TRECVID 2024 - Evaluating video search, captioning, and activity recognition

George Awad {gawad@nist.gov}, Jonathan Fiscus {jfiscus@nist.gov}, Afzal Godil, {godil@nist.gov}
Lukas Diduch {lukas.diduch@nist.gov}
Information Access Division, National Institute of Standards and Technology, USA

Yvette Graham {YGRAHAM@tcd.ie}
ADAPT Centre, Trinity College Dublin, Ireland

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

April 9, 2025

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. This year TRECVID has been merged back to TREC (Text Retrieval Conference[1]) and planned the following four tracks:

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

[1]https://trec.nist.gov/

The Vimeo Creative Commons collection dataset (V3C1 and V3C2) [Rossetto et al., 2019] of about 2300 hours in total and segmented into 1.5 million short video shots was continued to support the Ad-hoc video search track. 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 track also adopted a subset of 2000 short videos from the Vimeo V3C3 dataset.

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

The AVS results were judged by NIST human assessors, while the VTT track complete ground-truth was created by NIST human assessors and runs were scored automatically later using Machine Translation (MT) metrics.

The systems submitted for the ActEV track evaluations were scored by NIST using reference annotations created by Kitware, Inc.

This paper is an introduction to the tracks,

data, evaluation framework, and performance measures used in this year's evaluation campaign for the AVS, VTT, and ActEV tracks (readers should consult the MedVidQA separate overview paper [Gupta and Demner-Fushman, 2024]). 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 [TV24Pubs, 2024].

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 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.

2.2 MEVA Dataset

The TRECVID'24 ActEV Self-Reported Leaderboard (SRL) 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'24 ActEV Self-Reported Leaderboard (SRL) test dataset is a 16-hour collection of videos with 20 activities, which only consists of

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

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

[4] CC BY-4.0 license

Electro-Optics (EO) camera modalities from public cameras. The TRECVID'24 ActEV SRL test dataset is the same as the one used for TRECVID'22-23 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.

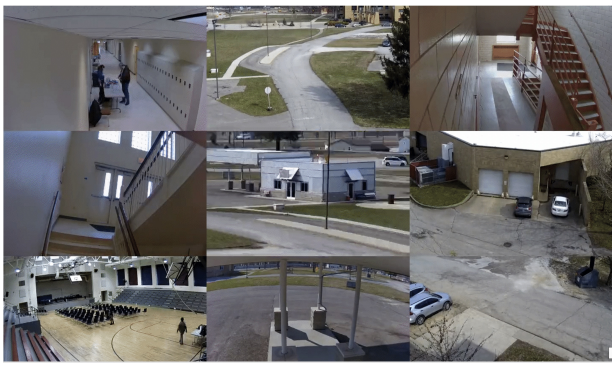

Figure 1: Montage of randomly selected video clips

2.3 TRECVID-VTT

This dataset contains short videos that are between 3 seconds and 15 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 6395 video files in webm format with Creative Commons Licenses. Those 6395 videos have been extracted from Flickr and the V3C2 dataset.

3 Evaluated Tasks

3.1 Ad-hoc Video Search

The Ad-hoc Video Search (AVS) track 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 track 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 were fixed for three years (2022-2024). Last year NIST evaluated progress runs submitted in 2022 and 2023, against 10 topics, so that teams can measure their progress against two years, while this year NIST measured their progress against three years over another 10 topics.

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 39 runs in the main task, while 8 teams submitted 79 progress runs between 2022 to 2024. One team submitted a novelty run this year. The 39 main runs consisted of 29 fully automatic, 6 manually-assisted runs, and 4 relevance feedback runs, while Progress runs consisted of 56 fully automatic and 19 manually-assisted runs.

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 to give them an opportunity to fix any judgment

mistakes.

In all, 126 903 clips were judged while 125 342 clips fell into the unjudged part of the overall samples. Total hits across the 30 topics reached 27 617 with 10 669 hits at submission ranks from 1 to 100, 11 208 hits at submission ranks 101 to 300, and 5740 hits at submission ranks between 301 to 1000. Table 1 presents information about the pooling and judging per topic.

Measures

Work at Northeastern University [Yilmaz and Aslam, 2006] has resulted in methods to estimate 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 that the metric inferred average precision (infAP) is a good estimator of average precision [Over et al., 2006]. This year, the mean extended inferred average precision (mean xinfAP) was used, allowing the 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 the adjustment of the sampling density to be greater among the highest-ranked items that contribute more average precision than those ranked lower. 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 the training type 'D', and no runs using the category 'E' or 'F' were submitted. It is encouraging to see relevance-feedback runs again this year. Tables 2, 3, and 4 show the results of all fully automatic (F), manually-assisted (M), and relevance-feedback (R) runs respectively for the main task. The *sample_eval* tool[5], a tool that implements xinfAP, was used to calculate inferred recall, inferred precision, inferred average precision, etc., for each result, given the sampling plan and a submitted run.

In general, for fully automatic results, new high scores and median (0.314) are reported with the majority of runs exceeding 30% score. For manually-assisted runs, we had four participating teams (WHU-NERCMS, VIREO, PolySmart, and NII_UIT). Overall, compared to automatic runs, manually-assisted runs performed lower (with a median score of 0.301) and performance varies between teams that participated using both run types. Regarding relevance-feedback runs, they all came from one team (WHU_NERCMS) with an overall median score of 0.332 and a top score (0.344) lower than the top automatic and manual runs.

