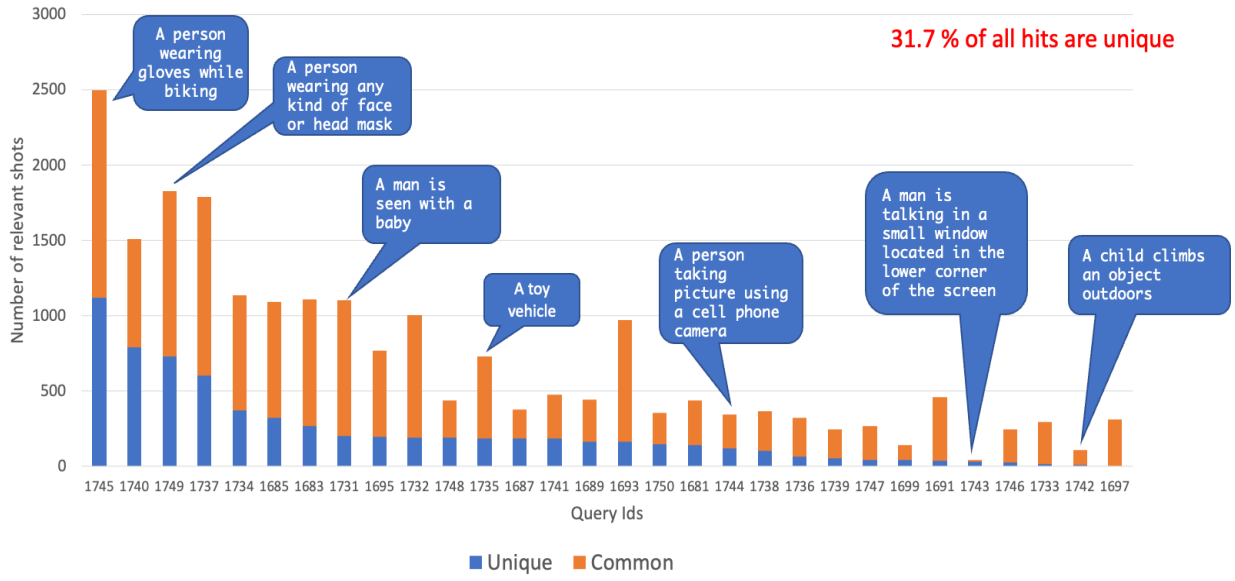

Figure 2: AVS: Unique vs overlapping results in main task

Run ID (appended with priority)	Mean xInfAP
C_VIREO.23_4	0.25
N_VIREO.23_5	0.222
C_VIREO.23_1	0.222
C_NII_UIT.23_4	0.189
C_NII_UIT.23_3	0.188
C_NII_UIT.23_2	0.187
C_NII_UIT.23_1	0.186
C_VIREO.23_3	0.072
C_VIREO.23_6	0.041
C_VIREO.23_2	0.002

Table 4: AVS: Sorted scores of 10 manually-assisted runs across all 20 main queries. All runs used training type "D". Run names are prefixed by "C" (common) or "N" (novelty)

high/low performance in the final scores as a good run must detect and rank results high as well.

Run ID (appended with priority)	Mean xInfAP
WHU_NERCMS.23_3	0.299
WHU_NERCMS.23_1	0.299
WHU_NERCMS.23_2	0.298
WHU_NERCMS.23_4	0.296

Table 5: AVS: Sorted scores of 4 relevance-feedback runs across all 20 main queries. All runs used training type "D".

Team	Relevant shots
VIREO	4167
NII_UIT	892
ITI_CERTH	534
WasedaMeiseiSoftbank	360
RUC_AIM37	357
RUCMM	230
WHU_NERCMS	138
kindai_ogu_osaka	52

Table 6: AVS: Sorted unique number of hits (true positive shots) by team.

Table 6 shows the number of unique clips found by the different participating teams. From this figure and the overall scores in Tables 3, 4, and 5, it can be shown that there is no clear relation between the teams who found the most unique shots and their to-

tal performance. The VIREO team contributed the most unique hits (similar to previous year). Although WHU_NERCMS, RUCMM, and Waseda teams performed well, their unique hits contributions were not very high.

Figures 3 show the performance of the top 10 runs across the 20 main queries for automatic runs. Note that each series in this plot represents a rank (from 1 to 10) of the scores, but all scores at a given rank do not necessarily belong to a specific team. A team's scores can rank differently across the 20 queries. Some samples of top and bottom performing queries are highlighted with the query text. Harder queries are those that include non-traditional combinations of concepts (e.g. A man with an earring in his left ear). In general, for automatic systems and topics not performing well, usually all top 10 runs are condensed together with low spread between their scores, while mid or high performing queries may vary in their range of performance.

The novelty run type encourages submitting unique (hard to find) relevant shots. Systems were asked to label their runs as either novelty type (N) or common type (C). The novelty metric was designed to score runs based on how good they are at detecting unique relevant shots. A weight was given to each topic and shot pair such as follows:

$$TopicX\_ShotY_{weight}(x) = 1 - \frac{N}{M}$$

where N is the number of times Shot Y was retrieved for topic X by any run submission, and M is the number of total runs submitted by all teams. For instance, a unique relevant shot weight will be close to 1.0 while a shot submitted by all runs will be assigned a weight of 0.

For a run R and for all topics, we calculate the summation S of all unique shot weights only, and the final novelty metric score is the mean score across all evaluated 20 topics. Figure 4 shows the novelty metric scores. The red bars indicate the single submitted novelty run.

For a team that submitted a novelty run, we removed all its other common runs submitted. The reason for doing this was the fact that usually for a given team there would be many overlapping shots within all its submitted runs. For other teams who did not submit novelty runs, we chose the best (top-scoring) run for each team for novelty metric calculations purposes. As shown in the figure, the novelty run (by VIREO team) scored best based on our metric. More runs are needed to conduct a better comparison within novelty systems.

Among the submission requirements, we asked teams to submit the processing time that was consumed to return the result sets for each query. Figure 5 plots the reported processing times vs the InfAP scores among all run queries for automatic runs.

It can be seen that spending more time did not necessarily help in most cases and few queries achieved high scores in less time. There is more work to be done to make systems efficient and effective at the same time. In general, most automatic systems reported processing time below 10 s.

The progress task results are shown in Table 7 for automatic and manually-assisted systems. In total, 7 teams participated in this progress task for the last two years. Comparing the best run in these two years for each team, we can see that for automatic Systems all teams submitted in both years achieved better in 2023, two teams submitted in 2022 but not in 2023, and one team submitted in 2023 but not in 2022. For manually-assisted systems, only VIREO submitted in both years being 2022 submission better.

To analyze in general which topics were the easiest and most difficult we sorted topics by the number of runs that scored above or below the midpoint score of xInfAP >= 0.5 for any given topic and assumed that those runs with 0.5 or above were the easiest topics, while topics with xInfAP < 0.5 were assumed hard topics. From this analysis, it can be concluded that the top 5 hard topics were: "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", and "A person opens a door and enters a location". On the other hand, the top 5 easiest topics were: "A person wearing gloves while biking", "A man is seen with a baby", "A person wearing any kind of face or head mask", and "A woman wearing (dark framed) glasses", and "A woman with red hair".

Ad-hoc Observations and Conclusions

Compared to the semantic indexing task that was conducted to detect single concepts (e.g., airplane, animal, bridge) from 2010 to 2015 it can be seen from running the ad-hoc task the last 7 years that it is still very hard and systems still have a lot of room to research methods that can deal with unpredictable queries composed of one or more concepts including their interactions, relationships and conditions. From

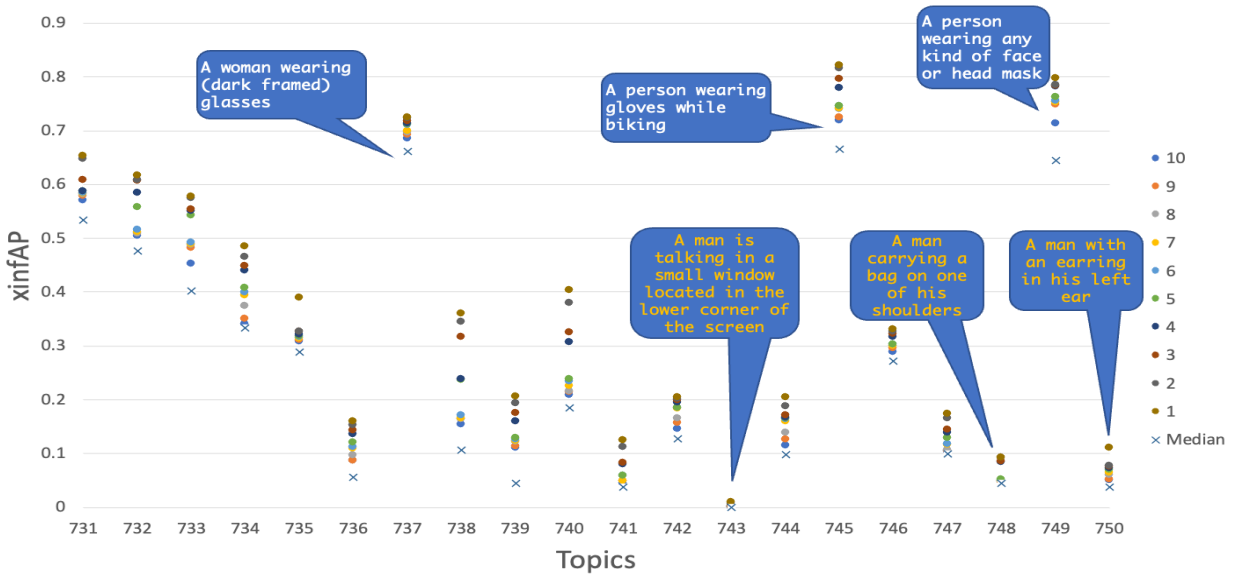

Figure 3: AVS: Top 10 runs (xinfAP) per query (fully automatic)

Team	Automatic systems	Manually-assisted systems
RUCMM (2022)	0.24	
RUCMM (2023)	0.26	
VIREO (2022)	0.14	0.149
VIREO (2023)	0.17	0.134
NII_UIT (2023)	0.15	0.15
ITI_CERTH (2022)	0.19	
ITI_CERTH (2023)	0.22	
RUCAIM3-Tencent (2022)	0.19	
kindai_ogu_osaka (2022)	0.21	
WasedaMeiseiSoftbank (2022)	0.26	
WasedaMeiseiSoftbank (2023)	0.29	

Table 7: AVS: Max performance (xInfAP score) per team on 10 progress queries

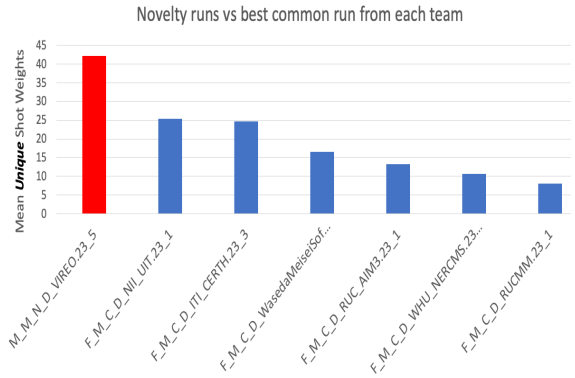

Figure 4: AVS: Novelty Runs Scores

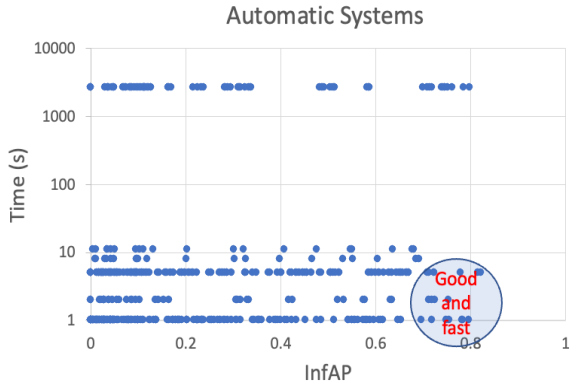

Figure 5: AVS: Processing time vs scores (fully automatic)

2016 to 2021 we concluded two cycles of six years running the Ad-hoc task using the Internet Archive (IACC.3) dataset [Awad et al., 2016] and the Vimeo Creative Commons Collection (V3C1). Starting in 2022, we are using a new sub-collection from Vimeo (V3C2) as the official testing dataset.

