A ASAP

A.1 Alignment Verification

We verify the accuracy of annotations made by ASAP by providing human annotators with a clip containing a contiguous sequence of events, and asking them to provide the timestamps in the video for when each event occurred. Additionally, all scorecard information is masked in each provided clip.

Verification of different sports For cricket, we built an AMT interface and asked annotators to provide both the timestamps and events that occurred in a clip for over 1200 events to verify that both the ASAP alignment process and video quality were sufficient, which we discuss more in Appendix B. For verifying and demonstrating the generality of ASAP pipeline, we annotate three different sports, namely, American football, football, and basketball, and verify it using a similar interface. Due to the limited mturk budget, we used two of the in-house annotators for the verification of these three sport's annotations by providing the humans with clips from 6 hours of match footage for each sport and had them verify (by annotating) 240 events for each sport.

American Football Alignment Issues We note that the reason why the verification accuracy for American football in Figure 4 is lower than the other sports is because for most standard plays, the timestamps provided are for when the play started. However, when a team scores or is given a penalty, the timestamp provided for the next play is either the end of the play, or when it happened. We were only able to have ASAP account for the touchdown instances, but not the penalty instances, which is generally what was marked incorrect during our verification process.

A.2 Annotation Event Details

Events for Different Sports In this section, we describe the events that we considered for each sport.

- **Cricket:** Each legal delivery was considered a valid event, where features such as the number of runs and the occurrence of a wide/out ball were marked as well. See Appendix B for further details.
- **American Football:** Each play was considered a valid event, so we considered *punts, field goals, complete passes, incomplete passes, run-plays, sacks, penalties,* and *spikes* as distinct.
- **Football/Soccer:** There are no distinct, sequential plays in football, so we based our events off of online commentary. We mark *shots off target, shots on target, shots on woodwork, goals, fouls, substitutions, yellow cards, red cards, corner kicks, free kicks, offsides, handballs,* and *saved/blocked balls* as distinct events to be annotated and aligned.
- **Basketball:** Like football/soccer, there are no distinct plays that happen, so we mark *fouls, jumper shots, layups, dunks, free throws*, and *regular shots* as distinct events that we annotate and align.

Granularity of Annotations Because the aligned annotations for different sports rely on the timestamps provided by the online commentary source, we observe that different sports are annotated with varying levels of granularity. Thus, when we verify the accuracy of an aligned annotation, we account for these differing levels of granularity with different margins for error. For example, in football, annotations are provided at a minute-level, so if the human annotator marks the event as occurring anywhere outside that range, we consider the annotation to be incorrect; however, for sports like basketball, where annotation timestamps are given by the second, we provide a margin of error of ± 1 second to the timestamp marked by the human. Similar to football, in cricket, an event lasts for 30-40 seconds, so if a human annotator is able to mark the event as occurring anywhere inside that range, we consider the annotation to be correct.

A.3 Raw Videos Source

All of the videos that we ran ASAP through were found on YouTube channels. For cricket we used 131 videos, and for all other three sports we annotated 3 videos each. The average video length of a cricket match is 7.5 hrs while for the other sports it is 1.5 hrs. We also provide the links to all the videos annotated with the supplementary document.

A.4 Qualitative Example

We provide a qualitative sample by attaching an annotation along with a sports (cricket) video snippet. The annotation is present as a `.srt` file and can be used as subtitle with the clip present to see the alignment accuracy of our pipeline.

B LCric

B.1 Primer on Cricket

In this section we further extend our primer to Cricket provided in Section 4.1 by describing the Batting/Bowling phases, as well as the primary objective of the game.

Overview: Cricket is a ball-and-bat sport played by two teams of eleven players each. Cricket is scored by "runs", and at the end of the game, the team with the most scored "runs" wins. The game is played in an inning-format, where one team is batting, and the other team is fielding. We describe the two phases below.

Bowling Phase: When a team is in the bowling phase, all 11 players stay on the field. One of the players is designated as the bowler, and their job is to deliver the ball to the batter (hitter) on the batting team. If the ball is struck by the batsman, the remaining players, called fielders, try to prevent the ball from reaching the boundary of the field and return the ball back to the pitch area. A single over consists of six deliveries bowled by the same player, and each team delivers a set number of overs depending on the tournament type in their bowling phase.

