

Building Scalable Video Understanding Benchmarks through Sports

Aniket Agarwal^{1*†} Alex Zhang^{2*} Karthik Narasimhan² Igor Gilitschenski³

Vishvak Murahari^{2‡} Yash Kant^{3‡}

¹IIT Roorkee

²Princeton University

³University of Toronto

<https://asap-benchmark.github.io/>

Abstract

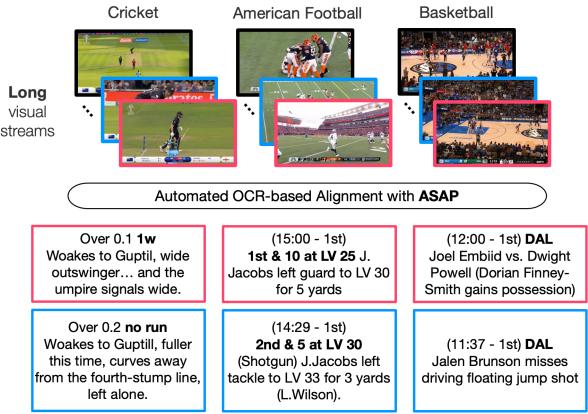
Existing benchmarks for evaluating long video understanding falls short on two critical aspects, either lacking in scale or quality of annotations. These limitations arise from the difficulty in collecting dense annotations for long videos, which often require manually labeling each frame. In this work, we introduce an automated Annotation and Video Stream Alignment Pipeline (abbreviated ASAP). We demonstrate the generality of ASAP by aligning unlabeled videos of four different sports with corresponding freely available dense web annotations (i.e. commentary). We then leverage ASAP’s scalability to create LCric, a large-scale long video understanding benchmark, with over 1000 hours of densely annotated long Cricket videos (with an average sample length of ~ 50 mins) collected at virtually zero annotation cost. We benchmark and analyze state-of-the-art video understanding models on LCric through a large set of compositional multi-choice and regression queries. We establish a human baseline that indicates significant room for new research to explore. Our human studies indicate that ASAP can align videos and annotations with high fidelity, precision, and speed. The dataset along with the code for ASAP and baselines can be accessed here: <https://asap-benchmark.github.io/>.

1. Introduction

Humans learn and master skills (e.g. playing guitar) by associating and reasoning over episodic memories captured over days, months, and years of failed and successful attempts [5]. Thus, building systems capable of understanding and reasoning over very long streams of visual data is a long-standing and crucial problem in Computer Vision.

Long-horizon Video Understanding (LVU) is the problem of reasoning over a long stream of video data, such as understanding the plot of a movie or analyzing the perfor-

Annotation and Video Stream Alignment Pipeline (ASAP) for sports



Annotations from sports websites

Figure 1: **Overview.** We propose the Annotation and Video Stream Alignment Pipeline (ASAP) for sports to align unlabelled sports videos with structured annotations publicly available on the web using an OCR based module. These newly aligned annotations can help compose structured queries that test long-horizon video understanding skills.

mance of a player in a lengthy game. Progress toward LVU has been greatly limited by the lack of densely annotated data. Creating an LVU benchmark requires manually annotating videos frame-by-frame, which is incredibly tedious and hard to scale. This constraint has limited the length of existing densely-annotated video understanding benchmarks (Table 1) from a few seconds [21, 36, 18, 47] to a few minutes [48, 25, 42, 51, 16, 2].

A line of previous works [19, 38, 26] in LVU have used readily available subtitles of TV shows or entire movies as dense annotations. While these videos are sufficiently long, manual annotations are still required to build non-trivial queries to evaluate LVU skills [38, 26], which greatly limits their scale. Another work [42] addresses this problem by extracting supervision from easily accessible YouTube metadata of nearly ~ 30 K movie clips spanning 1 – 3 minutes. However, the annotated clips are short, and the pro-

*Equal Contribution

†Done while interning at Princeton and University of Toronto

‡Equal Advising

posed prediction tasks rely on noisy (and obscure) attributes (e.g. YouTube views, like-to-dislike ratio, and so on).

Sports matches are a rich source of long videos (e.g. a one-day Cricket match lasts nearly 8 hours) and usually have a brief scorecard embedded in the screen (shown in Figure 1) that tracks the state of the match. Most sports matches also have dense annotations from experts available online (sports commentary describing major events in the game, e.g. [12, 11]). However, the annotations or the videos are not helpful individually unless they are aligned with each other.

Therefore, we introduce ASAP, an automated annotation and video stream alignment pipeline, to automatically generate video datasets with dense annotations (i.e natural language commentary, major events in the match) by aligning arbitrarily long sports matches with their commentary freely available on the web [12, 11]). ASAP automatically parses *match state* from the scorecard embedded in sports match videos using an OCR detector [17] and then uses this to automatically align the videos with dense annotations available on the web. To demonstrate the generality of ASAP, we align unlabeled videos of four distinct sports (Cricket, Football/Soccer, Basketball, and American Football) with their corresponding web annotations, with an average of 95% of the annotations being aligned within ± 1 second of their occurrence in the video.

We then leverage ASAP’s scalability to create LCric, a large-scale LVU benchmark with 1008 hours of densely annotated Cricket videos at virtually zero annotation cost, by auto-labeling 131 cricket matches of average length 7.5 hours, containing nearly 475 timestamp recordings (balls per match) on average. To our knowledge, LCric is the first automatically labeled sports video dataset that contains play-by-play annotations that span entire matches. To comprehensively evaluate LVU on LCric, we automatically curate multiple-choice (binary and N-way) and regression queries through simple composition with boolean operations, which require varying lengths of context to answer. These queries are complex and require context aggregation ranging anywhere from 5 minutes to an hour of continuous playtime (video). In the past, such compositional query building has been leveraged in popular vision and language datasets (e.g. CLEVR [22], GQA [20]).