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

Run ID (appended with priority)	Mean xInfAP
F_D_C_D_NII_UIT.24_1	0.425
F_D_C_D_NII_UIT.24_2	0.423
F_D_C_D_softbank-meisei.24_4	0.417
F_D_C_D_softbank-meisei.24_3	0.404
F_D_C_D_softbank-meisei.24_1	0.398
F_D_C_D_softbank-meisei.24_2	0.396
F_D_C_D_ruc_aim3.24_1	0.368
F_D_C_D_CERTH-ITI.24_1	0.360
F_D_C_D_ruc_aim3.24_2	0.358
F_D_C_D_CERTH-ITI.24_2	0.353
F_D_C_D_NII_UIT.24_3	0.352
F_D_C_D_NII_UIT.24_4	0.323
F_D_C_D_ruc_aim3.24_3	0.322
F_D_C_D_ruc_aim3.24_4	0.320
F_D_C_D_WHU-NERCMS.24_3	0.314
F_D_C_D_WHU-NERCMS.24_1	0.314
F_D_C_D_WHU-NERCMS.24_4	0.306
F_D_C_D_PolySmartAndVIREO.24_1	0.294
F_D_C_D_PolySmartAndVIREO.24_2	0.283
F_D_C_D_PolySmartAndVIREO.24_3	0.277
F_D_C_D_PolySmart.24_4	0.277
F_D_C_D_VIREO.24_1	0.275
F_D_C_D_CERTH-ITI.24_3	0.273
F_D_C_D_RUCMM.24_1	0.271
F_D_C_D_RUCMM.24_2	0.269
F_D_C_D_RUCMM.24_4	0.268
F_D_C_D_RUCMM.24_3	0.267
F_D_N_D_PolySmart.24_1	0.216
F_D_C_D_WHU-NERCMS.24_2	0.004

Table 2: 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 submitted runs, we applied a randomization test [Manly, 1997] to the top 10 runs for each category using a significance threshold of $p<0.05$.

For automatic runs, the analysis showed there is no statistical difference between NII_UIT team runs 1 and 2, while NII_UIT run 1 is better than CERTH-

Table 1: 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 %
1682	78986	66761	84.52	5770	8.64	211	3.66
1684	78979	67013	84.85	5454	8.14	2234	40.96
1686	78992	70021	88.64	8351	11.93	1141	13.66
1688	78995	70966	89.84	7269	10.24	808	11.12
1690	78986	69427	87.90	6439	9.27	137	2.13
1692	78959	66992	84.84	6765	10.10	4662	68.91
1694	78968	70205	88.90	7472	10.64	995	13.32
1696	78939	70275	89.02	5636	8.02	930	16.50
1698	78983	70483	89.24	7800	11.07	1552	19.90
1700	78948	71648	90.75	6295	8.79	1579	25.08
1751	39000	34086	87.40	2695	7.91	1024	38.00
1752	39000	32951	84.49	3068	9.31	1382	45.05
1753	39000	33004	84.63	3528	10.69	319	9.04
1754	39000	33033	84.70	4187	12.68	254	6.07
1755	39000	33943	87.03	3797	11.19	595	15.67
1756	39000	34527	88.53	3100	8.98	218	7.03
1757	39000	34470	88.38	2489	7.22	1396	56.09
1758	39000	34687	88.94	3293	9.49	924	28.06
1759	39000	31829	81.61	2134	6.70	358	16.78
1760	39000	32211	82.59	2244	6.97	1068	47.59
1761	39000	33984	87.14	2792	8.22	749	26.83
1762	39000	34160	87.59	2665	7.80	1157	43.41
1763	39000	34012	87.21	2579	7.58	705	27.34
1764	39000	35750	91.67	3721	10.41	505	13.57
1765	39000	34099	87.43	3029	8.88	635	20.96
1766	39000	32828	84.17	2803	8.54	451	16.09
1767	39000	34645	88.83	2725	7.87	462	16.95
1768	39000	34724	89.04	2953	8.50	522	17.68
1769	39000	35744	91.65	2837	7.94	440	15.51
1770	39000	32843	84.21	3013	9.17	204	6.77

ITI runs 1 and 2, ruc_aim3 runs 1 and 2, and softbank-meisei runs 1 and 2. On the other hand, NII_UIT run 2 is better than CERTH-ITI runs 1 and 2, and ruc_aim3 runs 1 and 2. For team softbank-meisei, their run 4 came better than their runs 1, 2, and 3, as well as CERTH-ITI run 1 and 2, and ruc_aim3 runs 1 and 2. softbank-meisei runs 1, 2 and 3 are all better than CERTH-ITI runs 1 and 2, and ruc_aim3 run 2. Finally, ruc_aim3 run 1 is better than ruc_aim3 run 2.

With respect to manually-assisted runs, the test indicated that NII_UIT runs 1 and 2 are both better than PolySmartAndVIREO runs 1 and 2, VIREO run 3, and wHU-NERCMS run 1, while there is no significant difference between the two runs of NII_UIT. Team WHU-NERCMS team run 1 came better than PolySmartAndVIREO runs 1 and 2, as well as VIREO run 3. For team, PolySmartAndVIREO, their run 2 is better than run 1, while for VIREO team, their run 3 is better than PolySmartAndVIREO run 1.

Finally, for R runs, it was indicated that there is no difference between the top 2 runs, while runs 1 and 2 are better than run 3.