Batting Phase: When is team is in the batting phase, only two players on the team stay on the field at a time. The batsman's job is to score runs and defend their wickets. A single

run is scored when the batsman hits the ball and runs from one end of the pitch to another. Another way to score runs is to hit the ball to the boundary of the field, which is called the **'boundary'**, giving 4 or 6 runs to the batting team. In total, each batting team has 10 wickets.

Objective: During an inning, the batting team wants to score as many runs as possible, while the bowling team wants to take as many wickets as possible to stop the batting team from scoring. In most single-day matches, the bowling team will bowl for 50 overs before the teams switch roles for the second half of the game. At this point, the goal of the new batting team is to outscore the previous team in runs before 50 overs or before losing all of their wickets.

B.2 Training and Implementation Details

We use consistent training schemes for both TQN (Zhang, Gupta, and Zisserman 2021a) and MeMViT (Wu et al. 2022a) to provide a fair comparison between the two baselines. Both models were trained for 50 epochs on 4 V100 GPUs with a batch size of 4. We used a base learning rate of $LR = 0.01$ with the Adam optimizer and default hyperparameters.

Baseline Implementations For setting up TQN as a baseline, we used the official code provided by the authors with some minor modifications to the output heads for answering **LCric** queries. For MeMViT, since there is no official implementation released at the time of writing, we implemented our own version using the same implementation details as the main paper. Our implementation is built on top of the official implementation of MViT (Fan et al. 2021), which is the base model used to create MeMViT.

B.3 LCric Queries

Query Set Generation Algorithm We describe our query set generation process in Algorithm 1, where we use logical operators and a set of possible atomic events form form different combinations of queries.

Binary Queries Statistics For our 10-over experiments, we formed a balanced set of 32 queries by taking queries from the set formed by Algorithm 1 and pruning them down so that given a random 10-over clip sampled uniformly from **LCric**, there would be a 0.5 ± 0.05 probability that the query would hold true on that clip. We list the set of all such queries and their corresponding probabilities in Table **??**.

Multi-choice Query Statistics We also generated a set of multi-choice queries for our 10-over experiments. These queries include a mix of common and less common event chains that generally occur between 0-9 (inclusive) times within any 10-over clip. The frequency of occurrence of these clips within our train set is provided in Figure 6.

B.4 AMT Interface

We built an AMT interface for verifying ASAP's alignment of cricket annotations to videos, with the full instructions and interface provided in Figure 7.

Algorithm 1: Query Set Generation

```
# The set of atomic events: [0,1,2,...,9,W,w]
Set of atomic events: A_e
# The number of queries for the query set
Size of the query set: n_q
query_set = []
for i in range(n_q) do
    # Step A: getting raw operators and combinators choice
    num_joins ~ [1,5]
    # total length for operators set being sampled for determining the query length
    ops = random.choices([atleast(), atmost(), inrange()], num_joins)  # sampling list of operators
    combine_op = random.choices([and, or], 1)  # sampling the combination operator
    # Step B: instantiating a query for query set for ??
    query = []
    for op in ops do
        # specify lower bound for atleast/inrange ops
        occ_min ~ [1,10]
        # specify upper bound bound for atmost/inrange ops
        occ_max ~ [occ_min,10]
        # sample atomic events in query
        atomic_event ~ [A_e]
        # Using the above variables for defining an occurrence pattern for atomic_event
        instanced_op = op(occ_min, occ_max, atomic_event)
        query.append(instanced_op)
    final_query = join_op(query)
    query_set.append(final_query)
```

Instruction Details Each annotator is given a set of instructions to read prior to beginning the main annotation task, called a HIT (Human Intelligence Task). For each task, the annotator is given a video clip from a sports match. The task is to classify each legal delivery/ball that occurred in the video, as well as the timestamp at which the annotator was able to gather enough information to answer this question. Additionally, we provide a set of examples for what each event looks like to the annotators, as well as a fully annotated example and video, as shown in Figure 8, 9.

Task Interface Details Each HIT contains a 1-over video and 6 rows, each corresponding to a legal delivery that occurred in the video. Each row consists of a dropdown for inputting the number of runs scored in that delivery, a checkbox for indicating an out ball occurred, a checkbox for indicating a wide ball occurred, and a field for writing the timestamp at which this information can be found. Figure 7 shows what the annotators initially see, as well as an example of how to fill it out.

LCric Annotation Verification A total of 205 overs with 1230 events spanning $\sim$1000 minutes were labeled by human annotators and compared to ground truth annotations from ESPNCricinfo. For each ball, we consider an event annotation to be correct if it was classified completely correctly. The timestamp annotation is marked as correct if it occurred anytime within the timestamp range specified by the ground truth ± 1 seconds.