We benchmark two recent state-of-the-art LVU models TQN [49], MemViT [43] on LCric, and find that their performance is significantly worse than human baseline ($\sim 38\%$ drop on query reasoning accuracy when evaluated on long clips containing ~ 50 minutes of playtime). This demonstrates significant room for new research to explore. In summary, we make the following contributions:

- We propose ASAP, an automated and scalable video labeling pipeline for aligning videos of sports matches of four different sports (Cricket, Football, Basketball,

Dataset	Avg clip	# Annot.	# Hours	Autolabel
VidSitu [18]	10s	145K	81	✗
VideoStory [16]	18s	123K	396	✓
MSR-VTT [47]	20s	200K	41	✗
Charades [36]	30s	28K	82	✗
TGIF [21]	30s	126K	86	✓
TVQA [26, 27]	75s	<u>152K</u>	460	✗
VTW [48]	90s	45K	213	✓
MovieClips [2]	120s	30K	1270	✓
LVU [42]	120s	11K	1270	✓
YouCook II [51]	<u>316s</u>	15K	176	✗
ActNet Captions [25]	180s	100K	849	✗
LCric (ours)	2778s	62K	1008	✓

Table 1: Comparison among annotated datasets for benchmarking video description and video understanding methods. LCric has an average clip length of ~ 2800 seconds, which is almost ten times larger than any previous work.

and American Football) with dense web annotations.

- Using ASAP, we create LCric, a large-scale LVU benchmark with 1008 hours of densely annotated Cricket videos with virtually zero annotation cost.
- We benchmark the performance of two recent video understanding models on our dataset, provide ablations, and establish a human baseline on LCric.

Finally, we note that the size of LCric is only limited by access to videos of Cricket matches, and we foresee the dataset being much larger given access to more videos.

2. Related Works

Existing benchmarks for LVU. The paper [42] introduces the large-scale LVU benchmark built on movie clips and metadata publicly available on YouTube. However, the videos only range from 1-3 minutes, and the annotations are limited due to their dependence on YouTube metadata. Other benchmarks for LVU include [31], which collected 29 hours of surveillance footage and bounding box annotations of major events but only have clips of length up to 3 minutes. Similarly, [7] collected 144 hours of surveillance footage by hiring actors to enact predefined scripts but only has clips of length up to 5 minutes. The [28] benchmark collected 430 videos, each 15 minutes long, and dense bounding box information for 80 different atomic actions. Although their videos are relatively long, the annotated videos are only up to 15 minutes long, which is shorter than our annotated videos which are up to 45 minutes long, and are generated with no additional cost. Additionally, [6] collected the LoL dataset comprising 230 clips from the League of Legends video game, with each clip ranging from 30 to 50 minutes. However, they collected video highlight annotations based on very noisy and unreliable audience chat statistics. Video games also tend to have easy visual cues before major highlights that incentivize models to learn spurious correlations. In contrast, our tasks, by construction, force models to reason over a long horizon of events in a match.

Collecting dense annotations for videos. Annotating video datasets is extremely expensive. Prior works [47, 18, 36] have collected expensive annotations through Amazon Mechanical Turks (AMT) to label their clips with an associated text description, which greatly limits their scale (Table 1). Another line of work bootstraps from pre-existing annotations to generate new annotations. Other prior works bootstrap from pre-existing annotations to generate new annotations. For example, [2] use existing captions on YouTube and IMDb metadata to label 30000 movie clips, but assume these labels span the entirety of their clips. Similarly, [48] take user-generated titles as labels for 18100 user-generated clips, but again assume that these labels span the entirety of their clips. While [16] temporally align sentences from paragraph captions to social media videos to form annotated clips. In contrast, our dataset is densely annotated by temporally aligning publicly available sports annotations, which offer more structure than text descriptions and are therefore hierarchically composable, enabling the creation of queries that require large but dense context. Although [30, 29, 46] also align sports videos to online commentary information, they use heuristic methods that are not as accurate and do not scale well for generating longer and more video matches.

Video datasets based on sports. Recent interest in using computer vision to drive sports analytics [39] suggests the importance of a dense annotation pipeline for sports videos. Current methods for producing sports datasets involve some form of manual annotations like in [34], where they manually label 4100 sports clips based on the given action. Several works [41, 1, 24] have used automatically generated, densely labeled pose annotations for sports videos but are not easily scalable because they run computationally expensive, frame-level models to generate their annotations. Larger datasets such as [37, 23] exist but primarily focus on action recognition over a single clip, rather than a full sports video. Our dataset focuses on producing play-by-play annotations spanning entire sports match. Our general annotation pipeline can also be easily extended to the creation of video datasets for other sports.

Video understanding models. Processing long videos is challenging, as it requires aggregating context over long horizons with limited computational and memory budgets. In SlowFast networks [15], they use a dual pathway operating at a low and high frame rate to enable the aggregation of context over longer horizons while capturing low-level visual attributes. Meanwhile, [14] introduce a simple technique for progressive architecture expansion (along axes such as temporal, depth, width, etc.), inspired by feature selection in machine learning to achieve efficient models. Taking advantage of the implicit nature of transformers to handle long-range data [4] proposes to adapt the standard transformer architecture for videos by enabling spa-