Run ID (appended with priority)	Mean xInfAP
M_D_C_D_NII_UIT.24_2	0.422
M_D_C_D_NII_UIT.24_1	0.417
M_D_C_D_WHU-NERCMS.24_1	0.322
M_D_C_D_VIREO.24_3	0.280
M_D_C_D_PolySmartAndVIREO.24_2	0.274
M_D_C_D_PolySmartAndVIREO.24_1	0.000

Table 3: AVS: Sorted scores of 6 manually-assisted runs across all 20 main queries. All runs used training type "D". Run names are prefixed by "C" (common) or "N" (novelty)

Run ID (appended with priority)	Mean xInfAP
R_D_C_D_WHU-NERCMS.24_2	0.344
R_D_C_D_WHU-NERCMS.24_1	0.337
R_D_C_D_WHU-NERCMS.24_4	0.328
R_D_C_D_WHU-NERCMS.24_3	0.324

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

Team	Relevant shots
VIREO	2851
NII_UIT	1035
RUC_AIM3	593
WHU-NERCMS	564
PolySmartAndVIREO	557
ITI_CERTH	426
RUCMM	294
Softbank-meisei	281
kindai_ogu_osaka	97

Table 5: AVS: Sorted unique number of hits (true positive shots) by team for main and progress tasks.

Table 5 shows the number of unique clips, for main and progress tasks, found by the different participating teams. From this table and the overall scores in Tables 2, 3, and 4, it can be shown that there is no clear relation between the teams that found the most unique shots and their total performance with the exception of team NII_UIT that achived top overall score as well as unique hits. The VIREO team contributed the most unique hits (similar to previous year). Although Softbank-meisei and ITI_CERTH teams performed well, their unique hits contributions were not very high.

Figures 2 shows 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 may rank differently across the 20 queries. Some samples of different performing queries are labeled with the query text. In general, queries that require more details and conditions to be satisfied tend to be more hard to retrieve.

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 3 shows the novelty metric scores. The red bars indicate the single submitted novelty run.

For 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 the PolySmart team) scored fourth place based on our metric, while NII_UIT run 1 ranked highest in novelty and also overall score. It can be shown this year that overall best-performing runs also retrieved many unique shots. 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 4 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

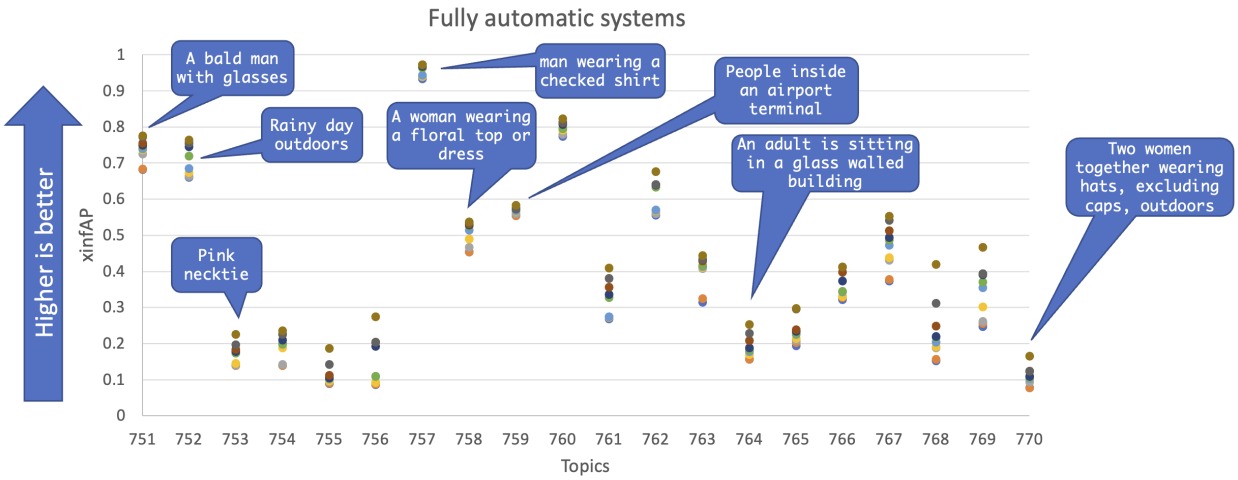

Figure 2: AVS: Top 10 runs (xinfAP) per query (fully automatic)

same time. In general, most automatic systems reported processing time below 10 s.

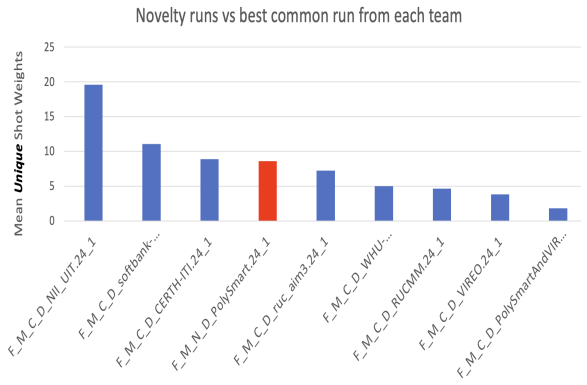

Figure 3: AVS: Novelty Runs Scores

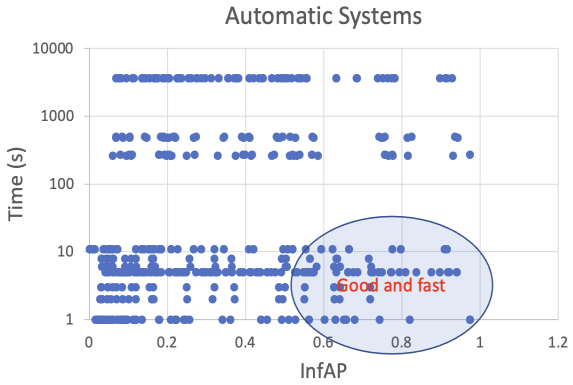

Figure 4: AVS: Processing time vs scores (fully automatic)