To summarize major observations in 2023 we can see that overall team participation and task completion rates are stable. All submitted runs were of training type "D", and no runs of type "E" or "E" were submitted. One novelty run type was submitted. Overall, 43 systems (29 automatic, 10 manually-assisted, and 4 relevance-feedback) were submitted in the main task including 1 novelty run, while 30 runs were submitted for the progress task. Overall, performance scores are higher than last year which is encouraging given that queries are still focused on fine-grained information. Few automatic systems are good and fast (< 10 sec). There exists a high similarity between automatic, manually-assisted, and relevance feedback systems in terms of query performance relative to each other. The top-scoring teams did not necessarily contribute a lot of unique true shots and vice-versa. About 32% of all hits are unique, while 68% are common hits across the submitted runs. Overall, 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 commonly found in the available training datasets). For low-performance queries, usually all systems are condensed in a small range. While for mid to high performance queries, the top 10 runs vary in their range of performance.

As a general high-level systems overview, we observe the use of multiple text-image and text-video common latent embedding approaches such as VSE++, CLIP and its various variants: SLIP, BLIP, BLIP-2, LaCLIP, OpenCLIP, and TeachCLIP. Some teams applied query expansion with ChatGPT, while others made use of Text-to-Image generative approaches. The majority of systems used a transformer-based extension of a cross-modal deep network architectures. An interesting approach was based on top-K feedback and proposed a new algorithm Quantum-Theoretic Interactive Ranking Aggregation (QT-IRA) that adjusts models' weight with relevance feedback.

No teams used the previous popular concept banks. However, the focus is more towards "dual task" (interpretable embeddings). In terms of datasets, multiple text-image and text-video annotated collections

such as MSR-VTT, TGIF, Vatex, Flickr8k/30k, MS-COCO, and Conceptual Captions were used for training systems. Teams experimented with several combinations and fusion approaches (e.g. normalization, averaging), as well as lightweight attentional feature fusion methods. Finally, it is hard to distinguish between data or feature effects and algorithmic effects.

For detailed information about the approaches and results for individual teams, we refer the reader to the reports [TV23Pubs, 2023] in the online workshop notebook proceedings.

3.2 Deep Video Understanding

Deep video understanding is a challenging task that requires systems to develop a deep analysis and understanding of the relationships between different entities in video, to use known information to reason about other, more hidden information, and to populate a knowledge graph (KG) representation with all acquired information [Curtis et al., 2020]. To work on this task, a system should take into consideration all available modalities (speech, image/video, and in some cases text). The aim of this task is to push the limits of multi-modal extraction, fusion, and analysis techniques to address the problem of analyzing long duration videos holistically and extracting useful knowledge to utilize it in solving different types of queries. The target knowledge includes both visual and non-visual elements. As videos and multimedia data are getting more and more popular and usable by users in different domains and contexts, the research, approaches and techniques we aim to be applied in this task will be very relevant in the coming years and near future.

Dataset

The Deep Video Understanding Training Set described in Table 8 consists of 19 Creative Commons (CC) license movies with a total duration of about 25 hours[6]. This training set has been annotated by human assessors and final ground truth, both at the overall movie level (Ontology of relations, entities, actions & events, Knowledge Graph, and names and images of all main characters), and the individual scene level (Ontology of locations, people/entities, interactions and their order between people, sentiments, and text summary) has been be provided to participating researchers for training and development of their systems. In summary, we hired 5 annotators in addition to a summer student. On average each movie took about 20 hours of work to annotate both movie and scenes. A sample from a scene-level knowledge graph annotation can be seen in Figure 7. For more detailed information about the annotation framework please refer to our paper at [Loc et al., 2022].

The DVU Test Set described in Table 9 contains 5 movies licensed from KinoLorberEdu[7] platform with a total duration of about 6 hours. Participants were required to complete a data access form in order to access these movies. The testing set was fully annotated by human annotators to the same degree as the training set. A set of queries, described in more detail in Section 3, were then automatically extracted from human annotations and released to participants, along with the set of movies and annotated images of the movie characters identified during annotation.

Further information about movies' genres and duration are provided below in Tables 8 and 9.

Annotation

Human assessors annotated each movie of the full DVU dataset. Full movies were annotated to a Knowledge Graph (KG) indicating the relationships and connections between every major character, entity, and concept in the movie. Images of each character and entity were also provided. Figure 6 shows an example of a movie-level KG. Following this, every scene within each movie was also annotated to a scene-level KG indicating the locations and characters within each scene, at least one sentiment label for that scene, non-neutral mental states of characters, the interactions between characters, and the ordering of the interactions as they happened in that scene. Figure 7 shows an example of a scene-level KG.

System task

The Deep Video Understanding task was as follows: given a whole original **movie** (e.g. 1.5 - 2hrs long), **image snapshots** of main entities (persons, locations, and concepts) per movie, and **ontology** of relationships, interactions, locations, and sentiments

[6] https://www-nlpir.nist.gov/projects/trecvid/dvu/dvu.development.dataset/

[7] `https://www.kinolorber.com/`

Movie	Genre	Duration
Honey	Romance	86 min
Let's Bring Back Sophie	Drama	50 min
Nuclear Family	Drama	28 min
Shooters	Drama	41 min
Spiritual Contact - The Movie	Fantasy	66 min
Super Hero	Fantasy	18 min
The Adventures of Huckleberry Finn	Adventure	106 min
The Big Something	Comedy	101 min
Time Expired	Comedy / Drama	92 min
Valkaama	Adventure	93 min
Bagman	Drama / Thriller	107 min
Manos	Horror	73 min
Road to Bali	Comedy / Musical	90 min
The Illusionist	Adventure/Drama	109 min
Chained for Life	Comedy / Drama	88 min
Liberty Kid	Drama	88 min
Calloused Hands	Drama	92 min
Like Me	Horror / Thriller	79 min
Losing Ground	Comedy / Drama	81 min

Table 8: The full DVU training set

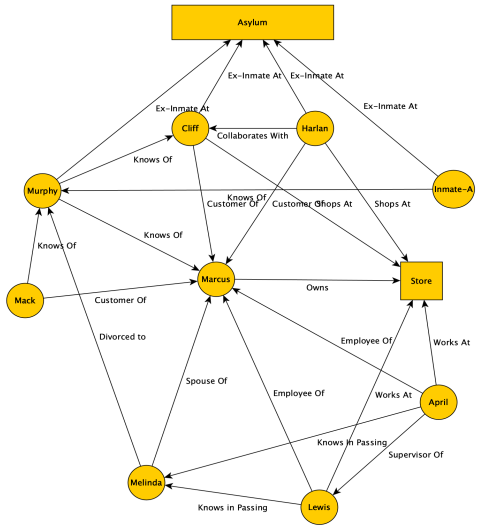

Figure 6: Movie-level KG sample

Movie	Genre	Duration
Archipelago	Drama	114 min
Bonneville	Drama	93 min
Heart Machine	Drama	85 min
Littlerock	Drama	82 min
Memphis	Drama	79 min

Table 9: The full DVU testing set

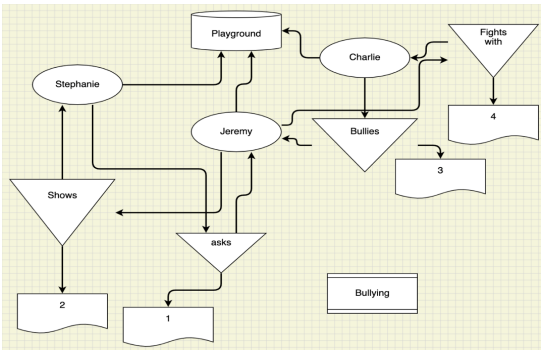

Figure 7: Scene-level KG sample

used to annotate each movie at global movie-level (relationships between entities) as well as on fine-grained scene-level (scene sentiment, interactions between characters, and locations of scenes), systems were expected to generate a knowledge-base of the main actors and their relations (such as family, work, social, etc) over the whole movie, and of interactions between them over the scene level. This representation would be used to answer a set of queries on the movie-level and/or scene-level (see below details about query types) per movie. The task supported two tracks (sub-tasks) where teams could join one or both tracks. The Movie track was comprised of queries on the whole movie level, and the Scene track was comprised of queries targeting specific movie scenes.

Query Topics & Metrics

Movie-level Track

- **Question Answering**
  This query type represents questions on the resulting knowledge base of the movies in the testing dataset. For example, we may ask 'How many children does Person A have?', in which case participating researchers should count the 'Parent Of' relationships Person A has in the Knowledge Graph. These queries also contain human-generated questions (open domain questions) which are not limited to the ontology. This query type takes a multiple choice questions format.

- **Fill in the Graph Space**
  Fill in spaces in the Knowledge Graph (KG). Given the listed relationships, events or actions for certain nodes, where some nodes are replaced by variables X, Y, etc., solve for X, Y etc. Example of The Simpsons: X Married To Marge. X Friend Of Lenny. Y Volunteers at Church. Y Neighbor Of X. Solution for X and Y in that case would be: X = Homer, Y = Ned Flanders.

Scene-level Track

- **Find Next or Previous Interaction**
  Given a specific scene and a specific interaction between person X and person Y, participants are asked to return either the previous interaction or the next interaction, in either direction, between person X and person Y. This can be specifically the next or previous interaction within the same scene, or over the entire movie. This query type takes a multiple choice questions format and it is considered a mandatory query in the scene-level track).

- **Find Unique Scene**
  Given a full, inclusive list of interactions, unique to a specific scene in the movie, teams should find which scene this is.

- **Find the 1-to-1 relationship between scenes and natural language descriptions**
  Given a set of scenes, and a set of natural language descriptions of movie scenes, match the correct natural language description for each scene.

- **Classify scene sentiment from a given scene** Given a specific movie scene and a set of possible sentiments, classify the correct sentiment label for each given scene.

Queries for this task were generated semi-automatically by parsing full annotations over the movie-level and the scene-level and populating a data structure with the full knowledge base. Four different sets of questions and accompanying answers for each query type were automatically generated. The TRECVID team then checked questions by hand, taking care to eliminate any questions that were duplicates or near-duplicates of previous questions, or where the question was considered not of sufficient quality to effectively evaluate systems performance.

Metrics

- **Movie-level Q1: Question Answering**
  Scores for this query were produced by calculated by the number of Correct Answers / number of Total Questions.

- **Movie-level Q2: Fill in the Graph Space**
  Results were treated as a ranked list of result items per each unknown variable, and the Reciprocal Rank score was calculated per unknown variable and Mean Reciprocal Rank (MRR) per query.

- **Scene-level Q1: Find Next or Previous Interaction**
  Scores for this query were produced by calculated by the number of Correct Answers / number of Total Questions.

- **Scene-level Q2: Find Unique Scene** Results were treated as a ranked list of result items per each unknown variable, and the Reciprocal Rank score was calculated per unknown variable and Mean Reciprocal Rank (MRR) per query.

- **Scene-level Q3: Find the 1-to-1 relationship between scenes and natural language descriptions** Scores for this query will be calculated by the number of Correct Answers / number of Total Questions.

- **Scene-level Q4: Classify scene sentiment from a given scene** Scores for this query will be calculated by the number of Correct Answers / number of Total Questions.

Evaluation

The advantage of automatically generating questions for this task was that evaluations could be performed automatically. A system was developed to parse correct answers for each query, as well as submitted answers from each participant team's submission. Answers were then compared and an itemized output was generated allowing participating teams to see the correct answers for each query in addition to their submitted query and assigned scores. In general, we divided each of the movie and scene tracks into two query type groups. For example, in scene-level, the first two query types focused on interactions were combined in one group, while the text to scene matching and scene sentiment classification were combined in another group. Each team were allowed to submit runs against any of the tracks and query groups within them.

In total 2 teams (NII_UIT and WHU_NERCMS) submitted runs this year. For details about their systems approaches, we refer the reader to the detailed teams' papers [TV23Pubs, 2023]

Results

Figures 8 and 11 show the overall summary scores of runs from both teams participating in these two subtasks. The WHU_NERCMS team achieved higher results in both movie and scene level queries. In general, the fill in the graph space query performed higher than the question answering queries in movie-level results, while group 2 scene queries (scene to text matching and sentiment classification) performed higher than group 1 consisting of interactions questions.

Figure 9 shows the movie-level results by individual movie for the fill in the graph space query. The movie "Archipelago" achieved the highest scores by both teams, "little_rock" scored the lowest overall, while both teams achieved similar scores on "Memphis". In summary, for this query type runs achieved a high median score just above 0.6.

Compared to fill in the graph space, Figure 10 shows the movie-level results by individual movie for the question answering query. WHU_NERCMS run 3 achieved the highest scores on all movies, while both runs by NII_UIT team performed similarly in 3 out of the 5 movies with very small changes in the other 2 movies. In general, across all runs and movies, a median score of 0.3 was observed indicating that this query type may have been difficult given the open domain questions by humans.