LCric Annotation Statistics We found that in total, $1185/1230(96.34\%)$ of balls were classified correctly, while $1213/1230(98.62\%)$ of ball timestamps were marked correctly. Additionally, assuming human annotators can aggregate and reason easily with logic, we aggregate their annotations to answer queries in our test set, which provides our human baseline. We find that the human annotations achieve an accuracy of $5541/5740(96.53\%)$ on the test query set – exceeding the TQN and MemViT baselines by a large margin.

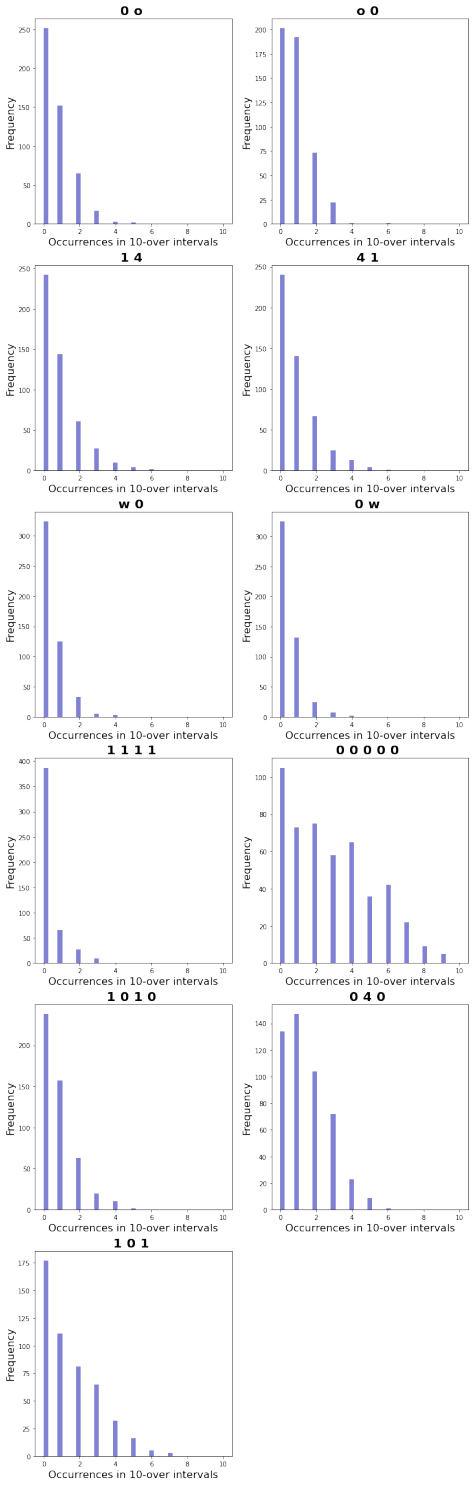


Figure 6: Ground truth output frequencies to queries used in multi-choice queries in the train set of **LCric**.

Queries	GT probability
atmost 7 1's	0.451
atleast 4 4's	0.523
atleast 5 1's AND atleast 3 4's	0.528
atleast 2 2's AND atleast 3 4's	0.452
atleast 4 4's AND atmost 5 o's	0.452
atleast 4 4's AND atmost 3 5's	0.456
atleast 4 2's OR atmost 2 4's	0.539
atleast 4 3's OR atmost 3 4's	0.544
atleast 5 2's OR atleast 4 4's	0.526
atleast 3 2's OR atleast 2 w's	0.485
atmost 3 4's AND atmost 2 6's	0.529
atmost 3 4's AND atmost 3 7's	0.544
atmost 2 0's OR atmost 3 4's	0.544
2 inrange [1, 6] AND 4 inrange [1, 4]	0.539
4 inrange [1, 6] AND o inrange [1, 4]	0.555
1 inrange [2, 7] OR 2 inrange [4, 5]	0.506
1 inrange [1, 2] OR 2 inrange [2, 3]	0.458
atleast 2 1's AND atleast 2 2's AND atleast 2 4's	0.542
atleast 4 4's OR atleast 4 o's OR atleast 4 w's	0.493
atleast 5 2's OR atleast 4 4's OR atleast 3 6's	0.535
atmost 4 3's AND atmost 3 4's AND atmost 2 5's	0.544
atmost 4 2's AND atleast 3 4's AND atmost 4 w's	0.546
atmost 5 1's OR atleast 5 3's OR atmost 2 4's	0.504
atmost 3 0's OR atleast 5 3's OR atmost 3 4's	0.544
atmost 3 0's OR atmost 4 1's OR atmost 2 4's	0.472
atmost 2 0's OR atmost 5 1's OR atmost 2 4's	0.504
1 inrange [2, 6] OR 2 inrange [3, 4] OR 3 inrange [6, 7]	0.528
atleast 4 0's AND atleast 3 1's AND atleast 2 2's AND atleast 2 4's	0.52
atleast 4 4's OR atleast 2 5's OR atleast 2 6's OR atleast 4 o's	0.518
atmost 3 2's AND atmost 4 4's AND atmost 3 6's AND atmost 5 w's	0.539
6 inrange [1, 7] OR 8 inrange [2, 4] OR o inrange [2, 3] OR w inrange [6, 7]	0.494
1 inrange [1, 6] OR 5 inrange [1, 2] OR o inrange [3, 6] OR w inrange [4, 6]	0.511