The progress task results are shown in Tables 7 and 6 for automatic and manually-assisted systems, respectively. Each table represents 10 progress queries. Set A progress queries were evaluated in 2023 and re-scored again in 2024 using ground truth built in 2023, while set B queries were fully assessed in 2024 using runs submitted in the last 3 years. In total, 7 teams participated in this progress task during the last three years, two teams submitted in 2022 only, and one team submitted in 2023 and 2024. Comparing the best run in these three years for each team, we can see that for automatic systems, all teams submitted in all three years achieved better in 2024 except for the RUCMM system of 2023 which scored higher for set B queries. For manually-assisted systems, only VIREO submitted in all years while team NII_UIT participated in two years. Both teams achieved better performance in 2024. For set B queries, it can be shown that the performance increase ranged from 19.6% to 37% with an average of 30% for automatic systems, and 36% to 65% and an average of 50% for manually-assisted systems, whereas for set A queries, the performance increase ranged between 0.7% to 59% with an average of 23% for automatic systems, and 33% to 60% with an average of 47% for manually-assisted systems.

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

Team	Automatic systems	Manually-assisted systems
RUCMM (2022)	0.275	
RUCMM (2023)	**0.311**	
RUCMM (2024)	0.305	
VIREO (2022)	0.177	0.186
VIREO (2023)	0.215	0.217
VIREO (2024)	**0.343**	**0.34**
NII_UIT (2023)	0.161	0.163
NII_UIT (2024)	**0.482**	**0.47**
ITI_CERTH (2022)	0.239	
ITI_CERTH (2023)	0.278	
ITI_CERTH (2024)	**0.389**	
RUCAIM3-Tencent (2022)	0.201	
kindai_ogu_osaka (2022)	0.217	
WasedaMeiseiSoftbank (2022)	0.278	
WasedaMeiseiSoftbank (2023)	0.335	
WasedaMeiseiSoftbank (2024)	**0.417**	

Table 6: AVS: Max performance (xInfAP score) per team on set 'B' 10 progress queries

Team	Automatic systems	Manually-assisted systems
RUCMM (2022)	0.237	
RUCMM (2023)	0.258	
RUCMM (2024)	**0.260**	
VIREO (2022)	0.137	0.149
VIREO (2023)	0.171	0.134
VIREO (2024)	**0.217**	**0.202**
NII_UIT (2023)	0.152	0.150
NII_UIT (2024)	**0.371**	**0.379**
ITI_CERTH (2022)	0.191	
ITI_CERTH (2023)	0.216	
ITI_CERTH (2024)	**0.268**	
RUCAIM3-Tencent (2022)	0.185	
kindai_ogu_osaka (2022)	0.205	
WasedaMeiseiSoftbank (2022)	0.256	
WasedaMeiseiSoftbank (2023)	0.286	
WasedaMeiseiSoftbank (2024)	**0.351**	

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

of xInfAP >= 0.5 for any given topic and assumed that those runs with 0.5 or higher were the easiest topics, while topics with xInfAP < 0.5 were assumed to be difficult topics. From this analysis, it can be concluded that the top 5 hard topics were "A person is rubbing part of their face using their hands", "A person's Hands with a red nail polish", "A person is pouring liquid into a type of container", "A round table", and "A person holding a long stick which is not a drum stick outdoors". On the other hand, the top 5 easiest topics were "A big building that is being camera panned or tilted from the outside", "A man wearing a checked shirt", "A rainy day outdoors", "A room with a wood floor", and "A man inside a workshop".

Ad-hoc Observations and Conclusions

Compared to detecting single concepts (e.g., airplane, animal, bridge), it can be seen from running the ad-hoc task for the last 9 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 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 the major observations in 2024, 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, 39 systems (29 automatic, 6 manually-assisted, and 4 relevance-feedback) were submitted in the main task including 1 novelty run, while 79 runs were submitted for the progress task through the past 3 years. Overall, performance scores are higher than the last two years 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. Few teams managed to retrieve unique shots while also achieving high performance. Overall, 21.7% of all judged shots across all queries are true positives, and about 21% of them are unique (submitted by a single team). Hard queries are the ones asked for unusual combinations of facets (compared to well-known concepts commonly found in the available training datasets). Progress task has been very useful to track system improvements for the last three years and the majority of systems reported higher scores that ranged on average 37%.

In terms of system's approaches, some general trends can be observed including extensive reliance on pre-trained transformer-based models like CLIP, BLIP, BEiT, and OpenCLIP for embedding text and video features, the emphasis on ensembles and combinations of models for better retrieval precision, usage of re-ranking techniques and normalization strategies to refine search results, and integration of cutting-edge multimodal and open-vocabulary detection models.

For detailed information about the approaches and results for individual teams, we refer the reader to the reports [TV24Pubs, 2024] in the online workshop notebook proceedings.

3.2 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 by 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 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 have been submit-

	Number of runs
BUPT_MCPRL	3
Kslab	4
PolySmart	7
RUC_AIM3	8
Softbank-Meisei	7

Table 8: VTT: List of teams participating and their submitted runs

ting results annually to measure and track system improvements 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. A robustness sub-task was also supported 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 29 runs including 18 runs in the main task and 11 runs in the robustness task. A summary of participating teams is shown in Table 8.

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 were 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 public 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 1757 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, our results will be reported for 1757 new videos (non-progress) and the 300 videos in progress set.