Finally, Figures 12 and 13 show the scene-level results by movie in both query groups. Group 1 queries received results from only 1 team (WHU_NERCMS) and it can be shown that they performed relatively higher in 3 movies while struggled in 2 movies (little_rock and Memphis) with overall median score of 0.29. In group 2 queries, WHU_NERCMS run performed higher than NII_UIT across all movies and we can observe that overall this query group achieved higher scores than group 1 with a median of 0.36. The difference in performance between the two query groups could be due to the difficulty of recognizing fine grained interactions asked in group 1 compared to scene textual summary matching or sentiment classification.

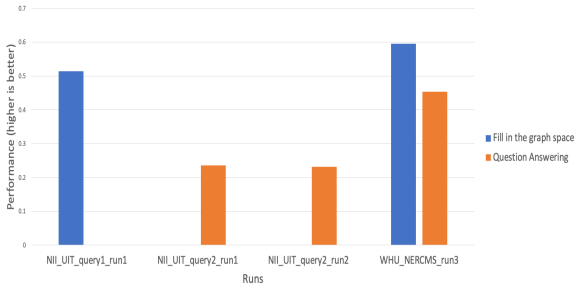

Figure 8: DVU: Overall run scores for movie-level in both query types

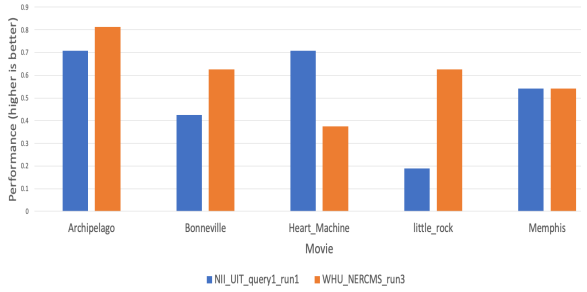

Figure 9: DVU: Movie-level fill in the graph query results by testing movie

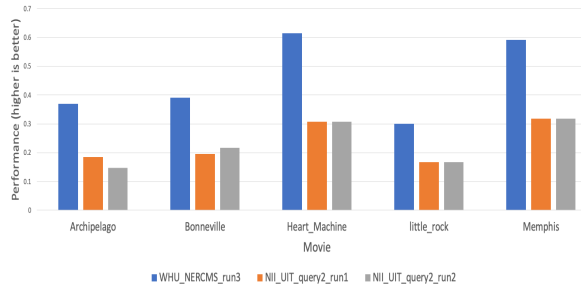

Figure 10: DVU: Movie-level question answering query results by testing movie

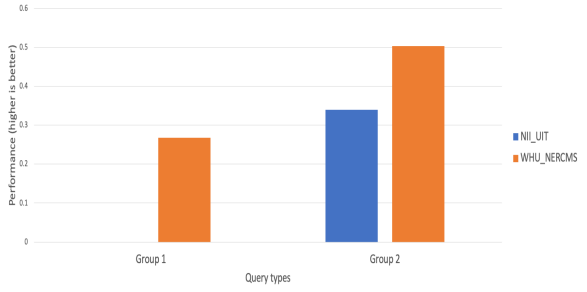

Figure 11: DVU: Overall team scores for scene-level in both query groups

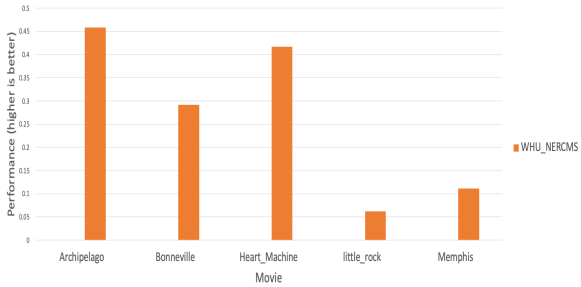

Figure 12: DVU: Scene-level group 1 query results by testing movie

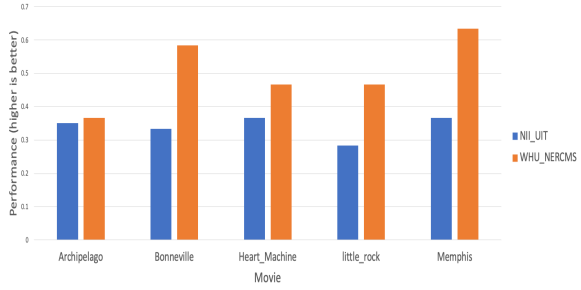

Figure 13: DVU: Scene-level group 2 query results by testing movie

Observations and conclusions

Task participation was low this year (2 out of 5 teams finished). In regard to movie-level queries, we found that fill in the graph space queries scored higher than question answering queries indicating QA queries are hard. On the other hand, for scene-level queries, group 2 (scene to text matching and sentiment classification) scored higher than group 1 (interactions focused). Comparing movie to scene queries in general, we found that movie-level results performed higher than scene-level results. While top system is consistently higher across most movies, the performance varies by movie and its complexity. Inspecting the teams' research papers we find that LLMs (Large Language Models) are being applied to answer DVU queries which is interesting direction given the multimodality aspect of long movies and how LLMs are becoming a major approach to solve a lot of natural language and vision problems. Given the low participation, the continuation of the task may not be feasible in 2024 unless the design of queries can be changed or a new derived task can be proposed and attract more participants.

3.3 Medical Video Question Answering

One of the key goals of artificial intelligence (AI) is the development of a multimodal system that facilitates communication with the visual world (image, video) using a natural language query. In recent years, significant progress has been achieved due to the introduction of large-scale language-vision datasets and the development of efficient deep neural techniques that bridge the gap between language and visual understanding. Improvements have been made in numerous vision-and-language tasks, such as visual captioning [Li et al., 2020, Luo et al., 2020], visual question answering [Zhang et al., 2021], and natural language video localization [Anne Hendricks et al., 2017]. Most of the existing works on language vision focused on creating datasets and developing solutions for open-domain applications. We believe medical videos may provide the best possible answers to many first aid, medical emergency, and medical education questions. With increasing interest in AI to support clinical decision-making and improve patient engagement [HHS, 2021], there is a need to explore such challenges and develop efficient algorithms for medical language-video understanding and generation. Towards this, we introduced new tasks to foster research toward designing systems that can understand medical videos to provide visual answers to natural language questions and are equipped with multimodal capability to generate instruction questions from the medical video. These tasks have the potential to support the development of sophisticated downstream applications that can benefit the public and medical practitioners.

System Task

- **Task A: Video Corpus Visual Answer Localization (VCVAL).** Given a medical query and a collection of videos, the task aims to retrieve the appropriate video from the video collection and then locate the temporal segments (start and end timestamps) in the video where the answer to the medical query is being shown or the explanation is illustrated in the video. The proposed VCVAL task can be considered as video retrieval and then finding a series of "*medical instructional activity-based frame localization*" where a potential solution first searches for all medical instructional activity for a given medical query and then localizes the activities in an untrimmed medical-instructional video. This task is the extension of the MVAL task introduced in MedVidQA-2022 [Gupta and Demner-Fushman, 2022], where we only focused on locating the segment from a given video. In contrast, the VCVAL task deals with relevant video retrieval followed by the visual answer segment localization (*cf.* Figure 14)[8]. The video retrieval system requires the ability to identify the medical instructional video and retrieve the most relevant video to the health-related query.
- **Task B: Medical Instructional Question Generation (MIQG).** Given a video segment and its subtitle, the task is to generate the instructional question for which the given video segment is the visual answer. This task comes under multimodal generation, where the system has to consider the video (visual) and subtitle (language) modality to generate the natural language question (*cf.* Figure 15)[9]. Given the data scarcity, annotation efforts, and necessity of expert involvement in the annotation to create ground-truth video question answering

[8]CC BY license
[9]CC BY license

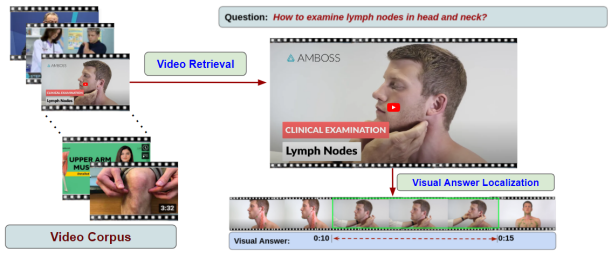

Figure 14: MedVidQA: Visualization of the proposed video corpus visual answer localization (VCVAL) task. The VCVAL task consists of two subtasks: video retrieval and visual answer localization.

(VidQA) dataset for the healthcare domain necessitate the effective and efficient MIQG system that can be used to generate additional VidQA datasets. The applications to the MIQG task are in creating an automatic human-computer dialogue system and developing intelligent tutor systems in a multimodal environment. A social or educational agent can be built that can generate appropriate and informative questions about a video or a collection of videos on a certain topic. Such generated questions can be used for promoting interactivity and persistence and test the knowledge of a student about a certain topic in a multimodal multi-turn dialogue setting.

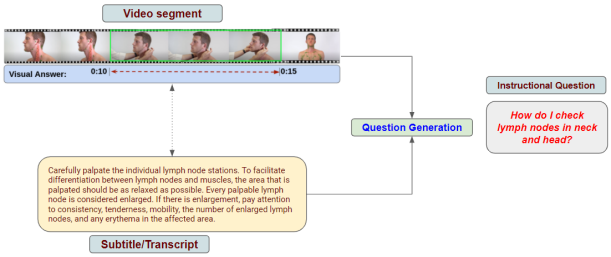

Figure 15: MedVidQA: Visualization of the proposed medical instructional question generation (MIQG) task.

Datasets

VCVAL: The VCVAL task comprises two subtasks: video retrieval and visual answer localization. For the video retrieval, we developed a video corpus considering the videos from '*Personal Care and Style*,' '*Health*,' and '*Sports and Fitness*' categories within the HowTo100M [Miech et al., 2019] dataset. We follow the strategy discussed in [Gupta et al., 2023b] to select the medical instructional videos from the HowTo100M dataset. This process yielded a total of 12,657 medical instructional videos, which we considered as video corpus to retrieve the relevant videos against the query. To facilitate the training and validation of the visual answer localization system, we use the MedVidQA collections [Gupta et al., 2023a] consisting of 3,010 human-annotated instructional questions and visual answers from 899 health-related videos.

We sampled a total of 60 videos from the video corpus and created forty (40) medical instructional questions. Out of these 40 questions, 20 questions were formulated according to the annotation guidelines discussed in [Gupta et al., 2023a]. We call this subset of the test questions 'Basic' questions. We aim to formulate another 20 questions that are more challenging compared to the 'Basic' questions. To create these questions, we asked the annotators to follow the following guidelines:

- Formulate such a question that cannot be answered with just the subtitles or captions available within the video (i.e., just listening to the video alone and not watching should not be enough to answer the question).
- The question should not be answered by reading the embodied text in the video.

The questions that can be answered with subtitles or embodied text in the video do not require visual information. Therefore, we formulated these questions, which required visual information to provide the answer, and called them "Visual Information Required" (VIR) questions.

MIQG: The MedVidQA dataset was also used for the MIQG task, as each sample in the dataset has the question and annotated start and end timestamps associated with it. To create a test collection, we sampled 100 videos from the video corpus discussed above, formulated 80 medical instructional questions, and marked the visual answer with answer start and end timestamps in the video. Similar to the VCVAL, we formulate two different test collections, 'Basic' and 'VIR', with 52 and 28 samples, respectively.

Judgment

The participants were asked to retrieve the relevant videos (up to 1000) for each question from the video corpus of having 12,657 videos. Additionally, the participants also had to provide the start and end

timestamps from each retrieved video against a given question, which can be considered a visual answer to the question. In order to judge the relevant videos and corresponding visual answers in the videos, we performed the manual judgments of all the submitted videos (943) and visual answers by the participants. We instructed a total of eight assessors with the following guidelines to assess the video:

Objective:

1. To judge relevant videos with respect to medical/healthcare instructional questions. A video can be called relevant if it has a visual answer to the question.
2. For each relevant video, provide the time stamps (start and end) where the answer is being shown or the explanation is illustrated in the video.