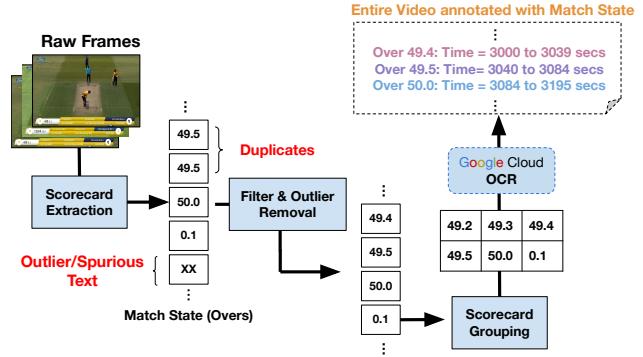


Figure 2: **Match State Extraction in Cricket.** **Left:** To extract the match state we begin by detecting scorecard bounding box across all video frames, and cropping its content out. **Middle:** We filter duplicates and outliers from extracted scorecards using L1 distance metric. **Right:** We batch the filtered patches in one image for the OCR call, and generate *match states* aligned in time with the full video.

tiotemporal feature learning directly from a sequence of frame-level patches. In MeMViT [43], they introduce a memory-augmented multi-scale vision Transformer, greatly improves temporal support with minimal memory overhead, and achieves state-of-the-art performance on a variety of video understanding benchmarks. While the trend shows the model’s capacity to handle longer and longer video clips more efficiently, the absence of a truly long-horizon dataset inhibits a fair comparison between these baselines and also inhibits the model’s transferability to real-world video understanding tasks.

Automated annotation pipelines. Automating annotation pipelines, even partially, is critical to developing large-scale datasets. For example, the SBU dataset [32] for image-text retrieval pruned and paired Flickr queries with a set of images, the Conceptual Captions dataset [35] for image captioning leveraged the “Alt-text” HTML attribute in web images, and the RedCaps dataset [9] harvested 12 million image-text pairs from curated subreddits. Meanwhile, [33] partially automate their annotation pipeline and collect multi-modal image annotations by asking annotators to describe an image through audio while simultaneously hovering their mouse over the region they are describing. We hope that our fully automated pipeline, ASAP, will help create long and densely annotated video datasets at an unprecedented scale.

3. ASAP: Annotation and Video Stream Alignment Pipeline for Sports Matches

Sports matches provide an abundant source of long videos, along with a rich source of corresponding play-by-play annotations (i.e. expert commentary of major events in the

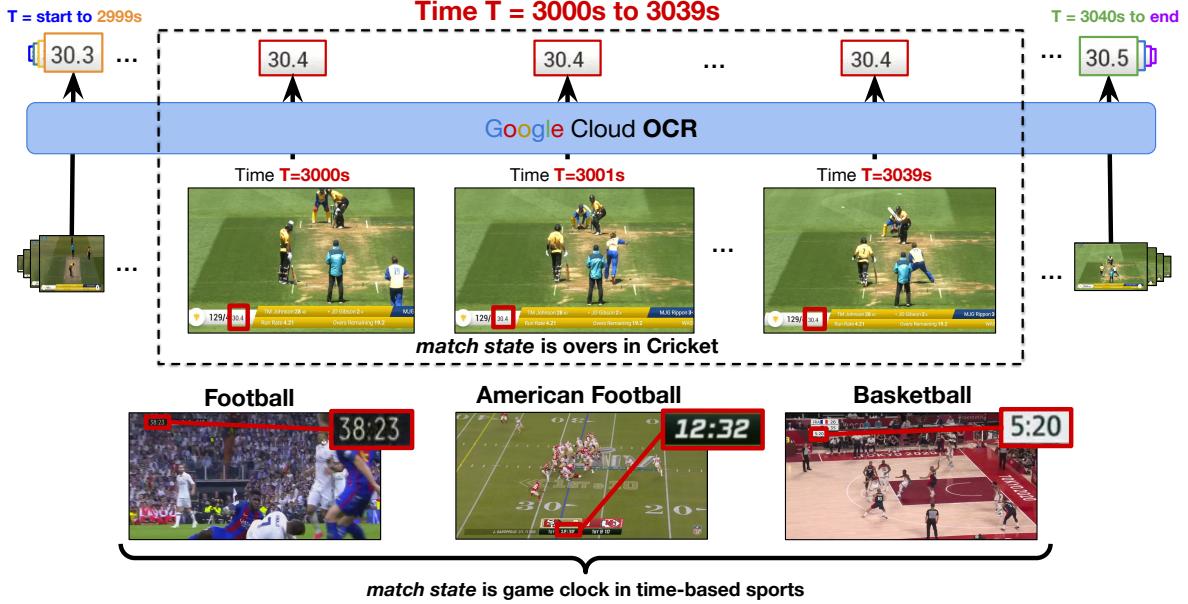


Figure 3: **Match States used by ASAP for different sports.** ASAP finds and stores the sequence of frames corresponding to each possible *match state*, which is necessary to temporally align web annotation data to its exact occurrence in each video. The process is as follows: **a) Top Row:** For each frame in a Cricket match video, ASAP locates and extracts the *match state* (30.4 in the above example) from the scorecard of each frame using an OCR module. Next, ASAP determines the sequence of frames (for *match state* = 30.4 in the example above, this would be the frames from $T=3000$ seconds to $T=3039$ seconds) that each *match state* corresponds to. **b) Bottom Row:** For other sports like American Football, Football, and Basketball, ASAP uses the onscreen timer as *match state*. More generally, any temporal marker can be used as *match state*.

match) easily accessible on the web [12, 11]. These play-by-play annotations are, however, not useful standalone as they are not aligned with the video of the match. We, therefore, introduce ASAP, a fully automated annotation pipeline for automatically aligning sports videos with their corresponding play-by-play annotations available on the web. ASAP automatically parses *match state* from the scorecard embedded in sports match video through an OCR detector (as seen in Figure 2) and then uses these states to align the video with dense annotations available on the web. Thus, ASAP enables the creation of long video datasets with unprecedented scale at virtually zero annotation cost. We first give a brief overview of ASAP and then describe the different stages of ASAP in more detail.