Table 4: The binary choice query set used for 10 over experiments and their associated ground truth (GT) probability of occurrence in the **LCric** train set.

Strongly recommended to know the game of Cricket/aware of the rules.

Description

Help us annotate cricket matches by filling in the events happening per ball in a clip.

Instructions

For each cricket match video, there will be up to 6 deliveries that you will need to label. For each delivery, you will need to report:

1. the **number of runs scored** in the delivery
2. whether or not there **was a wide ball or out ball (or neither)** in that delivery
3. **when in seconds** did the batsman play the delivery?

Apart from this, at the very end there is also last question prompt inquiring whether the clip given is sufficient for answering the given set of questions. Please answer it Yes/No accordingly.

Note: If a ball is wide, the ball subsequent to it will also be considered as a part of the same delivery. Also, please do not consider the wide towards the run tally. For instance, if during the second delivery, a bowler bowls a wide ball, then the batter gets 2 runs on the next ball, check “Wide?” and select “2” for the number of runs.

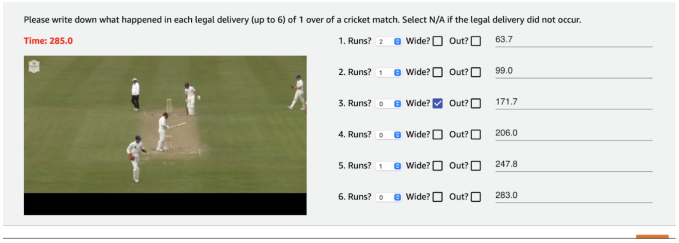
Please find the **timestamp info** for filling out the timestamp related question just above the clip in *red* color.

We request you to watch the full video carefully on a laptop or a computer to precisely answer the questions. The video player has a playback speed option which can be used to alter the playback speed up to 2x.

Please find the detailed instructions below where we cover the process with an example.

We provide an example video with a set of fully labeled annotations. We also walk through how we got each of the annotations labels.

We provide a fully annotated set of labels below for the video above.



Within a document, navigate to File > Page setup to switch between pages (the default format) and pageless (the new format). Changes to this setting are document-specific: everyone who interacts with your document will see it, but changing the setting for one document won’t impact other documents you own.

For annotating the above match the thinking used is as follows:

1. In the *first* delivery, the batsman hits the ball and begins running, resulting in two runs, **so we mark down 2 in the dropdown "Runs?".** We note that no out-balls or wide-balls occurred, so we **do not check either box labelled "Wide?" or "Out?".** We then pause the video at the point when the batsman hit the ball and started running and read the **red timer on the top left of the video** that shows the current time we are paused on, and **mark that time down in seconds [63.7]** in the right-most blank (*do a rough estimate of the time the batsman hit the ball to the best of your ability*).
2. In the *second* delivery, the batsman hits the ball and scores a single run, **so we mark down 1 in the dropdown "Runs?".** We note that no out-balls or wide-balls occurred, and **write down the time [99.0]** that the batsman hit the ball.
3. *In the *third* delivery, the batsman is first thrown a wide ball. So we check off the **wide-ball** label. Since the batsman was thrown a wide ball, we count the subsequent ball as part of the same delivery. In the next ball, the batsman scores 0 runs, **so we mark down 0 in the dropdown "Runs?".** We then **mark the time that the batsman hit the ball [171.7]** (you can mark either when the batsman was thrown the wide ball, or when the batsman hit/missed the subsequent ball).
4. In the *fourth* delivery, the batsman misses and scores no runs, **so we mark down 0 in the dropdown "Runs?".** We then **mark the time that the batsman swung at the ball [206.0].**
5. In the *fifth* delivery, the batsman hits the ball and scores a single run, **so we mark down 1 in the dropdown "Runs?".** We note that no out-balls or wide-balls occurred, and **write down the time [247.8]** that the batsman hit the ball.
6. In the *sixth* delivery, the batsman hits the ball and scores no runs, **so we mark down 0 in the dropdown "Runs?".** We note that no out-balls or wide-balls occurred, and **write down the time [283.0]** that the batsman hit the ball.