Evaluating videos for relevance : The videos are judged as being *"Definitely Relevant"*, *"Possibly Relevant"*, or *"Not Relevant"* to the given question. The assessors were presented with videos from the submitted runs. They were instructed to determine if the video was definitely relevant, possibly relevant, or not relevant to the question. In general, a video is definitely relevant if it contains a visual segment that can be considered a complete visual answer to the question. A video can be considered possibly relevant if it contains a visual segment that can be considered a partial/incomplete visual answer to the question. If the visual segments from the videos do not provide any visual answers to the question, the video can be marked as not relevant. The assessors were asked to provide the judgment with the following instructions:

- Only provide the time stamps for definitely relevant and possibly relevant videos.
- For each definitely relevant and possibly relevant video, provide the time stamps from the video that can be considered a visual answer.
- The time stamps should be the shortest span in the video, which can be considered as a complete (for definitely relevant video) or partially complete (for possibly relevant video) visual answer to the question.
- In case a video has multiple visual answers to the same question, assessors were asked to provide all the visual answers.

Metrics

Metrics for VCVAL Task The VCVAL task consists of two sub-tasks: video retrieval (VR) and visual answer localization (VAL). We evaluated the performance of the video retrieval system in terms of Mean Average Precision (MAP), Recall@k, Precision@k, and nDCG metrics with $k = \{5, 10\}$. We follow the `trec_eval`[10] evaluation library to report the performance of participating systems.

For the VAL task, if the predicted (retrieved by the system) video is from the list of relevant videos (marked by the assessor; we called it ground-truth video), then we compute the overlap between the retrieved and relevant video by the following metrics:

1. **Mean Intersection over Union (mIoU):** For a given question q_i, IoU is computed as the ratio of intersection area over union area between predicted and ground-truth temporal visual answer segments. It ranges from 0 to 1. A larger IoU means the predicted and ground-truth temporal visual answer segments match better, and IoU = 1.0 denotes an exact match. The mIoU is defined as the average temporal IoUs for all questions (N) in the test set. Formally,

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{i=N} \texttt{IoU}(q_i) \tag{1}$$

2. **IoU = μ** is another metric used to evaluate the performance of the VAL system. It denotes the percentage of questions for which, out of the top-n retrieved temporal segments, at least one predicted temporal segment having IoU with ground truth is larger than μ. Formally,

$$< R@n, IoU = \mu > = \frac{1}{N}\sum_{i=1}^{i=N} s(q_i, \mu), \text{ and} \tag{2}$$

$$s(q_i, \mu) = \begin{cases} 1, & \text{if } IoU(q_i) \geq \mu \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We evaluated the participants' submission by considering $\mu = \{0.3, 0.5, 0.7\}$ and for brevity, we denote the $< R@n, IoU = \mu >$ metric with IoU=μ $n = \{1, 3, 5, 10\}$

Metrics for MIQG Task For the task of MIQG, we followed the language generation evaluation metrics and evaluated the performance of the MIQG systems in terms of BLEU [Papineni et al., 2002], Rouge

[10] `https://github.com/usnistgov/trec_eval`

Team Name	Team Affiliations	VCVAL	MIQG
MI_TJU	Tianjin University	✓	✗
VPAI	Hunan University/CAS	✓	✓
UNCWAI	University of North Carolina Wilmington	✓	✗
UMBCVQA	University of Maryland Baltimore County	✗	✓
doshisha_uzl	Doshisha University and University of Lubeck	✗	✓

Table 10: MedVidQA: Participating teams and their task participation at MedVidQA@TRECVID 2023

[Lin, 2004] and BERTScore [Zhang et al., 2019] metrics.

Participating Teams

In total, 5 teams from Asia (China, Japan), Europe (Germany), and North America (USA) continents participated in the MedVidQA and submitted 5 and 8 individual runs for the VCVAL and MIQG tasks, respectively. We have provided (*cf.* Table 10) the team name, affiliations, and their participation in VCVAL and MIQG tasks.

Results

VCVAL Task: The VCVAL task consists of video retrieval and visual answer localization subtasks. We presented the results of the video retrieval subtask in Table 11. We reported the results in terms of MAP, R@5, R@10, P@5, P@10, and nDCG. Since the relevancy of the videos is judged in terms of multi-level judgment, we consider nDCG as the primary metric for video retrieval subtask. Team MI_TJU achieved the best nDCG for the video retrieval subtask with a score of 0.5448. We also compare the performance of the participating teams between Basic and VIR questions, which are shown in the results in Figure 16. We noticed that, out of the five runs submitted for the video retrieval subtask, three performed better on VIR questions than on Basic questions.

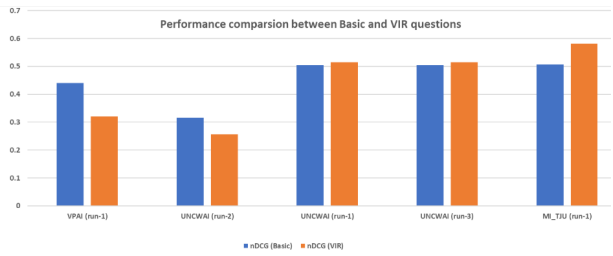

Figure 16: MedVidQA: Performance comparison of participating teams between Basic and VIR questions on video retrieval subtask of the VCVAL task.

The participating teams' visual answer localization subtask results are reported in Table 12. The table exhibits the detailed results with varying numbers of n and multiple evaluation metrics. We consider IoU=0.7 the primary metric for this subtask as it is the most strict metric, which signifies >= 70% overlap between the predicted and ground-truth visual answer segments. Team MI_TJU achieved the best IoU=0.7 for the visual answer localization subtask with a score of 50 ($n = 1$). Additionally, we examined the performance of participating teams on both Basic and VIR questions, as illustrated in Figure 17. We observed that, among the five runs submitted for the visual answer localization subtask, only two demonstrated better performance on VIR questions compared to Basic questions.

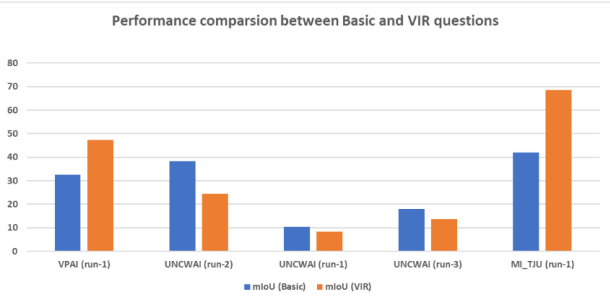

Figure 17: MedVidQA: Performance comparison of participating teams between Basic and VIR questions on visual answer localization ($n = 1$) subtask of the VCVAL task.

MIQG Task The results of medical instructional question generation by the participating teams are presented in Table 13. The table provides a detailed breakdown of the performance using various evaluation metrics. Following prior research [Du et al., 2017, Dong et al., 2019] on question generation, we adopted BLEU-4 as the primary metric for this task due to its strict measurement, indicating the extent of 4-gram overlap between the generated and ground-truth questions. Team doshisha_uzl achieved the highest BLEU-4 score for the MIQG task with a score of 0.05153. Similar to the previous task, we assessed the performance of participating teams on both Basic and VIR questions, as depicted in Figure 18. Among the eight runs submitted for the MIQG task, only one exhibited superior performance on VIR questions compared to Basic questions.

Team	RunID	MAP	R@5	R@10	P@5	P@10	nDCG
VPAI	run-1	0.2427	0.2489	0.2489	0.31	0.155	0.3804
UNCWAI	run-2	0.1839	0.1903	0.1903	0.29	0.145	0.2858
UNCWAI	run-1	0.3669	0.2221	0.3654	0.395	0.3575	0.5094
UNCWAI	run-3	0.3669	0.2221	0.3654	0.395	0.3575	0.5094
MI_TJU	run-1	0.404	0.3549	0.4132	0.545	0.3625	0.5448

Table 11: MedVidQA: Official results of the participating teams on video retrieval subtask of the VCVAL task.

n	Team	RunID	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
1	VPAI	run-1	57.5	35	25	39.97
	UNCWAI	run-2	42.5	32.5	22.5	31.37
	UNCWAI	run-1	10	7.5	0	9.32
	UNCWAI	run-3	25	10	5	15.78
	MI_TJU	run-1	67.5	62.5	50	55.24
3	VPAI	run-1	65.0	45.0	32.5	46.98
	UNCWAI	run-2	55.0	40.0	35.0	42.87
	UNCWAI	run-1	27.5	12.5	0.0	19.32
	UNCWAI	run-3	37.5	25.0	10.0	25.25
	MI_TJU	run-1	85.0	85.0	65.0	73.22
5	VPAI	run-1	65.0	45.0	32.5	46.98
	UNCWAI	run-2	55.0	40.0	35.0	42.87
	UNCWAI	run-1	37.5	15.0	0.0	24.34
	UNCWAI	run-3	37.5	27.5	12.5	29.58
	MI_TJU	run-1	87.5	87.5	75.0	77.29
10	VPAI	run-1	65.0	45.0	32.5	46.98
	UNCWAI	run-2	55.0	40.0	35.0	42.87
	UNCWAI	run-1	52.5	27.5	2.5	31.83
	UNCWAI	run-3	50.0	32.5	17.5	36.52
	MI_TJU	run-1	87.5	87.5	75.0	77.71

Table 12: MedVidQA: Official results of the participating teams on visual answer localization subtask of the VCVAL task.

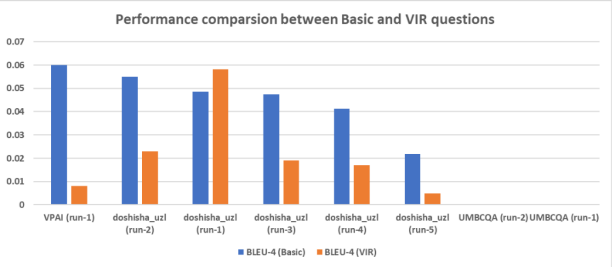

Figure 18: MedVidQA: Performance comparison of participating teams between Basic and VIR questions on the MIQG task.

Findings and Conclusion

Despite the expectation that VIR questions would be more challenging for the system, sometimes we observe submitted runs performed better on the Basic Questions compared to the VIR questions. We hypothesize that our approach to constructing these questions may explain this observation. VIR questions are generated by watching a specific video, and answering them requires visual information from that particular video. However, this requirement may not be applicable to other videos for the same question. The maximum nDCG of 0.5448 signifies the challenges of instructional video retrieval for the medical domain. The team MI_TJU achieved the best performance (55.24 mIoU) on the visual answer localization subtask with a multimodal approach. On the other hand, the team UNCWAI utilized only the textual modality for the VCVAL task and reported a performance of 31.37 (mIoU). The team doshisha_uzl achieved the best performance (0.05153 BLEU-4) on the MIQG task with a combination of mono and multimodal approaches. The lower performance on the MIQG task signifies the challenge of generating the instructional questions from the video segment. This underscores the need for a more sophisticated approach that involves a comprehensive multimodal understanding of the task.

3.4 Video to Text

Automatic annotation of videos using natural language text descriptions has been a long-standing goal of computer vision. The task involves understanding many concepts such as objects, actions, scenes, person-object relations, the temporal order of events throughout the video, to mention a few. In recent years there have been major advances in computer vision techniques that enabled researchers to start practical work on solving the challenges posed in automatic video captioning.

There are many use-case application scenarios that can greatly benefit from the technology, such as video summarization in the form of natural language, facilitating the searching and browsing of video archives

Team	RunID	BLEU	BLEU-4	ROUGE-2	ROUGE-L	BERTScore
doshisha_uzl	run-1	0.15828	0.05153	0.27845	0.47822	0.91092
doshisha_uzl	run-2	0.14352	0.04546	0.24372	0.44667	0.90523
VPAI	run-1	0.12969	0.04331	0.27329	0.47979	0.90981
doshisha_uzl	run-3	0.14593	0.03875	0.27379	0.47418	0.91099
doshisha_uzl	run-4	0.13289	0.03404	0.24227	0.45566	0.9078
doshisha_uzl	run-5	0.093	0.01627	0.20113	0.4085	0.90248
UMBCQA	run-2	0	0	0.12253	0.26042	0.85332
UMBCQA	run-1	0	0	0.1317	0.31554	0.87683

Table 13: MedVidQA: Official results of the participating teams on MIQG task.

	Number of runs
BUPT_MCPRL	4
Kslab	4
MLVC_HDU	1
RUC_AIM3	8
WasedaMeiseiSoftbank	8

Table 14: VTT: List of teams participating and their submitted runs

using such descriptions, describing videos as an assistive technology, etc. In addition, learning video interpretation and temporal relations among events in a video will likely contribute to other computer vision tasks, such as the prediction of future events from the video.

The Video to Text (VTT) task was introduced in TRECVID 2016. Since then, there have been substantial improvements in the dataset and evaluation. Essentially, each year's testing dataset is being appended to previous year's development dataset. In addition, since 2021, a subset of videos has been dedicated to a progress sub-task for which the ground truth is withheld and participants submit results from 2021 to 2023. They will then be able to compare their systems across the three years to measure improvement over the years on the same set of videos.