ASAP Overview. ASAP first extracts the scorecards embedded in a sports video and automatically parses the state of the match. For instance, Basketball and Football games have a running game clock, whereas, Cricket matches have information about the *ball* being played (Section 4.1). We crop out the scorecard from each video frame in the match and extract the *match state* contained within it using an Optical Character Recognition (OCR) system (shown in Figure 3). We then use the *match state* to get the corresponding commentary from the web annotations (Sec. 3.2) and assign one of the discrete events for that sport (Refer to Ap-

pendix A.2 for a list of events). This multi-stage approach allows us to align every video frame in the match to the corresponding commentary and discrete event.

3.1. Stage 1: Match State Extraction

Extracting Match State. In a sports video containing N video frames denoted as $[f_1, \dots, f_N]$, we start by detecting a bounding box that encapsulates the *match state* (scorecard) information. For this, we sample a few frames uniformly throughout the video and run *Google OCR* on them. Next, we determine the bounding box where text changes gradually across frames (i.e. the bounding box containing the scorecard). Once these bounds are detected, we crop the scorecard contained within them across all the frames. We show example cropped scorecards in Figure 3.

Match State in Cricket (Overs). For Cricket, we represent the match state by the *ball* that is currently being delivered. For example, in the top row of Figure 3, we extract "30.4" at frame f_t , which reveals that 4th ball of 30th over is being played in this frame. Assuming this ball (event) lasted for n frames, we detect that at frame f_{t+n} the match state changes to "30.5", which means the event that took place on the 4th ball of 30th over lasted across frames $[f_t, f_{t+n}]$. In this manner, we can label all frames of the video sequence with a corresponding match state, and also club the consec-

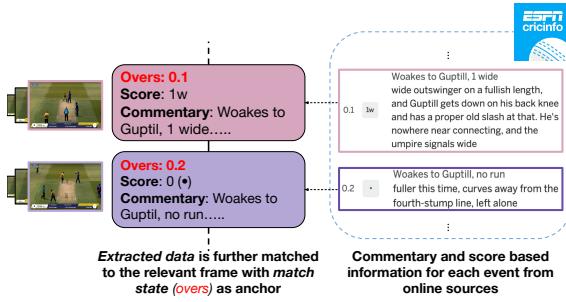


Figure 4: **ASAP Stage 2 for Cricket matches.** ASAP scrapes dense play-by-play annotations from the web and parses them into discrete events. ASAP then uses the extracted *match state* from Stage 1 to align the annotations with the video of the match.

utive frames across which the same match state persists.

Match State in American Football, Football, and Basketball (Timer). For remaining sports, we represent the match state by the onscreen timer usually displaying minutes and seconds passed. For example, in the bottom row of Figure 3, we show the time extracted across different sports. As we use videos with a high frame rate of 30 FPS, the same match state persists across multiple consecutive frames and we, therefore, club these frames together like in Cricket.

We find that locating the scorecard across frames, as well as extracting *match state* from it is non-trivial – due to the noisy and dynamic nature of the scorecard. We list some key challenges we encountered (and resolved) below.

Occlusion, and changing attributes. We found that scorecards can get occluded by advertisements, move to a different position on the screen, or change their attributes (format, shape, color) during gameplay. To address issues, we use a *reference scorecard*, which is a template image containing an unoccluded crop of the scorecard. We compare this reference image against all cropped scorecards in the video using L1 distance, and remove spurious that lie far away from the reference. These are the frames 1) being occluded by advertisements or other miscellaneous patches, 2) having varying attributes of the scorecard (format, shape, color, etc.). We put frames rejected by this filtering in Appendix A.

OCR calls for every frame in a video can be expensive. Sports matches last for many hours and passing all frames of a 30 FPS video through the paid *Google OCR* becomes prohibitively expensive. To reduce the costs many-fold, we stack multiple scorecards into a single image and get multiple annotations with a single *OCR* call. We also skip frames that contain the same *match state* as an older frame by detecting changes in the scorecard between consecutive frames using a *L1* distance metric.

3.2. Stage 2: Aligning dense annotations with videos

Dense play-by-play annotations (i.e expert commentary, major events, etc.) are often easily available on the web [12, 11]. Since these play-by-play annotations are indexed by the *match state* (precise *play time* or *ball* information), we map annotations based on their *match state* to their exact timestamps (frames) found in the first stage of ASAP.

In addition to aligning the annotations with the video, ASAP also processes the sequence of play-by-play annotations into a sequence of discrete events, which we refer to as an *event chain* (Figure 4). While some sports (Cricket and Football) already contain discrete events (e.g. ‘foul’, ‘wicket’, ‘boundary’ etc.) in their annotations, for other sports (American Football and Basketball), we use string-matching to parse the commentary and assign each play to a fixed event that we define (e.g. ‘incomplete pass’). These extracted event chains can then be used as ground truth for evaluating LVU models. Models can be queried on different segments of the event chain of varying lengths – to test both short and long-horizon reasoning. We discuss the use of event chains for evaluation in Section 4.3.

3.3. ASAP is faithful and fast

How accurate is ASAP? To verify ASAP’s ability to align dense annotations on the web with videos of sports matches, we conduct a study with human annotators on Amazon Mechanical Turk (AMT). We randomly sample clips from sports matches corresponding to 6 contiguous events in the event chain generated by ASAP. For all the generated clips, we ask human annotators to provide timestamps for all 6 events and then check whether the provided timestamps belong to the intervals generated for those events by ASAP. We plot the resulting accuracy of the timestamps in Figure 5 and find that ASAP is highly accurate, with an average accuracy of 95.3% across four very different sports, each differing in visual attributes, number of events, and length of plays. The drop in accuracy for American football annotations can be explained by the inconsistency in the timestamps provided by ESPN. For regular plays, the timestamp indicates when the play begins; however, for ‘touchdowns’, the timestamp indicates when the team scored and not when the play begins. Additionally, penalties may affect the game clock, which we use to align our annotations, which sometimes leads to slight alignment issues for ASAP. We present more details in Appendix A.