It is important for a good dataset to have a diverse

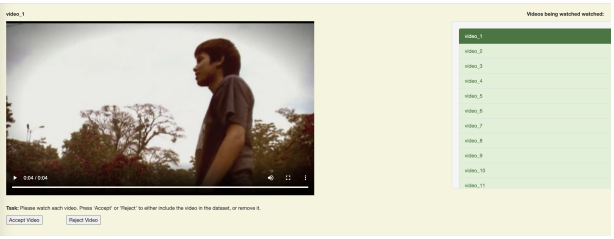

Figure 5: VTT: Screenshot of video selection tool.

set of videos, so we reviewed around 9000 videos and selected 2000 videos. Figure 5 shows a screenshot[6] of the video selection tool that was used to decide whether a video was to be selected or not. We tried to ensure that the videos covered a large set of diverse topics including spatial and temporal description aspects. If we came across 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	22.03	1757
2	20.62	2000
3	26.98	1893
4	20.8	2000
5	36.99	1764

Table 9: VTT: Average number of words per sentence for all the annotators. The table also shows the number of videos watched by each annotator.

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

Annotation Process The videos were divided among 5 annotators, with each video being annotated once by each to create 5 annotations per video. Due to time limitations, three out of the 5 annotators could not finish annotating the full 2000 videos and therefore we selected the 1757 videos for which we had 5 annotations.

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 9 shows the average number of words per sentence for each of the annotators. The average sentence length varies from 20 words to 36 words, emphasizing the difference in descriptions provided by the annotators. The overall average sentence length for the dataset is 20.7 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 2.69 (on a scale of 1 to 5), showing that the annotators thought the videos were close medium level of difficulty on average. The average score for the second question was 2.39 (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 -0.7.

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, 29 runs were submitted and distributed as follows: 8 runs were of type "VA" (Audio and visual features from video datasets), and 21 runs were 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 self-critical reinforcement learning, self-critical reinforcement learning, categorical crossentropy, and contrast_loss. Figure 7 shows a sample of a video with captions by human annotators as well as submissions by automatic systems.

Evaluation and Metrics

The description generation task scoring was done automatically using different popular metrics borrowed from machine translation and image captioning domains as mentioned below.

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.

Results

The metric score for each run is calculated as the average of the metric scores for all the descriptions within that run. Table 10 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 that team BUPT_MCPRL performed the highest in most metrics, followed by RUC_AIM3 and Softbank-Meisei teams.

Table 11, on the other hand, shows the results for the three teams that participated in the robustness sub-task (introducing noise to the testing dataset). It can be shown that most systems performed lower on the robustness task with few exceptions such as softbank-meisei and PolySmart teams' scores on the BLEU metric which was slightly higher than main task scores.

Table 12 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 0.854 to 0.988.

Teams were asked to provide a confidence score for each generated sentence. Figure 6 shows the submitted average confidence scores for each run against each metric score. There seems to be a weak correlation between confidence and some metric scores.

Table 13 shows the automatic metrics scores for the progress sub-task which evaluated runs on 300 fixed videos between 2021 and 2024. The table shows only teams who submitted in at least two years. It can be shown that all teams performed in 2024 better than previous years with one exception for team MLV_HDU, where they performed consistently in 2022 better than in 2023. Overall, average improvements ranged between 4% for CIDER and METEOR metrics to 37% for BLEU metric.

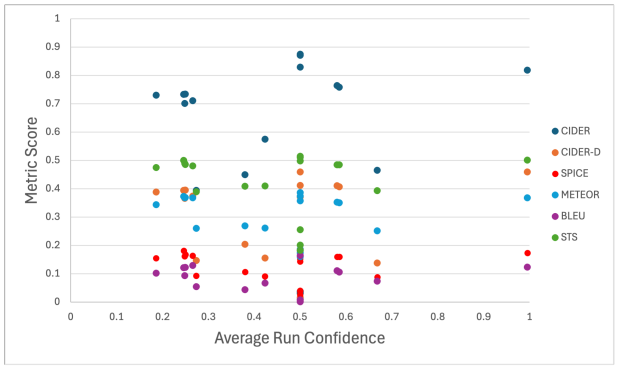

Figure 6: VTT: system reported sentence confidence scores against the various metric scores.

Task observations and conclusions

The VTT task has been running since 2016. 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.

	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
Kslab	0.073	0.269	0.575	0.204	0.106	0.409
PolySmart	0.007	0.186	0.038	0.016	0.040	0.255
RUC_AIM3	0.123	0.368	0.818	**0.459**	0.173	0.500
BUPT_MCPR	**0.183**	**0.386**	**0.875**	**0.459**	0.166	**0.514**
Softbank-Meisei	0.129	0.372	0.734	0.395	**0.181**	0.500

Table 10: VTT: Top score by each team for all automatic metrics (Main task).

	BLEU	METEOR	CIDER	CIDER-D	SPICE	STS
RUC_AIM3	0.121	**0.364**	**0.813**	**0.450**	**0.170**	**0.469**
Softbank-Meisei	**0.134**	0.360	0.724	0.373	0.160	0.451
PolySmart	0.008	0.183	0.036	0.014	0.040	0.150

Table 11: 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	0.960	0.948	0.968	0.932	0.978
CIDER-D	0.960	1.000	0.980	0.984	0.898	0.938
SPICE	0.948	0.980	1.000	0.988	0.854	0.956
METEOR	0.968	0.984	0.988	1.000	0.912	0.963
BLEU	0.932	0.898	0.854	0.912	1.000	0.883
STS	0.978	0.938	0.956	0.963	0.883	1.000

Table 12: VTT: Correlation between overall run scores for automatic metrics (Main task).