System Task

For each video, automatically generate a text description of 1 sentence independently from any previously generated sentences. Up to 4 runs are allowed per team. New this year is the introduction of a robustness sub-task where we added noise to the main task test data in both the audio and video channels.

For this year, 5 teams participated in the VTT task. The 5 teams submitted a total of 17 runs in the main task and 8 runs in the robustness task. A summary of participating teams is shown in Table 14.

Data

When the VTT task started the testing dataset consisted of Twitter Vine videos, which generally had a duration of 6 seconds. In 2019, we supplemented the dataset with videos from Flickr. During the years of 2020, 2021, and 2022 the VTT data was selected from the V3C1 and V3C2 data collection. The V3C dataset [Rossetto et al., 2019] is a large collection of videos from Vimeo. It also provides us with the advantage that we can distribute the videos rather than links, which may not be available in the future. This year, the testing dataset was selected from the V3C3 collection which is another subset of the bigger V3C dataset and shares all V3C1 and V3C2 characteristics.

For the purpose of this task, we only selected video segments with lengths between 3 and 15 seconds. A total of 2000 video segments were annotated manually by multiple annotators for this year's task. Since we have selected 300 videos for our progress set in 2021 and 2022, our results will be reported for 2000 new videos (non-progress) and the 300 videos in progress set.

It is important for a good dataset to have a diverse set of videos. We reviewed around 8000 videos and selected 2000 videos. Figure 19 shows a screenshot[11] of the video selection tool that was used to

[11]all videos are subset of V3C dataset and CC licensed

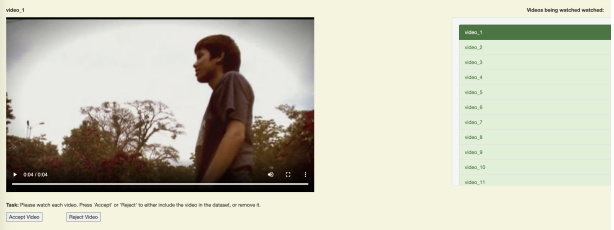

Figure 19: VTT: Screenshot of video selection tool.

decide whether a video was to be selected or not. We tried to ensure that the videos covered a large set of topics including spatial and temporal description aspects. If we came across a large number of videos that looked similar to previously selected clips, they were rejected. We also removed the following types of videos:

- Videos with multiple, unrelated segments that are hard to describe, even for humans.
- Any animated videos.
- Other videos that may be considered inappropriate or offensive.

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

Table 15: VTT: Average number of words per sentence for all the annotators. The table also shows the number of videos watched by each annotator.

Annotation Process The videos were divided among 5 annotators, with each video being annotated once by each to create 5 annotations per video.

The annotators were asked to include and combine into 1 sentence, if appropriate and available, four facets of the video they are describing:

- **Who** is the video showing (e.g., concrete objects and beings, kinds of persons, animals, or things)?
- **What** are the objects and beings doing (generic actions, conditions/state or events)?
- **Where** was the video taken (e.g., locale, site, place, geographic location, architectural)?
- **When** was the video taken (e.g., time of day, season)?

Different annotators provide varying amounts of detail when describing videos. Some people try to incorporate as much information as possible about the video, whereas others may write more compact sentences. Table 15 shows the average number of words per sentence for each of the annotators. The average sentence length varies from 20 words to 29 words, emphasizing the difference in descriptions provided by the annotators. The overall average sentence length for the dataset is 24.56 words.

Furthermore, the annotators were also asked the following questions for each video:

- Please rate how difficult it was to describe the video.
  1. Very Easy
  2. Easy
  3. Medium
  4. Hard
  5. Very Hard
- How likely is it that other assessors will write similar descriptions for the video?
  1. Not Likely
  2. Somewhat Likely
  3. Very Likely

The average score for the first question was 2.22 (on a scale of 1 to 5), showing that the annotators thought the videos were close to medium level of difficulty on average. The average score for the second question was 2.52 (on a scale of 1 to 3), meaning that they thought that other people would write a similar description as them for most videos. The two scores are negatively correlated as annotators are more likely to think that other people will come up with similar descriptions for easier videos. The Pearson correlation coefficient between the two questions is -0.53.

Submissions

Systems were required to specify the run types based on the types of training data and features used.

The list of training data types is as follows:

- 'I': Training using image captioning datasets only.

- 'V': Training using video captioning datasets only.
- 'B': Training using both image and video captioning datasets.

The feature types can be one of the following:

- 'V': Only visual features are used.
- 'A': Both audio and visual features are used.

In total, 25 runs were submitted and distributed as follows: 16 runs were of type "BV" (visual features from both image and video datasets), 2 runs of type "IV" (used image datasets with visual features), 2 runs of type "IA" (image data with audio features from video dataset), and 5 runs are of type "VV" (video datasets with visual only features).

Teams were also asked to specify the loss function used for their runs. Loss functions reported were mainly based on cross-entropy (8 runs). Four runs reported language-based loss function, while four other runs applied self-critical reinforcement learning loss.

Evaluation and Metrics

The description generation task scoring was done automatically using different popular metrics. We also used a human evaluation metric on selected runs to compare with the automatic metrics.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee and Lavie, 2005] and BLEU (BiLingual Evaluation Understudy) [Papineni et al., 2002] are standard metrics in machine translation (MT). BLEU was one of the first metrics to achieve a high correlation with human judgments of quality. It is known to perform poorly if it is used to evaluate the quality of individual sentence variations rather than sentence variations at a corpus level. In the VTT task the videos are independent and there is no corpus to work from. Thus, our expectations are lowered when it comes to evaluation by BLEU. METEOR is based on the harmonic mean of unigram or n-gram precision and recall in terms of overlap between two input sentences. It redresses some of the shortfalls of BLEU such as better matching synonyms and stemming, though the two measures seem to be used together in evaluating MT.

The CIDEr (Consensus-based Image Description Evaluation) metric [Vedantam et al., 2015] is borrowed from image captioning. It computes TF-IDF (term frequency inverse document frequency) for each n-gram to give a sentence similarity score. The CIDEr metric has been reported to show high agreement with consensus as assessed by humans. We also report scores using CIDEr-D, which is a modification of CIDEr to prevent "gaming the system".

The SPICE (Semantic Propositional Image Caption Evaluation) metric [Anderson et al., 2016] is another metric that has gained popularity in image captioning evaluation. The metric uses scene graph similarity between generated captions and the ground truth instead of n-grams.

The STS (Semantic Textual Similarity) metric [Han et al., 2013] was also applied to the results, as in the previous years of this task. This metric measures how semantically similar the submitted description is to one of the ground truth descriptions.

In addition to automatic metrics, the description generation task includes human evaluation of the quality of automatically generated captions. Recent developments in Machine Translation evaluation have seen the emergence of DA (Direct Assessment), a method shown to produce highly reliable human evaluation results for MT and Natural Language Generation [Graham et al., 2016, Mille et al., 2020]. DA now constitutes the official method of ranking in main MT benchmark evaluations [Bojar et al., 2017, Barrault et al., 2020].

With respect to DA for evaluation of video captions (as opposed to MT output), human assessors are presented with a video and a single caption. After watching the video, assessors rate how well the caption describes what took place in the video on a 0–100 rating scale [Graham et al., 2018]. Large numbers of ratings are collected for captions before ratings are combined into an overall average system rating (ranging from 0 to 100 %). Human assessors are recruited via Amazon's Mechanical Turk (AMT), with quality control measures applied to filter out or downgrade the weightings from workers unable to demonstrate the ability to rate good captions higher than lower quality captions. This is achieved by deliberately "polluting" some of the manual (and correct) captions with linguistic substitutions to generate captions whose semantics are questionable. For instance, we might substitute a noun for another noun and turn the manual caption "A man and a woman are dancing on a table" into "A *horse* and a woman are dancing on a table", where "horse" has been substituted for "man". We expect such automatically-polluted captions to be rated poorly and when an AMT worker correctly does this, the ratings for that worker are improved.

DA was first used as an evaluation metric in TRECVID 2017. This metric has been used every year since then to rate each team's primary run.

Results

The metric score for each run is calculated as the average of the metric scores for all the descriptions within that run. Table 16 shows the top performance per team across all automatic metrics. The STS metric allows the comparison between two sentences. For this reason, the captions are compared to a single ground truth description at a time, resulting in 5 STS scores. We report the average of these scores as the STS score. It can be shown the two teams (RUC_AIM3 and WasedaMeiseiSoftbank) performed the highest in all metrics, followed by BUPT_MCPRL and then the two teams Kslab and MLVC_HDU performed lowest. Table 17 ,on the other hand, shows the results for the two teams that participated in the robustness sub-task (introducing noise to the testing dataset). We can see that both teams' performance in general did not get affected to a big extent. Specifically, RUC_AIM3 team surprisingly improved slightly in 3 metrics and worsened in the other 3 metrics (CIDER-D, SPICE and STS). The WasedaMeiseiSoftbank performed slightly lower in all metrics. The metric that reported the lowest performances in both teams was the STS metric (about a 50% decrease).

Table 18 shows the correlation between the different metric scores for all the runs. The metrics correlate very well, which shows that they agree on the overall scoring of the runs. The correlation scores ranged between 0.876 and 0.989.

Teams were asked to provide a confidence score for each generated sentence. Figure 20 shows the submitted average confidence scores for each run against each metric score. There seems to be some correlation (not very strong) between confidence and metric scores. It can be shown that a few runs (by RUC_AIM3) achieved very high CIDER scores at mid level confidence.

Figure 21 shows the average DA score per system after it is standardized per individual AMT worker's mean and standard deviation score. The DA raw scores are micro-averaged per caption, and then averaged over all videos. The DA experiment was conducted on only 1 primary run per team that they selected when submitted their runs. The HUMAN systems represent manual captions provided by assessors. As expected, captions written by assessors outperform the automatic systems. They are followed by two systems (RUC_AIM3 and BUPT_MCPRL) outperforming the other 3 systems based on the DA experiments. To check how significant are system performance in comparison to each other and human captions, a significance testing indicated that the top 4 human systems as well as the 2 top automatic systems (RUC_AIM3 and BUPT_MCPRL) were not significantly better than each other while they are all better than the other 3 systems (WasedaMeiseisoftbank, Kslab, and MLVC_HDU). In addition, WasedaMeiseisoftbank and Kslab is significantly better than MLVC_HDU.

Table 19 shows the correlation between different overall metric scores for the primary runs of all teams and the 'DA_Z' metric score (DA_Z is the standardized score per individual worker's mean and standard deviation score) generated by humans. The score correlates positively with all metrics. The correlation ranged between 0.81 to 0.98 with CIDER and STS achieving the highest correlation with DA.

Table 20 shows the automatic metrics scores for the progress sub-task which evaluated runs on 300 fixed videos between 2021 and 2023. The table shows only teams who submitted in at least two years. It can be shown that most teams performed in 2023 better than 2022 or 2021 with one exception for team MLV_HDU were they performed consistently in 2022 better than 2023. Finally, the DA experiment was also conducted on the progress sub-task videos for the primary runs submitted in the three years. Figure 22 shows the results where it can be seen that the top 3 teams (RUC_AIM, BUPT_MCPRL, and WasedaMeiseSoftbank) 2023 results are better than other 2022 and 2021 progress results based on the human evaluation in DA metric.

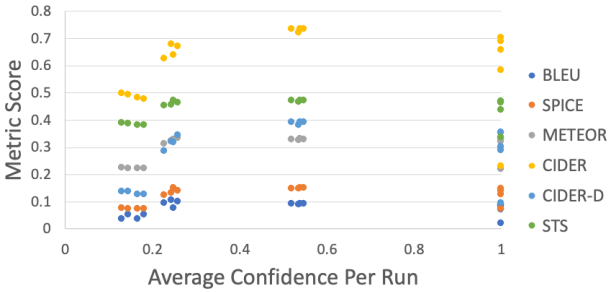

Figure 20: VTT: system reported sentence confidence scores against the various metric scores.

	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
Kslab	0.054	0.227	0.501	0.14	0.078	0.39
MLVC_HDU	0.023	0.220	0.233	0.096	0.077	0.339
RUC_AIM3	0.094	0.332	0.738	0.394	0.152	0.474
BUPT_MCPR	0.091	0.332	0.706	0.357	0.150	0.471
WasedaMeiseiSoftbank	0.108	0.335	0.682	0.348	0.152	0.475

Table 16: VTT: Top score by each team for all automatic metrics.