How fast is ASAP? ASAP generates annotations with very high speed and requires just 10 minutes to align and process around 7 hours of video at 30FPS on a single machine.

4. Generating the LCric dataset with ASAP

In this section, we describe how we leverage ASAP to build a long video understanding (LVU) dataset from Cricket

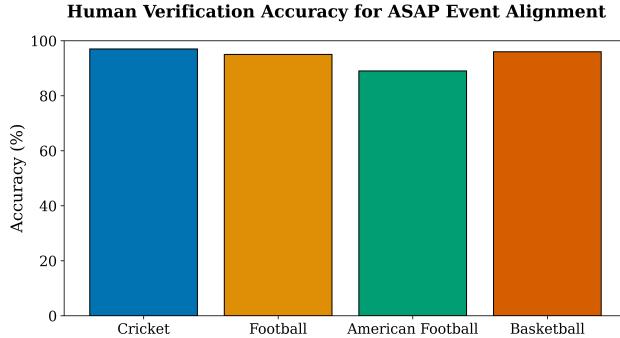


Figure 5: **Accuracy of ASAP.** Our human studies indicate ASAP alignment to be highly accurate, with an average of 95% of the annotations being correctly aligned to the corresponding video events (± 1 second) across 4 different sports.

videos. The ASAP pipeline takes in videos of Cricket matches along with play-by-play commentary annotations from the web¹. ASAP then aligns every video frame with the corresponding commentary annotation. Using the frame-aligned annotations produced by ASAP, we describe a scalable approach for generating structured and compositional queries in Section 4.3 to evaluate LVU. Our LCric dataset is a collection of Cricket videos with play-by-play annotations and a set of auto-generated queries.

4.1. Introduction to Cricket

We present a brief introduction to the sport of Cricket before discussing LCric. Cricket is played by two teams of 11 players each that alternate between *batting* and *fielding* throughout the game. The batting team aims to score *runs* by hitting a ball bowled by the fielding team out of the playing field. Meanwhile, the fielding team aims to prevent the batting team from scoring runs and dismiss all players in the batting team by taking their *wickets*. Each exchange where the fielding team bowls a valid ball and the batting team attempts to hit the ball to score runs is called a *ball* (or *delivery*) and a sequence of 6 *balls* is called an *over*. Each *ball* is an *atomic event* and there are 12 distinct possible events per ball, listed below:

- The batting team scores n runs, where $n \in \{0, \dots, 9\}$
- A wicket is taken and the current batsman is dismissed.
- A wide (invalid) ball is bowled, giving the batting team an extra run and another ball.

We present a detailed discussion of different phases of Cricket in Appendix B.

4.2. LCric: Overview

Using ASAP’s scalability, we create LCric, a large-scale LVU benchmark with 1008 hours of densely annotated

Cricket videos at virtually zero annotation cost, by auto-labeling 131 cricket matches of average length ~ 7.5 hours, containing nearly 475 timestamp recordings (balls per match) on average. ASAP automatically labels all the balls in a match with 1 of 12 events (Section 4.1) to generate a sequence of events (i.e. *event chain* shown in Figure 4) for a cricket match. We then generate annotated video clips by segmenting the videos along with the aligned event chain into a contiguous sequence of 10-over (~ 50 minutes) clips.

4.3. LCric: Testing LVU via compositional queries

Task Definition and LVU Capabilities. In an LVU task, the evaluated model is given a very long video clip, from a sports match of nearly 50 mins in our case, and it is tasked to answer a question (query) about it, e.g. “how many times did a wide ball occur in this video?”. An LVU system needs to possess two types of skills: a) the ability to reliably detect local (short-term) events – e.g., classifying an atomic event in Cricket (say wide, wicket, or run), and b) the ability to aggregate information across these local events given a task (which we refer to as a query) – e.g., counting the total number of runs scored by the batting team in an arbitrarily long video. We evaluate these LVU skills on LCric, by automatically generating binary, multiple-choice, and regression queries, and evaluating them on long video segments.

Min-Max occurrence query. To test a model’s ability to detect and remember events, we construct queries such as “for a given video, did a wide ball (an event) occur between 3 and 5 times inclusive?”. We generate these queries by sampling an atomic event from the set of all possible events, and then sampling two numbers, o_{min} and o_{max} , to denote the minimum and the maximum number of occurrences needed for this query to be *true*.

Binary queries by chaining occurrence queries. To increase query diversity and complexity, we sample n_{chain} different min-max occurrence queries and combine them using [and]/[or] operators. For example, for a given video spanning 10 overs (~ 50 mins), “did a wide ball occur between 3 to 5 times [and] did a ball with 2 runs scored occur 1 to 3 times?”. All **binary** queries in LCric are formed by chaining 1-5 different min-max occurrence queries.

Multiple-choice queries by counting occurrences. We expand upon the binary occurrence queries by generating multiple-choice occurrence queries, which ask models to directly predict the number of occurrences rather than predicting membership in a range. An example of such a query is – given a video, how many times did a wide ball occur after a ball with 4 runs? We note that these events are sequential, but not necessarily contiguous. As most non-trivial multiple-choice events in LCric occur between 0-9 times in a given clip, we use $\{0, \dots, 9\}$ as our answer choices.