	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
RUC_AIM3 (2024)	**0.144**	**0.418**	**0.933**	**0.599**	**0.195**	**0.529**
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
WasedaMeiseiSoftbank (2024)	**0.136**	**0.409**	**0.848**	**0.523**	**0.199**	**0.525**
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
Kslab (2024)	**0.099**	**0.297**	**0.642**	**0.307**	**0.123**	**0.425**
BUPT_MCPRL (2023)	0.091	0.278	0.62	0.267	0.1	0.39
BUPT_MCPRL (2024)	**0.185**	**0.403**	**0.949**	**0.569**	**0.175**	**0.531**
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 13: VTT: Top score by each team for 300 progress videos (measured from 2021 to 2024) for all automatic metrics (Main task).

GT:
1- An adult hand holds a baby's hand and lets it go
2- An individual holds the hand of a baby in a crib
3- Man's hand holds infant's hand, moving it slightly
4- An adult hand playing with a baby's hand
5- Someone is holding a baby's hand indoors

Submissions:
1- A person is putting a white substance on the nose of a person laying down
2- A person is **holding** a child's **hand**
3- A person is **holding** a pen, possibly writing or drawing
4- A person's **hands** are holding a small object in a dark room
5- A white man is **holding** a white woman's **hand**

Figure 7: VTT: Sample of video captions by humans (green box) vs submitted sentences by systems (red)

This was the second year using the V3C3 test data as well as the seconnd year to introduce a robustness sub-task. The robustness sub-task this year incorporated more real world harder transformations such as change in lighting, camera shaking, etc. which compared to last year show that given this transformations, most systems could not cope well and their performance was lower than the main task on the same videos.

The progress sub-task (300 fixed videos across four years) concludes that this year's systems are better than the previous three years with significant average improvements across all metrics. High correlation exists between all automatic metrics. Few runs reported the employment of audio features in their runs.

General trends across this year's systems include heavy reliance on large pre-trained multimodal models (BLIP2, BLIP3, LLAVA, EVA-CLIP), data augmentation played a central role in improving performance (back-translation using Google Translate API, and augmentation with GPT-3.5 to enhance training data), use of advanced segmentation techniques (KTS) to optimize keyframe selection, and integration of models for specific tasks such as captioning (BLIP2/3, BART) and reranking (EVA-CLIP).

For detailed information about the approaches and results for individual teams' performance and runs, we refer the reader to the site reports [TV24Pubs, 2024] in the online workshop notebook proceedings.

3.3 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'24 ActEV Self-Reported Leaderboard (SRL) Challenge is based on the Multiview 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'23 ActEV SRL and TRECVID'23 ActEV SRL evaluations. 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) evalu-

ation, 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 8 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.

Figure 8: Example of activities for MEVA dataset used ActEV SRL evaluation. IRB (Institutional Review Board): ITL-00000755

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'24 ActEV SRL submissions and discuss observations and conclusions.

Task and Dataset

In the TRECVID'24 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: Activity and Object Detection (AOD). 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: Activity Detection (AD). 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 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 [ActEV24, 2023].

Table 14 lists the 20 activity names for TRECVID'24 ActEV SRL evaluation, based on the

Table 14: A list of activity names for TRECVID ActEV SRL evaluation, there were 20 activities based on the MEVA dataset.

person_closes_vehicle_door	person_reads_document
person_enters_scene_through_structure	person_sits_down
person_enters_vehicle	person_stands_up
person_exits_scene_through_structure	person_talks_to_person
person_exits_vehicle	person_texts_on_phone
person_interacts_with_laptop	person_transfers_object
person_opens_facility_door	vehicle_starts
person_opens_vehicle_door	vehicle_stops
person_picks_up_object	vehicle_turns_left
person_puts_down_object	vehicle_turns_right

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 the video reviewed.

For both the AOD (primary) and AD (secondary) tasks for TRECVID'24 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 high-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 9 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

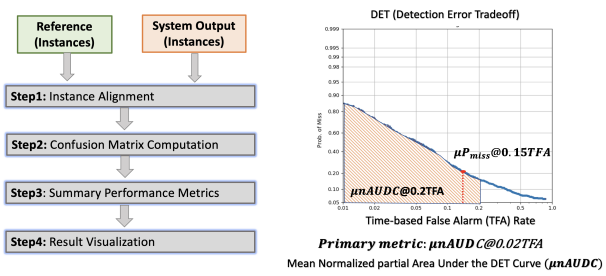

Figure 9: Performance measure calculation and Detection Error Tradeoff (DET) curves

x-axis is the Time-based False Alarm (TFA) Rate and the 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 10, 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 the DET curve is computed separately for each activity over all the 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 AOD metric is calculated by considering both the temporal overlap and the bounding boxes overlap of all the objects associated with the activities of the reference and system output instances. This is covered in further detail in the evaluation plan [ActEV24, 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. $\text{CD}_t(\tau)$ is the count of

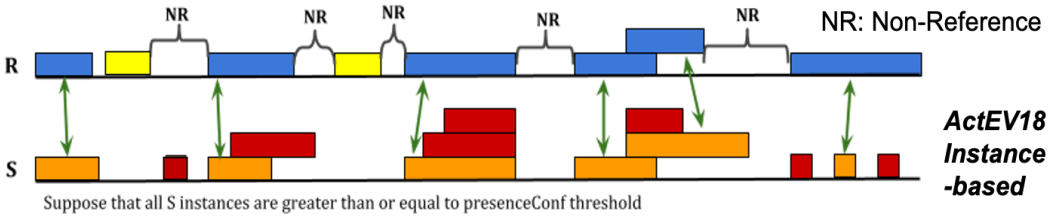

Figure 10: Illustration of activity instance alignment. R is the set of reference instances and S is the set of the system instances. Green arrows connect R and S instances that are determined to be aligned and thus labeled correct detections.