	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
RUC_AIM3	0.099	0.332	0.739	0.390	0.151	0.206
WasedaMeiseiSoftbank	0.105	0.331	0.677	0.340	0.151	0.224

Table 17: VTT: Top score by each team participated in the robustness sub-task for all automatic metrics.

	CIDER	CIDER-D	SPICE	METEOR	BLEU	STS
CIDER	1.000	0.931	0.877	0.886	0.912	0.959
CIDER-D	0.931	1.000	0.971	0.966	0.916	0.959
SPICE	0.877	0.971	1.000	0.989	0.876	0.963
METEOR	0.886	0.966	0.989	1.000	0.923	0.971
BLEU	0.912	0.916	0.876	0.923	1.000	0.925
STS	0.959	0.959	0.963	0.971	0.925	1.000

Table 18: VTT: Correlation between overall run scores for automatic metrics.

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

Table 19: VTT: Correlation between DA and automatics metrics for the primary runs only

	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
RUC_AIM3 (2021)	0.042	0.335	0.651	0.387	0.128	0.454
RUC_AIM3 (2022)	**0.113**	0.384	0.85	0.545	0.173	**0.488**
RUC_AIM3 (2023)	0.094	**0.397**	**0.906**	**0.552**	**0.181**	0.474
WasedaMeiseiSoftbank (2022)	0.036	0.271	0.417	0.216	0.09	0.378
WasedaMeiseiSoftbank (2023)	**0.108**	**0.398**	**0.82**	**0.499**	**0.178**	**0.475**
Kslab (2021)	0.005	0.204	0.163	0.07	0.047	0.26
Kslab (2022)	**0.085**	**0.295**	0.607	0.261	0.099	**0.40**
Kslab (2023)	0.054	0.278	**0.62**	**0.267**	**0.1**	0.39
MLVC_HDU (2022)	**0.071**	**0.283**	**0.364**	**0.201**	**0.1**	**0.367**
MLVC_HDU (2023)	0.023	0.272	0.32	0.189	0.096	0.339

Table 20: VTT: Top score by each team participated in the progress sub-task (2021 to 2023) for all automatic metrics.

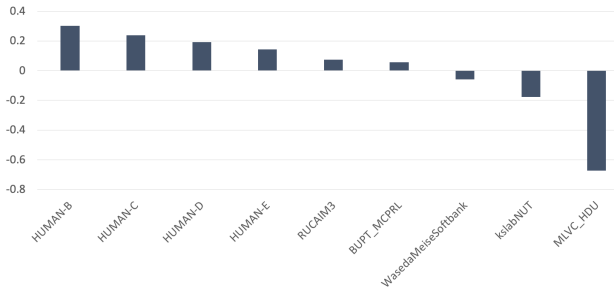

Figure 21: VTT: Average DA score per system after standardization per individual worker's mean and standard deviation score.

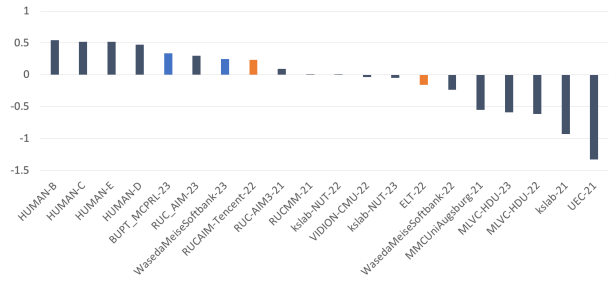

Figure 22: VTT: Average DA score per system, for progress task, after standardization per individual worker's mean and standard deviation score.

Task observations and conclusions

The VTT task continues to have stable participation. Given the challenging nature of the task, and the increasing interest in video captioning in the computer vision community, we hope the dataset resources generated from the task as well as algorithms by teams inspire more improvements for the task in the future.

This was the first year using the V3C3 test data as well as the first year to introduce a robustness subtask. The robustness sub-task setup will need to be updated to incorporate more real world harder transformations such as change in lighting, camera shaking, etc. This year's robustness results proved that systems were able to cope with the introduced noise and performance did not change significantly. The progress subtask concludes that this year's systems are better than the previous two years in most cases. High correlation exists between all automatic metrics. Audio features were used by only two runs. Based on the DA evaluation, human captions are still better than the best automatic system but automatic systems are generally getting better and closer to human captions. With increasing interest in video captioning, participants have many options of open datasets available to train their systems.

For detailed information about the approaches and results for individual teams' performance and runs, we refer the reader to the site reports [TV23Pubs, 2023] in the online workshop notebook proceedings.

3.5 Activities in Extended Video

The Activities in Extended Video (ActEV) evaluation series is designed to accelerate the development of robust, multi-camera, automatic human activity detection systems for forensic and real-time alerting applications. In this evaluation, an activity is defined as "one or more people performing a specified movement or interacting with an object or group of objects (including driving)", while an instance indicates an occurrence (time span of the start and end frames) associated with the activity. This year's TRECVID'23 ActEV Self-Reported Leaderboard (SRL) Challenge is based on the Multi-view Extended Video with Activities (MEVA) Known Facility (KF) dataset [Kitware, 2020]. The large-scale MEVA dataset is designed for activity detection in multi-camera environments. The same MEVA dataset was used for TRECVID'22 ActEV SRL evaluation. The ActEV task evaluations in 2021 and 2020

used the VIRAT dataset which had 35 target activities [Oh et al., 2011]. The NIST TRECVID ActEV series was initiated in 2018 to support the Intelligence Advanced Research Projects Activity (IARPA) Deep Intermodal Video Analytics (DIVA) Program.

The TRECVID 2018 ActEV (ActEV18) evaluated system detection performance on 12 activities for the self-reported evaluation and 19 activities for the leaderboard evaluation using the VIRAT V1 and V2 datasets [Lee et al., 2018]. For the self-reported evaluation, the participants ran their software on their hardware and configurations and submitted the system outputs with the defined format to the NIST scoring server.

The ActEV18 evaluation addressed two different tasks: 1) identify a target activity along with the time span of the activity (AD: activity detection), 2) detect objects associated with the activity occurrence (AOD: activity and object detection).

For the TRECVID 2019 ActEV (ActEV19) evaluation, we primarily focused on 18 activities and increased the number of instances for each activity. ActEV19 included the test set from both VIRAT V1 and V2 datasets and the systems were evaluated on the activity detection (AD) task only.

The TRECVID 2020 ActEV (ActEV20) SRL is based on the VIRAT V1 and V2 datasets with 35 activities with updated names to make it easier to use the MEVA dataset to train systems for TRECVID ActEV leaderboard. The TRECVID 2021 ActEV (ActEV21) was based on the same 35 activities as ActEV20 and on the VIRAT V1 and V2 datasets and systems are evaluated on the activity detection (AD) task only.

Figure 23 illustrates an example of representative activities that were used in the TRECVID 2023 ActEV SRL based on the MEVA dataset.

All these evaluations are primarily targeted for forensic analysis applications that process an entire corpus prior to returning a list of detected activity instances.

In this section, we first discuss the task and datasets used and introduce the metrics to evaluate algorithm performance. In addition, we present the results for the TRECVID'23 ActEV SRL submissions and discuss observations and conclusions.

Figure 23: Example of activities for MEVA dataset used ActEV SRL evaluation. IRB (Institutional Review Board): ITL-00000755

Task and Dataset

In the TRECVID'23 ActEV SRL evaluation, there are two tasks for systems; the primary task is Activity and Object Detection (AOD) and the secondary task is Activity Detection (AD)

Task1 (AOD): for the AOD task, given the predefined activity classes, the objective is to automatically detect the presence of the target activity, spatiotemporally localize all instances of the activity, and provide a confidence score indicating the strength of evidence that the activity is present. This task requires spatiotemporal localization of objects involved in the activity (as one bounding box per frame that encompasses people, vehicles, and other objects). For a system-identified activity instance to be evaluated as correct, the activity class must be correct and the spatiotemporal overlap must fall within a minimal requirement. The evaluation tool, ActEV_Scorer, transforms the localization bounding boxes of both the system and reference files on the fly so that developers have the flexibility to spatially localize individual objects or a single encompassing box.

Task2 (AD): for the AD task, given the predefined activity classes, the objective is to automatically detect the presence of the target activity, temporally localize all instances, and provide a presence confidence score indicating the strength of evidence that the activity is present. This task does not require spatiotemporal localization of objects. For a system-identified activity instance to be evaluated as correct, the activity class must be correct and the temporal overlap must fall within a minimal requirement.

The ActEV SRL evaluation is based on the Known Facilities (KF) data from the Multiview Extended Video with Activities (MEVA) dataset. The KF data was collected at the Muscatatuck Urban Training

Table 21: A list of activity names for TRECVID ActEV SRL evaluation, there were 20 activities based on the MEVA dataset.

person_closes_vehicle_door	person_closes_vehicle_door
person_enters_scene_through_structure	person_enters_scene_through_structure
person_enters_vehicle	person_enters_vehicle
person_exits_scene_through_structure	person_exits_scene_through_structure
person_exits_vehicle	person_exits_vehicle
person_interacts_with_laptop	person_interacts_with_laptop
person_opens_facility_door	person_opens_facility_door
person_opens_vehicle_door	person_opens_vehicle_door
person_picks_up_object	person_picks_up_object
person_puts_down_object	person_puts_down_object

Center (MUTC) with a team of over 100 actors performing in various scenarios. The KF dataset has two parts: (1) the public training and development data and (2) SRL test dataset.

For this evaluation, we used 20 activities from the MEVA dataset and the activities were annotated by Kitware, Inc. The CVPR'22 ActivityNet ActEV SRL test dataset is a 16-hour collection of videos that only consists of Electro-Optics (EO) camera modalities from public cameras. The ActEV SRL test dataset is the same as the one used for WACV'22 HADCV workshop ActEV SRL challenge and for the CVPR ActivityNet 2022 ActEV SRL challenge. The detailed definition of each activity and evaluation requirements are described in the evaluation plan [ActEV23, 2023].

Table 21 lists the 20 activity names for TRECVID ActEV SRL evaluation, based on the MEVA dataset.

Performance Measures

ActEV is not a discrete detection task unlike speaker recognition [Greenberg et al., 2020] and fingerprint identification [Karu and Jain, 1996], it is a streaming detection task where multiple activity instances can overlap temporally or spatially and is similar to keyword spotting in audio [Le et al., 2014]. From a metrology perspective, the difference between discrete and streaming detection tasks is that non-target trials (i.e., test probes not belonging to the class) are not countable for streaming detection because the number of unique temporal/spatial instances is practically infinite. To account for this difference, the ActEV evaluations used two methods to normalize the measured false alarm performance. The first, "Rate of False Alarms" (R_{fa}), is an instance-based false alarm measure that uses the number of video minutes as an estimate of the number of non-target trials as the false alarm denominator. The second, "Time-based False Alarms" (T_{fa}), is a time-based false alarm measure that uses the sum of non-target time as the denominator. The two variations correspond to two views concerning the impact false alarms have on a user reviewing detections. The former is instance-based which implies the user effort would scale linearly with the detected instances and the latter is time-based which implies the user effort would scale linearly with the duration of video reviewed.

For both the AOD (primary) and AD (secondary) tasks for TRECVID'23 ActEV SRL, the submitted results are measured by Probability of Missed Detection (Pmiss) at a Rate of Fixed False Alarm (R_{fa}) of 0.1 (denoted Pmiss@0.1RFA). RateFA is the average number of false alarm activity instances per minute. Pmiss is the portion of activity instances where the system did not detect the activity within the required temporal (AD) and spatio-temporal (AOD) overlap requirements. Submitted results are scored for Pmiss and RateFA at multiple thresholds (based on confidence scores produced by the systems), creating a detection error tradeoff (DET) curve.

The primary measure of performance for TRECVID ActEV21 was the normalized, partial Area Under the DET Curve ($nAUDC$) from 0 to a fixed value a , denoted $nAUDC_a$, representing a Rate of False Alarms (R_{fa}) nAUDC R_{FA} which is a different metric than used for the TRECVID ActEV20 and ActEV19 evaluations which used T_{fa}. The switch to R_{fa} coincided with a new experimental finding. T_{fa}-optimized systems tend to hyper-segment detections to maximize performance on the metrics. When evaluators reviewed the detections of top systems, the number of detections

to review overwhelmed the reviewer. Consequently, changing the primary metric to use R_{fa} greatly penalized hyper fragmentation and produced systems with fewer, higher quality detections. All ActEV performance measurements were on a per-activity basis and then performance was aggregated by averaging over activities. While presence confidence scores were used to compute performance, cross-activity presence confidence score normalization was not required nor evaluated.