Filtering unbalanced queries. We can compose many LVU multiple-choice and binary queries using the above

¹<https://www.espnccricinfo.com/>

formulation, however, not all queries are balanced. Due to the rarity of certain events occurring in Cricket, some queries are far easier to guess correctly than others. For example, in a 45 minutes clip (spanning ~ 10 overs), the query – “did a ball with 2 runs occur between 0 to 10 times” is true with a probability of 87%. We filter such queries based on the probability of their occurrence in training matches and ensure the average probability of occurrence of the selected queries to be between 0.45 – 0.55 to avoid bias.

Regression query for counting runs. Lastly, we also experiment with a single regression query that asks the model to predict the number of runs scored as a regression output for a given video sequence.

4.4. LCric: Statistics and Dataset Splits

Statistics. LCric currently includes 1008 hours of cricket match videos across 131 unique matches (average length of 7.5 hours), along with 61957 ball-by-ball annotations. All the videos are preprocessed at a resolution of 360p and we provide links to the source videos of higher resolution.

Dataset splits To effectively test generalization, we split all the matches in LCric into train, validation, and test splits and ensure a 3:1:1 ratio of the number of hours in each split. Due to a limited compute availability, we present ablations on a subset of LCric and refer to this as LCric-Mini, which has around 420 hours of labeled Cricket matches, and therefore enables us to train ablations experiments in a shorter duration (2-3 days per experiment). We generate splits for LCric-Mini identically to LCric.

5. Experiments on LCric

In this section, we describe our experiments benchmarking state-of-the-art LVU models on LCric. We also provide a human baseline on LCric and demonstrate significant room for improvement. Finally, we analyze the performance of our baselines and present key insights to spur future modeling improvements on LCric.

5.1. Experimental Setup

Preprocessing. We follow past work in LVU [50] by sampling videos at a lower FPS to make training over long videos feasible. We process our longest clips (containing 10 overs of the match) at 0.1 FPS, and process clips of 2-8 overs at 0.5 FPS respectively. We remove the scorecard from all frames to prevent annotation leakage and process frames at a resolution of 128 x 128.

Evaluation Metrics. We compute following evaluation metrics 1) Classification accuracy for binary (**Binary Accuracy**) and multi-choice (**Multiple Accuracy**) queries 2) Average L1 norm for regression queries (**Regression L1 Norm**).

5.2. Baseline Models and Training Scheme

Previous works [13, 15, 14] in LVU use pretrained CNNs [3] and Transformers [40] paired with explicit memory modules for modeling long contexts. However, none of these methods can scale to video clips longer than a few minutes. Since our query set requires reasoning over contexts ranging up to an hour, we choose two state-of-the-art video understanding models, Temporal Query Network (TQN) [50] and MeMViT [44] as benchmarks – given their effective caching mechanisms for processing long videos. **Temporal Query Network (TQN)** [50] TQN uses stochastic memory banks to efficiently model long horizon videos. They also introduce a transformer-based multi-query head to generate responses for multiple queries with a single pass through the network. TQN uses S3D [45] for the visual backbone network and processes non-overlapping contiguous sequences of 8 frames with a temporal stride of 1. Please note that the term *query* used in TQN is a modeling component and is not the same as our use of query in Section 4.

Memory-augmented Multiscale Vision Transformer (MeMViT) [44] MeMViT applies a memory caching strategy by processing videos in an online fashion, allowing the model to efficiently store context to reason over a long horizon. MeMViT builds upon ViT [10] by using a novel pooling method and a dynamic patch resolution approach to reduce computational costs while processing long clips. We adapt MeMViT to handle our multi-query setting by leveraging the multi-query head introduced in TQN.

Training Schemes. We employ two different training schemes – 1) **Homogeneous training** where we train different models for the three different types of queries (binary, multi-choice, and regression) and 2) **Mixed training** where we train a single model for all three types of queries.

5.3. Human Baseline

To quantify the room for modeling improvements on LCric, we measure the accuracy of human annotators through AMT [8]. We provide annotators with video clips from LCric and ask them to predict the sequence of ball-by-ball events (4.1). To compute human performance on our queries, we assume that given an event chain, humans can answer these queries by applying logical operators without mistakes. We detail our AMT setup in Appendix B.4.

5.4. Key Results

Performance of both TQN and MeMViT degrades rapidly and approaches random for very long clips To understand the impact of length of the videos on task performance, we train different baseline models for clips with over-lengths ranging from 2 overs (~ 10 minutes) to 10 overs (~ 50 minutes). Figure 6 shows that performance rapidly decreases with increasing clip length and ap-

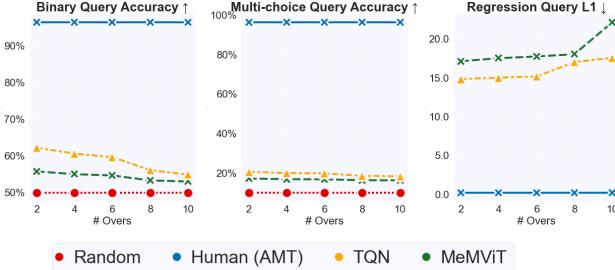


Figure 6: **Performance drops as video length (# of overs) increase.** We plot the accuracy achieved by our baseline models for binary, multiple-choice, and regression queries across clips varying from 2-10 overs (length). LVU models severely degrade in performance with longer time horizons and perform much worse than humans.

proaches the random baseline for binary and multi-choice queries. This result, in addition to the strong human baseline, shows significant room for modeling improvements.