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 is as follows:

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

where 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 three teams from academia and industry from 3 countries participated in the TRECVID'24 ActEV SRL evaluation. Each participant was allowed to submit multiple system outputs and a total of 68 submissions were received. Table 15 lists the participating teams along with results ordered by $mean\_P_{miss}0.1RFA$ values scores for the top performing system per team along with $nAUDC@0.2RFA$ values. The top $mean\_P_{miss}0.1RFA$ performance on activity detection is by Mlvc_hdu at 82.32% followed by hsmw at 83.30% and M4D_team_2024 is third at 98.38%.

Figure 11 shows the performance based on the Activity and Object Detection (AOD) DET Curve for the 3 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 the best performance for $mean\_P_{miss}@.1RFA$ of 82.3% for team MLVC_HDU.

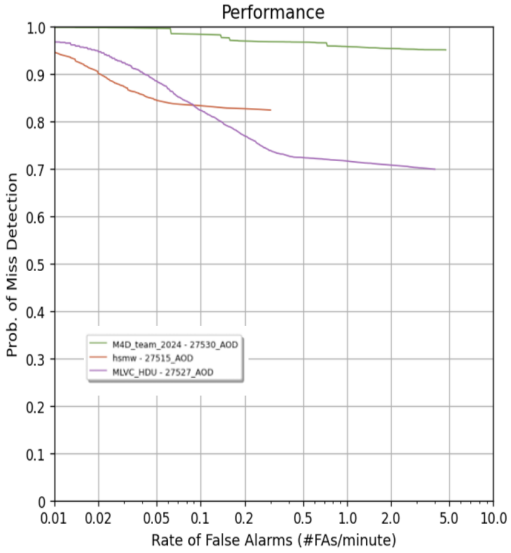

Figure 11: Activity and Object Detection (AOD) DET Curve for the three teams.

Figure 12 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 13 shows the AD vs. AOD Detection Performance for the three 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

Table 15: Summary of participants' information and results ordered by AOD, $\mu nAUDC$ values. The AOD values of $mean\_P_{miss}$@.1RFA values along with the $nMODE$@0.1RFA are also presented. We also present the AD values of $nAUDC$@0.2RFA and $mean\_P_{miss}$@0.1RFA. 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
MLVC_HDU	Hangzhou Dianzi University, China	0.8232	0.0382	0.7685	0.7903
hsmw	Hochschule Mittweida University	0.8330	0.0622	0.7771	0.7961
M4D_team_2024	Centre for Research and Technology Hellas	0.9838	0.0132	0.9643	0.9588