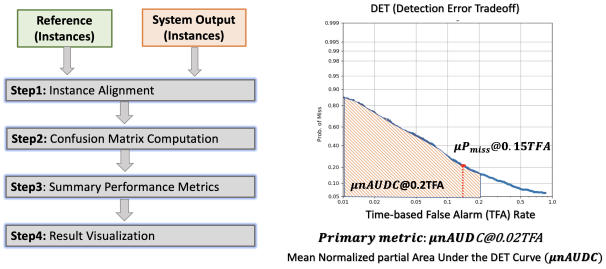

Figure 24: Performance measure calculation and Detection Error Tradeoff (DET) curves

Figure 24 shows a summary of performance metric calculation. For given reference annotation and system output, the steps are 1) Align the reference activity instance with each relevant system's instance; 2)Compute detection confusion matrix; 3)Compute summary performance metrics; and 4) Visualize the results such as DET curve shown here, which the x-axis is the Time-based False Alarm (TFA) Rate and y-axis is the probability of missed detection. For both the AOD (primary) and AD tasks, the submitted results are measured by the Probability of Missed Detection (Pmiss) at a Rate of Fixed False Alarm (RateFA) of 0.1 (Pmiss@0.1RFA). RateFA is the average number of false alarm activity instances per minute. Pmiss is the portion of activity instances where the system did not detect the activity within the required temporal (AD) and spatio-temporal (AOD) overlap requirements. For TRECVID'23 ActEV SRL evaluation primary metric was the AOD mean Normalized partial Area Under the DET Curve $nAUDC$.

As shown in Figure 25, the detection confusion matrix is calculated with an alignment between reference and system output instances per target activity; Correct Detection (CD) indicates that the reference and system output instances are correctly mapped (instances marked in blue). Missed Detection (MD) indicates that an instance in the reference has no correspondence in the system output (instances marked in yellow) while False Alarm (FA) indicates that an instance in the system output has no correspondence in the reference (instances marked in red). After calculating the confusion matrix, we summarize system performance: for each instance, a system output provides a confidence score that indicates how likely the instance is associated with the target activity. The confidence scores are not used as a decision threshold. Rather, a decision threshold is applied to the scores to determine the error counts (N_{FA} and N_{miss}).

In the ActEV22 evaluation, a probability of missed detections (P_{miss}) and a rate of false alarms (R_{FA}) were used and computed at a given decision threshold:

$$P_{\text{miss}}(\tau) = \frac{N_{\text{MD}}(\tau)}{N_{\text{TrueInstance}}}$$

$$\text{R}_{\text{FA}}(\tau) = \frac{N_{\text{FA}}(\tau)}{\text{VideoDurInMinutes}}$$

where $N_{\text{MD}}(\tau)$ is the number of missed detections at the threshold τ, $N_{\text{FA}}(\tau)$ is the number of false alarms, and *VideoDurInMinutes* is the video duration in minutes. $N_{\text{TrueInstance}}$ is the number of reference instances annotated in the sequence per activity. Lastly, the Detection Error Tradeoff (DET) curve [Martin et al., 1997] is used to visualize system performance.

To understand system performance better and to be more relevant to the human review use case, we used the normalized, partial area under the DET curve ($nAUDC$) from 0 to a fixed (R_{fa}) to evaluate algorithm performance. The partial area under DET curve is computed separately for each activity over all videos in the test collection and then is normalized to the range [0, 1] by dividing by the maximum partial area. $nAUDC_a = 0$ represents a perfect score. The $nAUDC_a$ is defined as:

$$nAUDC_a = \frac{1}{a}\int_{x=0}^{a} P_{miss}(x)dx, x = R_{fa}$$

where x is integrated over the set of R_{fa} and P_{miss} as defined above.

In the AOD task, a system detects the target activity, temporally localizes it, and also spatio-temporally localizes the objects that are associated with a given activity by providing the coordinates of object bounding boxes and object presence confidence scores.

The primary metric is similar to AD, however, the instance alignment step uses an additional alignment term for object detection congruence to opti-

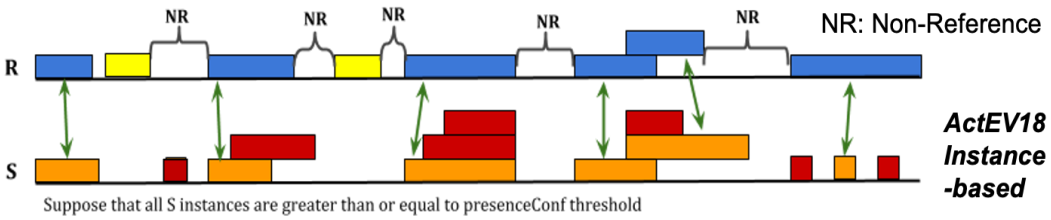

Figure 25: Illustration of activity instance alignment. R is the set of reference instances and S is the set of the system instances. Green arrows connect R and S instances that are determined to be aligned and thus labeled correct detections.

mally map reference and system output instances—this is covered in further detail in the evaluation plan [ActEV23, 2023].

For the object detection (secondary) metric, we employed the Normalized Multiple Object Detection Error (N_MODE) described in [Kasturi et al., 2009] and [Bernardin and Stiefelhagen, 2008]. N_MODE evaluates the relative number of false alarms and missed detections for all objects per activity instance. Note that the metric is applied only to the frames where the system overlaps with the reference. The metric also uses the Hungarian algorithm to align objects between the reference and system output at the frame level. The confusion matrix for each frame t is calculated from the confidence scores of the objects' bounding boxes, referred to as the object presence confidence threshold τ. $\mathrm{CD}_t(\tau)$ is the count of reference and system output object bounding boxes that are correctly mapped for frame t at threshold τ. $\mathrm{MD}_t(\tau)$ is the count of reference bounding boxes not mapped to a system object bounding box at threshold τ. $\mathrm{FA}_t(\tau)$ is the count of system bounding boxes that are not aligned to reference bounding boxes. The equation for N_MODE follows:

$$N_{\mathrm{MODE}(\tau)} = \sum_{t=1}^{N_{\mathrm{frames}}} \frac{(C_{\mathrm{MD}} \times MD_t(\tau) + C_{\mathrm{FA}} \times FA_t(\tau))}{\sum_{t=1}^{N_{\mathrm{frames}}} N_R^t}$$

N_{frames} is the number of frames in the sequence for the reference instance and N_R^t is the number of reference objects in frame t. For each instance-pair, the minimum N_MODE value (minMODE) is calculated for object detection performance and P_{Miss} at R_{FA} points are reported for both activity-level and object-level detections. For the activity-level detection, we used the same operating points P_{miss} at $R_{\mathrm{FA}} = 0.1$ and P_{miss} at $R_{\mathrm{FA}} = .2$ while P_{miss} at $R_{\mathrm{FA}} = 0.1$ was used for the object-level detection. We used 1- minMODE for the object detection congruence term to align the instances for the target activity detection. In this evaluation, the spatial object localization (that is, how precisely systems can localize the objects) is not addressed.

ActEV Results

A total of six teams from academia and industry from 3 countries participated in the ActEV23 evaluation. Each participant was allowed to submit multiple system outputs and a total of 38 submissions were received. Table 22 lists the participating teams along with results ordered by $mean\_P_{miss}@.1RFA$ values scores for the top performing system per team along with $nAUDC@0.2RFA$ values. The top $mean\_P_{miss}@.1RFA$ performance on activity detection is by BUPT-MCPRL at 57.81% followed by Mlvc_hdu at 89.52% and hsmw is third at 98.41%.

Figure 26 shows the performance based on the Activity and Object Detection (AOD) DET Curve for the 5 teams. The x-axis is the Rate of False Alarms, the y-axis is the Probability of Missed Detection and a smaller value is considered better performance. We observed that the new low for $mean\_P_{miss}@.1RFA$ of 57.8% for team BUPT-MCPRL, states a relative reduction of 8.4% from the previous year.

Figure 27 shows the AOD performance for all individual activities for all the teams. The x-axis shows the 20 activities and the y-axis shows the $mean\_P_{miss}@.1RFA$. The vehicles activities remain easier than people only activities and people and object interaction activities.

Figure 28 shows the AD vs. AOD Detection Performance for the six teams for all the activities. The x-axis shows the scores for AD and AOD tasks and the y-axis shows the $mean\_P_{miss}@.1RFA$. As expected for every team, their AOD system has higher $mean\_P_{miss}@.1RFA$ rates than AD.

To examine the localization performance for correct AOD instances, Figure 29 shows the localization performance varies across the 6 teams that participated in AOD evaluations. The x-axis shows the 20 activities and the y-axis shows the localization performance $nMODE@0.1RFA$. The missing points in the graph indicate no correct AOD detections. The BUPT-MCPRL team localizes well for most of the activities.

Table 22: Summary of participants' information and results ordered by AOD, $\mu nAUDC$ values. The AOD values of $mean\_P_{miss}$@.1RFA values along with the $nMODE$@.1RFA are also presented. We also present the AD values of $nAUDC$@.2RFA and $mean\_P_{miss}$@.1RFA. Each team was allowed to have multiple submissions.

Team	Organization	Primary Task: Activity and Object Detection (AOD)		Secondary Task: Activity Detection (AD)	
		Pmiss @0.1RFA	nMODE @0.1RFA	Pmiss @0.1RFA	nAUDC @0.2RFA
BUPT-MCPRL	Beijing University of Posts and Telecommunications, China	0.5781	0.0206	0.5145	0.5611
mlvc_hdu	Hangzhou Dianzi University	0.8952	0.3167	0.8746	0.885
HSMW (late)	University of Applied Sciences	0.9841	0.1349	0.9641	0.9669
Waseda_Meisei _Softbank	Waseda University, Meisei University, SoftBank Corporation	0.9985	0.0614	0.9940	0.9948
FDU_AWS	Fudan University, Amazon Web Service	0.9999	0.0	0.9999	0.9916
QWER	Fudan University, Amazon Web Service			1.0	1.0

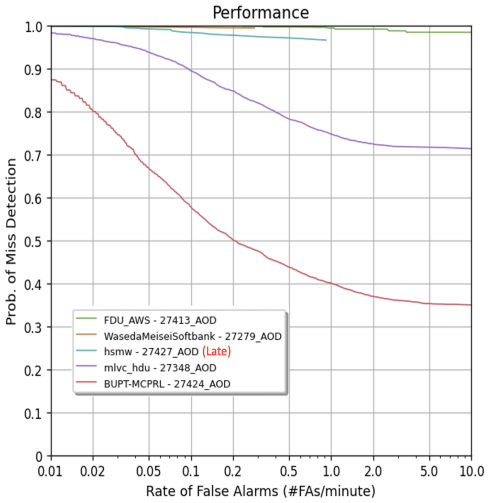

Figure 26: Activity and Object Detection (AOD) DET Curve for the six teams.

Summary

In this section, we presented the TRECVID'23 ActEV SRL evaluation task, the performance metric and results for human activity detection for both the Activity and Object Detection and the Activity Detection tasks. We primarily focused on the activity detection task only and the time-based false alarms were used to have a better understanding of the system's behavior and to be more relevant to the use cases. The TRECVID'23 ActEV evaluation was based on the MEVA [Kitware, 2020] dataset and had 20 target activities in total. This was the fourth time the MEVA dataset has been used for a ActEV evaluation. Six teams from 3 countries participated in the ActEV SRL evaluation and made a total of 118 submissions. We observed that, given the datasets and systems, the vehicles activities remain easier than people and people and object interaction activities. The teams MLVC_hdu and WadsedaMeiselSoftbank participated for the first time in the ActEV evaluation. The BUPT team had the top performing system followed by the mlvc_hdu team. The BUPT AOD performance improved 8.4% relative to last year. The Detection and Localization (AOD) still remains a more difficult task for the teams.