Finally, we scroll down and answer the last question. Because we were able to answer all of the given questions using the video, we answer "Yes".

Figure 7: AMT instructions page given to annotators prior to starting the task.

Examples of various different kinds of balls

Below we provide some example snippets of various different kinds of balls that can be seen in the video snippets for our task.

1. **Dot ball (where run scored is 0):**

As can be seen from the clip, the runs scored in the ball is 0. By definition, this can happen either if the batsman does not hit the ball or if he/she hits the ball but is not able to run from one end of the pitch to another.

2. **1 Run Scored:**

As can be seen from the clip, the runs scored in the ball is 1. By definition, if a batsman is able to hit the ball and run from one end of the pitch to another, their team is awarded one run. Similarly, a player can score other possibilities of runs such as 2,3, etc.

3. **4 Run (boundary) Scored:**

As can be seen from the clip, the runs scored in the ball is 4. By definition, it happens if the batsman hits the ball and the ball hits the ground before reaching the stadium boundary.
`

4. **6 Run (boundary) Scored:**

Figure 8: Instructions page for AMT interface for Cricket. Each of the 12 events is described in gif format.

As can be seen from the clip, the runs scored in the ball is 6. By definition, it happens if the batsman hits the ball and the ball reaches the stadium boundary without hitting the ground.

5. **Out ball**

As can be seen from the clip, the ball leads to the player getting out. By definition, an out can happen on multiple accounts.

- Leg Before Wicket: If a ball delivery hits any part of the body and is adjusted to have been hitting the stumps.
- Run Out: A batsman is deemed run out if a member of the fielding team puts down the wicket while the batsman is out of their crease/ground.
- Bowled Out: A batsman is considered bowled out if a delivery strikes their wicket and puts it down.
- Caught: If a ball is hit by the batsman is caught by the opposing team before it hits the ground, it is considered an out ball as well.

For this task of annotation, we request you to consider the ball where an Out occurs as one where runs scored is also 0.

6. **Wide ball**

As can be seen from the clip, the ball is a wide one. By definition, a ball is considered wide if it is bowled too wide to be played by a batsman. Also, a wide ball leads to another ball being played on the same ball number and 1 run also being awarded.

Figure 9: Instructions page for AMT interface for Cricket. Each of the 12 events is described in gif format.

References

Andriluka, M.; Iqbal, U.; Milan, A.; Insafutdinov, E.; Pishchulin, L.; Gall, J.; and Schiele, B. 2017. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. *CoRR*, abs/1710.10000.

Bain, M.; Nagrani, A.; Brown, A.; and Zisserman, A. 2020. Condensed Movies: Story Based Retrieval with Contextual Embeddings. arXiv:2005.04208.

Bengio, Y.; and Lecun, Y. 1997. Convolutional Networks for Images, Speech, and Time-Series.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *International Conference on Machine Learning*, 813–824. PMLR.

Cheng-Yang Fu, M. B., Joon Lee; and Berg, A. C. 2017. Video Highlight Prediction Using Audience Chat Reactions. In *EMNLP*.

Corona, K.; Osterdahl, K.; Collins, R.; and Hoogs, A. 2021. MEVA: A Large-Scale Multiview, Multimodal Video Dataset for Activity Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1060–1068.

Desai, K.; Kaul, G.; Aysola, Z.; and Johnson, J. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

ESPN. 2022a. ESPN Soccer Commentary. https://www.espn.in/football/commentary.

ESPN. 2022b. ESPNCricinfo. www.espncricinfo.com/.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6824–6835.

Feichtenhofer, C. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *Proc. CVPR*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gella, S.; Lewis, M.; and Rohrbach, M. 2018. A Dataset for Telling the Stories of Social Media Videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 968–974.