Models need to localize and detect events accurately to perform well on LCric To understand the importance of localizing and detecting events for LVU models, we first train a TQN ‘event classifier’ model to predict 1 of 12 events (Section 4.1) in a video clip. The model is trained on ground truth annotations from LCric-Mini and has a fairly high test accuracy of 84.79%. We then divide the clips into a contiguous sequence of ‘event segments’ by either using ground truth segmentations from LCric-Mini (labeled ‘Ground Truth’ in Table 3) or by uniformly dividing the clips into 60 contiguous segments (labeled ‘Uniform’ in Table 3), as each 10-over clip contains 60 events. Finally, we leverage the learned ‘event classifier’ model to generate event chains by sequentially predicting events on the ‘event segments’ and evaluating different queries on these event chains. We report performance in Table 3 and make two observations – **1)** Access to ground truth event segments leads to $\sim 20\%$ improvement over uniformly generated segments on binary queries and therefore shows the importance of event localization. **2)** While access to ground truth event segments leads to better performance, as it aids the event classifier in making more accurate predictions, the performance is still $\sim 16\%$ worse than the human baseline on binary queries. Therefore, even with perfect event localization, models need performant event detection capabilities.

TQN performs better than MeMViT on all query types Table 2 shows that TQN performs much better than MeMViT across different query types and training schemes (+2% on binary and multi-choice queries). While we expected MeMViT to perform better due to the clever caching mechanism in its architecture, we think adapting MeMViT with the TQN multi-query head led to the performance drop. We hope future work can better integrate the caching mechanism of MeMViT into the multi-query setting as it is

#	Model	Training	Binary Accuracy ↑	Multiple Accuracy ↑	Regression L1 Norm ↑
1	TQN	Mixed	57.68%	19.05%	17.21
2	MeMViT	Mixed	54.31%	16.71%	21.79
3	TQN	Homogeneous	60.74%	20.19%	10.63
4	MeMViT	Homogeneous	56.53 %	17.79%	11.95
5	Human		96.34 %	96.29%	0.215

Table 2: **Baseline performance on the full LCric split with 10 over clips.** TQN outperforms MeMViT in different training schemes across different query types. We find that training models under the Homogeneous training scheme greatly improves performance, especially for the regression query.

#	Model	Clip Segments	Binary Accuracy ↑	Multiple Accuracy ↑	Regression L1 Norm ↑
1	TQN	Ground Truth	80.34%	36.32%	6.89
2	TQN	Uniform	60.29%	15.18%	19.31
3	Human		96.34 %	96.29%	0.215

Table 3: **Both localization and detection of events are important.** Using ground truth clip segments (aligned by ASAP) for event prediction leads to significant performance gain (Rows 1 vs 2). Improving event detection within clips can further improve performance (Rows 2 vs 3).

infeasible to train one model for every new query.

Homogeneous training scheme leads to superior performance than Mixed training Our results in Table 2 indicate that training different models for different query types leads to much better performance than training a single model for all queries. We also observe that performance on the regression query improves by ~ 7 points and we posit that the model needs the extra representational capacity to answer the regression queries with high precision.

6. Conclusion

In this work, we introduce ASAP, a fully automated annotation and video stream alignment pipeline for sports matches. ASAP automatically aligns unlabeled videos of sports matches with corresponding dense annotations (*i.e.* commentary) freely available on the web. We demonstrate the generality of ASAP by aligning unlabeled matches of four very different sports with their corresponding annotations on the web. ASAP is highly accurate (as judged by human annotators), and is robust to varying visual attributes, number of events, and length of plays. We then demonstrate ASAP’s potential to generate large-scale video datasets with *no additional annotation cost* by generating LCric, a large-scale long video understanding benchmark with over 1000

hours of densely annotated long Cricket videos (having an average sample length of \sim 50 minutes). We extensively benchmark state-of-the-art LVU models and establish a human baseline on LCric. Our strong human baseline, coupled with the poor performance of state-of-the-art models, validates LCric as an effective benchmark for the next generation of LVU models. We hope that future work extends, improves, and leverages ASAP to generate annotated video datasets at an unprecedented scale and cost efficiency.

References

- [1] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *CoRR*, abs/1710.10000, 2017. 3
- [2] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 2020. 1, 2, 3
- [3] Y. Bengio and Yann Lecun. Convolutional networks for images, speech, and time-series. 11 1997. 7
- [4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 3
- [5] John H. Byrne. *Learning and Memory: A Comprehensive Reference*. Elsevier, Oxford UK, 2008. 1
- [6] Mohit Bansal, Cheng-Yang Fu, Joon Lee and Alexander C. Berg. Video highlight prediction using audience chat reactions. In *EMNLP*, 2017. 2
- [7] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. 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)*, pages 1060–1068, January 2021. 2
- [8] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Anol Bhattacherjee and Brian Fitzgerald, editors, *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 7
- [9] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 3
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7
- [11] ESPN. Espn soccer commentary, 2022. <https://www.espn.in/football/commentary>. 2, 4, 5
- [12] ESPN. Espncricinfo, 2022. www.espncricinfo.com/. 2, 4, 5
- [13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 7, 13
- [14] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proc. CVPR*, 2020. 3, 7
- [15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3, 7
- [16] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 968–974, 2018. 1, 2, 3
- [17] Google. Google cloud optical character recognition, 2022. <https://cloud.google.com/vision/docs/ocr>. 2
- [18] Arka Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *CVPR 2021*, 2021. 1, 2, 3
- [19] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1
- [20] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 2
- [21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proc. CVPR*, 2017. 1, 2
- [22] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2
- [23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 3
- [24] Vahid Kazemi and Josephine Sullivan. Using richer models for articulated pose estimation of footballers. In *BMVC*, 2012. 3
- [25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2
- [26] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 1, 2
- [27] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019. 2
- [28] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 2
- [29] Chao Liang, Yu Jiang, Jian Cheng, Changsheng Xu, Xiaowei Luo, Jinqiao Wang, Yu Fu, Hanqing Lu, and Jian Ma. Personalized sports video customization for mobile devices. In *Proceeding of International Conference on Multimedia Modeling (MMM)*, pages 614–625, 2010. 3
- [30] Chao Liang, Changsheng Xu, and Hanqing Lu. Personalized sports video customization using content and context analysis. In *International Journal of Digital Multimedia Broadcasting (IJDMB)*, 2010. 3