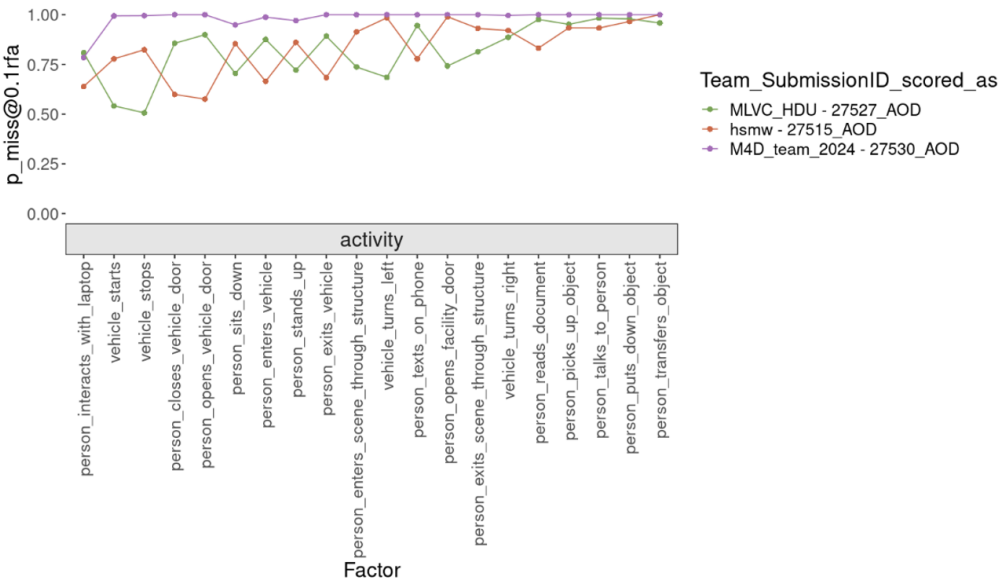

Figure 12: The AOD Activity Specific Performance for the three teams

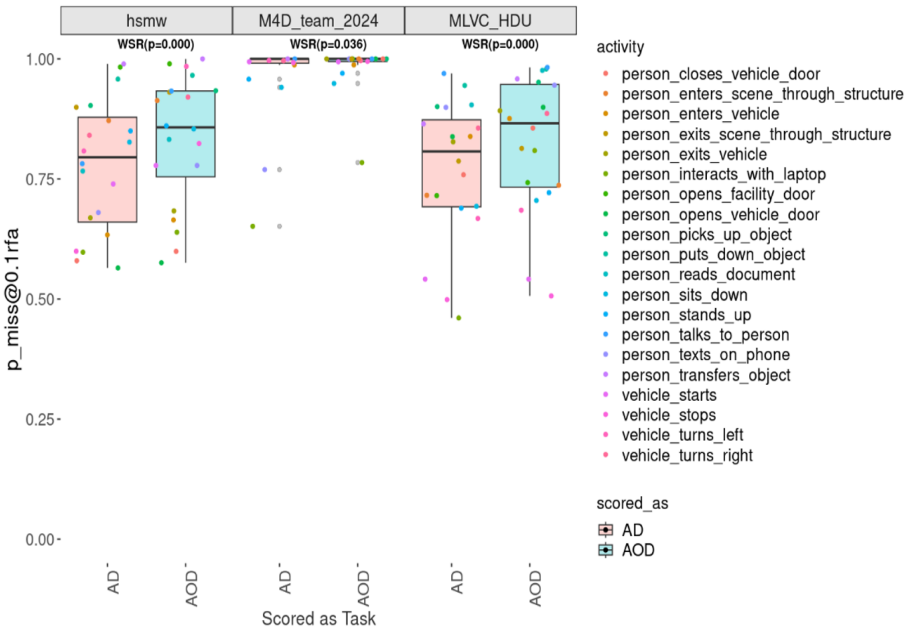

Figure 13: AD vs. AOD Detection Performance

AOD instances, Figure 14 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 most of the activities well.

Summary

In this section, we presented the TRECVID'24 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'24 ActEV evaluation was based on the MEVA [Kitware, 2020] dataset and had 20 target activities in total. This was the fifth time the MEVA dataset has been used for the ActEV evaluation. Three teams from 3 countries participated in the ActEV SRL evaluation and made a total of 68 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 team had the top-performing system followed by the hsmw team. The Detection and Localization (AOD) still remains a more difficult task for the teams.

The TRECVID'24 ActEV SRL evaluation provided researchers an opportunity to evaluate their activity detection algorithms on a self-reported leaderboard. We hope the TRECVID'24 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 2024, we provided basic information for all tasks we run this year and, particularly, on the goals, data, evaluation mechanisms, metrics used, and high-level results analysis.

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 [TV24Pubs, 2024]. Finally, we look forward to continuing a new evaluation cycle in 2025 after refining the current tasks and introducing any potential new tasks.

5 Authors' note

TRECVID would not have happened in 2024 without support from the National Institute of Standards and Tech-

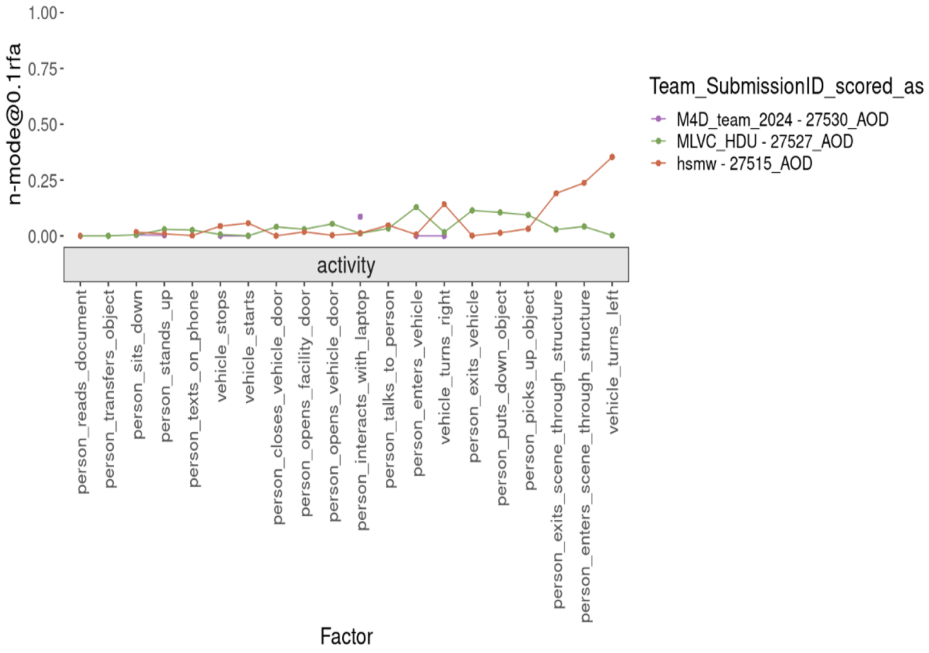

Figure 14: Localization Performance for Correct AOD Instances

nology (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.

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

[ActEV24, 2023] ActEV24 (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*.

[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.

[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.

[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 and Demner-Fushman, 2024] Gupta, D. and Demner-Fushman, D. (2024). Overview of trec 2024 medical video question answering (medvidqa) track. *arXiv preprint arXiv:2412.11056*.

[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.

[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`.

[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.

[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.

[TV24Pubs, 2024] TV24Pubs (2024). `https://trec.nist.gov/proceedings/proceedings.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.

A Ad-hoc 2024 main task query topics

751 A bald man with glasses
752 A rainy day outdoors
753 A pink necktie
754 A white sweater
755 A person is wiping themselves or an object using their bare hands or other object.
756 A man is putting on a jacket or a t-shirt
757 A man wearing a checked shirt
758 A woman wearing a floral top or dress
759 People inside an airport terminal
760 A man inside a workshop
761 A traffic light seen at an intersection of a road or street
762 A map seen on a wall indoors
763 At least two persons in a hallway are seen walking
764 An adult is sitting in a glass walled building
765 An adult is wrapped in a blanket
766 A person holding a pen
767 A seated person reading from a paper or book outdoors during daytime
768 A woman wearing a silver necklace around her neck
769 Two or more persons indoors with coffee cups or mugs seen with them.
770 Two women together wearing hats, excluding caps, outdoors

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