Google. 2022. Google Cloud Optical Character Recognition. https://cloud.google.com/vision/docs/ocr.

Gupta, A.; Yatskar, M.; Nevatia, R.; and Kembhavi, A. 2021. Visual Semantic Role Labeling for Video Understanding. In *CVPR 2021*.

Huang, Q.; Xiong, Y.; Rao, A.; Wang, J.; and Lin, D. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proc. CVPR*.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732.

Kazemi, V.; and Sullivan, J. 2012. Using Richer Models for Articulated Pose Estimation of Footballers. In *BMVC*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2019. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *Tech Report, arXiv*.

Li, A.; Thotakuri, M.; Ross, D. A.; Carreira, J.; Vostrikov, A.; and Zisserman, A. 2020. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*.

Liang, C.; Jiang, Y.; Cheng, J.; Xu, C.; Luo, X.; Wang, J.; Fu, Y.; Lu, H.; and Ma, J. 2010. Personalized Sports Video Customization for Mobile Devices. In *Proceeding of International Conference on Multimedia Modeling (MMM)*, 614–625.

Liang, C.; Xu, C.; and Lu, H. 2010. Personalized Sports Video Customization Using Content and Context Analysis. In *International Journal of Digital Multimedia Broadcasting (IJDMB)*.

Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.-C.; Lee, J. T.; Mukherjee, S.; Aggarwal, J. K.; Lee, H.; Davis, L.; Swears, E.; Wang, X.; Ji, Q.; Reddy, K.; Shah, M.; Vondrick, C.; Pirsiavash, H.; Ramanan, D.; Yuen, J.; Torralba, A.; Song, B.; Fong, A.; Roy-Chowdhury, A.; and Desai, M. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, 3153–3160.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Pont-Tuset, J.; Uijlings, J.; Changpinyo, S.; Soricut, R.; and Ferrari, V. 2020. Connecting Vision and Language with Localized Narratives. In *ECCV*.

Safdarnejad, S. M.; Liu, X.; Udpa, L.; Andrus, B.; Wood, J.; and Craven, D. 2015. Sports Videos in the Wild (SVW): A Video Dataset for Sports Analysis. In *Proc. International Conference on Automatic Face and Gesture Recognition*. Ljubljana, Slovenia.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), 2556–2565. Melbourne, Australia: Association for Computational Linguistics.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. ECCV*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, abs/1212.0402.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tuyls, K.; Omidshafiei, S.; Muller, P.; Wang, Z.; Connor, J.; Hennes, D.; Graham, I.; Spearman, W.; Waskett, T.; Steel, D.; Luc, P.; Recasens, A.; Galashov, A.; Thornton, G.; Elie, R.; Sprechmann, P.; Moreno, P.; Cao, K.; Garnelo, M.; Dutta, P.; Valko, M.; Heess, N.; Bridgland, A.; Pérolat, J.; De Vylder, B.; Eslami, S. M. A.; Rowland, M.; Jaegle, A.; Munos, R.; Back, T.; Ahamed, R.; Bouton, S.; Beauguerlange, N.; Broshear, J.; Graepel, T.; and Hassabis, D. 2021. Game Plan: What AI can do for Football, and What Football can do for AI.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. NeurIPS*.

Voeikov, R.; Falaleev, N.; and Baikulov, R. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. *CoRR*, abs/2004.09927.

Wu, C.-Y.; and Krahenbuhl, P. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1884–1894.

Wu, C.-Y.; Li, Y.; Mangalam, K.; Fan, H.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022a. MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. *arXiv preprint arXiv:2201.08383*.

Wu, C.-Y.; Li, Y.; Mangalam, K.; Fan, H.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022b. MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. *arXiv preprint arXiv:2201.08383*.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proc. ECCV*.

Xu, C.; Wang, J.; Wan, K.; Li, Y.; and Duan, L. 2006. Live sports event detection based on broadcast video and web-casting text. In *Proceeding of ACM International Conference on Multimedia*, 221–230.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5288–5296.

Zeng, K.-H.; Chen, T.-H.; Niebles, J. C.; and Sun, M. 2016. Title Generation for User Generated Videos. volume 9906. ISBN 978-3-319-46474-9.

Zhang, C.; Gupta, A.; and Zisserman, A. 2021a. Temporal Query Networks for Fine-grained Video Understanding. *arXiv preprint*.

Zhang, C.; Gupta, A.; and Zisserman, A. 2021b. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4486–4496.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*.