The TRECVID'23 ActEV SRL evaluation provided researchers an opportunity to evaluate their activity detection algorithms on a self-reported leaderboard. We hope

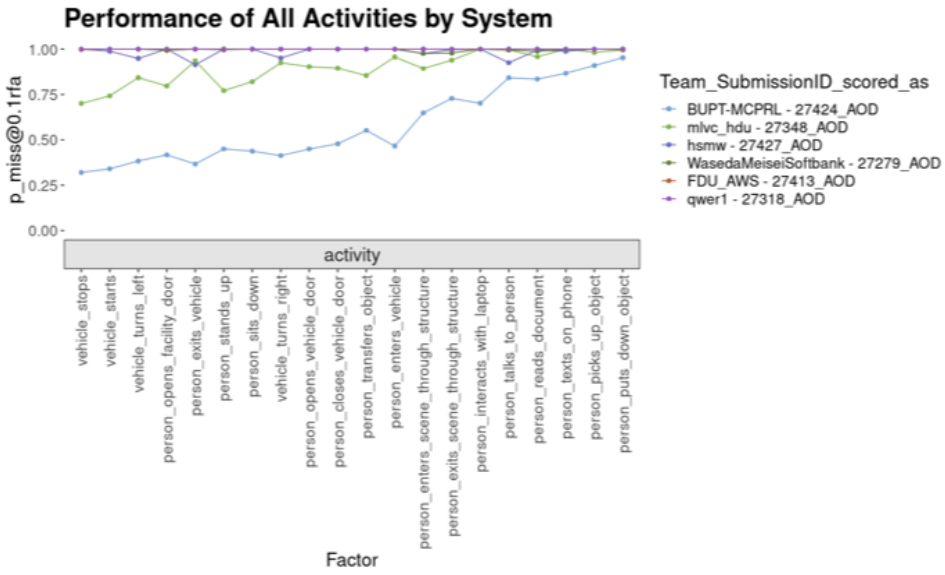

Figure 27: The AOD Activity Specific Performance for the six teams

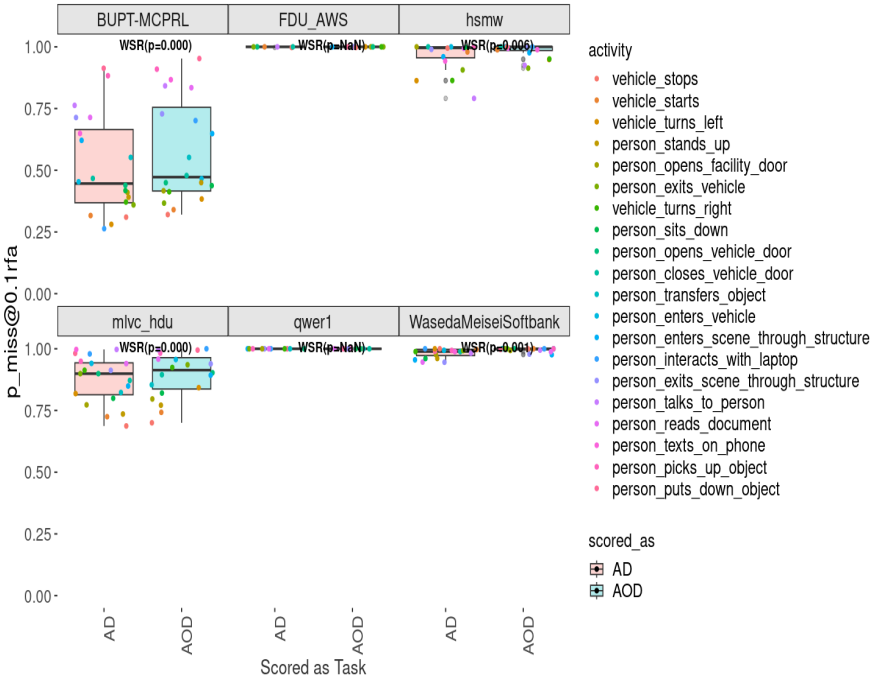

Figure 28: AD vs. AOD Detection Performance

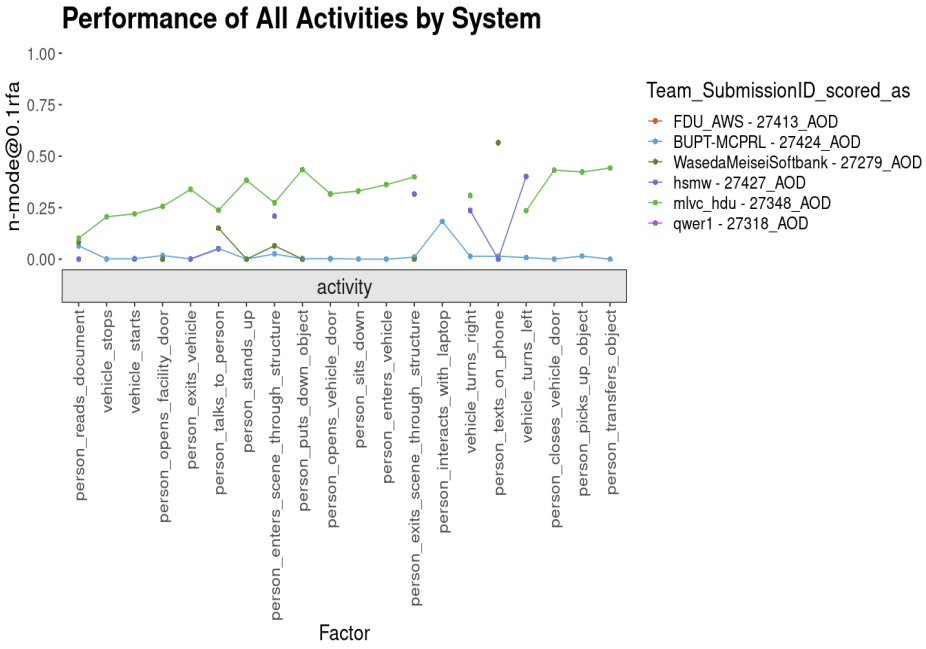

Figure 29: Localization Performance for Correct AOD Instances

the TRECVID'23 ActEV SRL evaluation, and the associated datasets will facilitate the development of activity detection algorithms. This will in turn provide an impetus for more research worldwide in the field of activity detection in videos.

4 Summing up and moving on

In this overview paper to TRECVID 2023, we provided basic information for all tasks we run this year and particularly on the goals, data, evaluation mechanisms, and metrics used. Further details about each particular group's approach and performance for each task can be found in that group's site report. The raw results for each submitted run can be found in the online proceedings of the workshop [TV23Pubs, 2023]. Finally, we are looking forward to continuing a new evaluation cycle in 2024 after refining the current tasks and introducing any potential new tasks.

5 Authors' note

TRECVID would not have happened in 2023 without support from the National Institute of Standards and Technology (NIST). The research community is very grateful for this. Beyond that, various individuals and groups deserve special thanks:

- Koichi Shinoda of the TokyoTech team agreed to host a copy of IACC.2 data.
- Georges Quénot provided the master shot reference for the IACC.3 videos.
- The LIMSI Spoken Language Processing Group and Vocapia Research provided ASR for the IACC.3 videos.
- Luca Rossetto of University of Basel for providing the V3C dataset collection.
- Baptiste Chocot of NIST associate for supporting the previous ActEV task.

Finally, we want to thank all the participants and other contributors on the mailing list for their energy and perseverance.

6 Acknowledgments

The ActEV NIST work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA), agreement IARPA-16002. The authors would like to thank Kitware, Inc. for annotating the dataset. The Video-to-Text work has been partially supported by Science Foundation Ireland (SFI) as a part of the Insight Centre at Dublin City University (12/RC/2289) and grant number 13/RC/2106 (ADAPT Centre for Digital Content Technology, `www.adaptcentre.ie`) at Trinity College Dublin. We would like to thank Tim Finin and Lushan

Han of University of Maryland, Baltimore County for providing access to the semantic similarity metric. Finally, the TRECVID team at NIST would like to thank all external coordinators for their efforts across the different tasks they helped to coordinate.

References

[ActEV23, 2023] ActEV23 (2023). Actev self-reported leaderboard (srl) challenge draft evaluation plan. `href="https://actev.nist.gov/uassets/Draft_ActEV_SRL_Eval_Plan_May10.pdf"`.

[Anderson et al., 2016] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *ECCV*.

[Anne Hendricks et al., 2017] Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. (2017). Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

[Awad et al., 2016] Awad, G., Fiscus, J., Joy, D., Michel, M., Kraaij, W., Smeaton, A. F., Quénot, G., Eskevich, M., Aly, R., Ordelman, R., Ritter, M., Jones, G. J., , Huet, B., and Larson, M. (2016). TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

[Barrault et al., 2020] Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

[Bernardin and Stiefelhagen, 2008] Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008:1.

[Bojar et al., 2017] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

[Curtis et al., 2020] Curtis, K., Awad, G., Rajput, S., and Soboroff, I. (2020). HLVU: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 355–361.

[Dong et al., 2019] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

[Du et al., 2017] Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

[Graham et al., 2018] Graham, Y., Awad, G., and Smeaton, A. (2018). Evaluation of automatic video captioning using direct assessment. *PloS one*, 13(9):e0202789.

[Graham et al., 2016] Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

[Greenberg et al., 2020] Greenberg, C. S., Mason, L. P., Sadjadi, S. O., and Reynolds, D. A. (2020). Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech & Language*, 60:101032.

[Gupta et al., 2023a] Gupta, D., Attal, K., and Demner-Fushman, D. (2023a). A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

[Gupta et al., 2023b] Gupta, D., Attal, K., and Demner-Fushman, D. (2023b). Towards answering health-related questions from medical videos: Datasets and approaches. *arXiv preprint arXiv:2309.12224*.

[Gupta and Demner-Fushman, 2022] Gupta, D. and Demner-Fushman, D. (2022). Overview of the medvidqa 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274.

[Han et al., 2013] Han, L., Kashyap, A., Finin, T., Mayfield, J., and Weese, J. (2013). UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 44–52.

[HHS, 2021] HHS (2021). Artificial intelligence (ai) strategy. U.S. Department of Health and Human Services.

[Karu and Jain, 1996] Karu, K. and Jain, A. K. (1996). Fingerprint classification. *Pattern recognition*, 29(3):389–404.

[Kasturi et al., 2009] Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., and Zhang, J. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336.

[Kitware, 2020] Kitware (2020). MEVA Data Website. `https://www.mevadata.org`. Accessed: 2020-03-12.

[Le et al., 2014] Le, V.-B., Lamel, L., Messaoudi, A., Hartmann, W., Gauvain, J.-L., Woehrling, C., Despres, J., and Roy, A. (2014). Developing stt and kws systems using limited language resources. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[Lee et al., 2018] Lee, Y., Godil, A., Joy, D., and Fiscus, J. (2018). TRECVID 2019 actev evaluation plan. `https://actev.nist.gov/pub/Draft_ActEV_2018_EvaluationPlan.pdf`.

[Li et al., 2020] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

[Lin, 2004] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[Loc et al., 2022] Loc, E., Curtis, K., Awad, G., Rajput, S., and Soboroff, I. (2022). Proceedings of lrec2022 workshop "people in language, vision and the mind" (pvlam2022). In *Proceedings of LREC2022 Workshop" People in language, vision and the mind"(P-VLAM2022)*.

[Luo et al., 2020] Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., and Zhou, M. (2020). Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

[Manly, 1997] Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman & Hall, London, UK, 2nd edition.

[Martin et al., 1997] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings*, pages 1895–1898.

[Miech et al., 2019] Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

[Mille et al., 2020] Mille, S., Belz, A., Bohnet, B., Castro Ferreira, T., Graham, Y., and Wanner, L. (2020). The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.

[Oh et al., 2011] Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 3153–3160. IEEE.

[Over et al., 2006] Over, P., Ianeva, T., Kraaij, W., and Smeaton, A. F. (2006). TRECVID 2006 Overview. `www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6overview.pdf`.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Rossetto et al., 2019] Rossetto, L., Schuldt, H., Awad, G., and Butt, A. A. (2019). V3C–a research video collection. In *International Conference on Multimedia Modeling*, pages 349–360. Springer.

[TV23Pubs, 2023] TV23Pubs (2023). `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.23.org.html`.

[Vedantam et al., 2015] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

[Yilmaz and Aslam, 2006] Yilmaz, E. and Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management (CIKM)*, Arlington, VA, USA.

[Yilmaz et al., 2008] Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, New York, NY, USA. ACM.

[Zhang et al., 2021] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

[Zhang et al., 2019] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Ad-hoc 2023 main task query topics

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

B Ad-hoc query topics - 20 progress topics

681 A woman with a ponytail
682 A person's Hands with a red nail polish
683 A building with balconies seen from the outside during daytime
684 A room with a wood floor
685 A wooden bridge
686 A round table
687 A person is throwing an object away
688 A person is washing oneself or another thing
689 A man wearing a lanyard around his neck
690 A man is seen at a gas station
691 A vehicle driving under a tunnel
692 A big building that is being camera panned or tilted from the outside
693 A person is lying on the ground outdoors
694 A person is rubbing part of their face using their hands
695 A man holding a gun but not shooting
696 A person is pouring liquid into a type of container
697 A man holding a fishing rod while being dipped in a body of water
698 A person holding a long stick which is not a drum stick outdoors
699 A person wearing a ring in their nose
700 A man wearing a dark colored hooded jacket outdoors