[31] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, 2011. 2

[32] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. 3

[33] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 3

[34] Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udupa, Brooks Andrus, John Wood, and Dean Craven. Sports videos in the wild (svw): A video dataset for sports analysis. In *Proc. International Conference on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia, May 2015. 3

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 3

[36] Gunnar A Sigurdsson, Gü̈l Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. ECCV*, 2016. 1, 2, 3

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 3

[38] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[39] Karl Tuyls, Shayegan Omidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steel, Pauline Luc, Adria Recasens, Alexandre Galashov, Gregory Thornton, Romuald Elie, Pablo Sprechmann, Pol Moreno, Kris Cao, Marta Garnelo, Praneet Dutta, Michal Valko, Nicolas Heess, Alex Bridgland, Julien Pérolat, Bart De Vylder, S. M. Ali Eslami, Mark Rowland, Andrew Jaegle, Remi Munos, Trevor Back, Razia Ahmed, Simon Bouton, Nathalie Beauguerlange, Jackson Broshear, Thore Graepel, and Demis Hassabis. Game plan: What ai can do for football, and what football can do for ai, May 2021. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017. 7

[41] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. Ttnet: Real-time temporal and spatial video analysis of table tennis. *CoRR*, abs/2004.09927, 2020. 3

[42] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 1, 2

[43] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. *arXiv preprint arXiv:2201.08383*, 2022. 2, 3, 13

[44] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. *arXiv preprint arXiv:2201.08383*, 2022. 7

[45] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proc. ECCV*, 2018. 7

[46] C.S. Xu, J. Wang, K. Wan, Y. Li, and L. Duan. Live sports event detection based on broadcast video and web-casting text. In *Proceeding of ACM International Conference on Multimedia*, page 221–230, 2006. 3

[47] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 1, 2, 3

[48] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Title generation for user generated videos. volume 9906, 10 2016. 1, 2, 3

[49] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. *arXiv preprint*, 2021. 2, 13

[50] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 7

[51] Luwei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, 2018. 1, 2

A. ASAP

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 5 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.1. Qualitative Video

We provide qualitative samples under ‘qualitative_video_asap’ directory in supplementary material by attaching some cricket match clips auto labelled using our ASAP pipeline.

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 across YouTube. 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.

B. LCric

B.1. Primer on Cricket [Video]

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. A brief overview video for the game explaining the game can be found in [here](#).



Figure 7: Some examples of frames rejected by ASAP. In all of these frames the scorecard information either get obstructed by screen overlays or shifted from their usual position.

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 a 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 [49] and MeMViT [43] 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 [13], which is the base model used to create MeMViT.

Attached Code We have also attached the codebase with the supplementary zip and can be found in the ‘code’ repo. The codebase contains implementations for both our ASAP pipelines and our baseline implementations.

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 from different combinations of queries.

Binary Query 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 4.

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 11.

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 8.

Algorithm 1: Query Set Generation

```

1 # The set of atomic events: [0,1,2,...,9,W,w]
2 Set of atomic events:  $A_e$ 
3 # The number of queries for the query set
4 Size of the query set:  $n_q$ 
5 query.set = []
6 for i in range( $n_q$ ) do
7     # Step A: getting raw operators and
    # combinator's choice
8     num_joins ~ [1,5]
9     # total length for operators set being sampled
10    for determining the query length
11    ops = random.choices([atleast(), atmost(),
12                          inrange()], num_joins) # sampling list of
13                          operators
14    combine_op = random.choices([and, or], 1) # sampling the combination operator
15    # Step B: instantiating a query for query set
    # for matching
16    query = []
17    for op in ops do
18        # specify lower bound for atleast/inrange
        ops
19        occ_min ~ [1,10]
20        # specify upper bound bound for
        atleast/inrange ops
21        occ_max ~ [occ_min,10]
22        # sample atomic events in query
23        atomic_event ~ [ $A_e$ ]
24        # Using the above variables for defining
        # an occurrence pattern for atomic_event
25        instanced_op = op(occ_min, occ_max,
                           atomic_event) query.append(instanced_op)
26    final_query = join_op(query)
27    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 9, 10.

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 8 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 ~ 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.

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 0's	0.452
atleast 4 4's AND atmost 3 5's	0.456
atleast 4 2's OR atleast 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 0 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 atleast 4 1's OR atmost 2 4's	0.472
atmost 2 0's OR atleast 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 0 inrange [2, 3] OR w inrange [6, 7]	0.494
1 inrange [1, 6] OR 5 inrange [1, 2] OR 0 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.

Please write down what happened in each legal delivery (up to 6) of 1 over of a cricket match. Select N/A if the legal delivery did not occur.

Time: 285.0

1. Runs? <input checked="" type="checkbox"/> 2 <input type="checkbox"/> Wide? <input type="checkbox"/> Out? <input type="checkbox"/> 63.7
2. Runs? <input type="checkbox"/> 1 <input checked="" type="checkbox"/> Wide? <input type="checkbox"/> Out? <input type="checkbox"/> 99.0
3. Runs? <input type="checkbox"/> 0 <input checked="" type="checkbox"/> Wide? <input checked="" type="checkbox"/> Out? <input type="checkbox"/> 171.7
4. Runs? <input type="checkbox"/> 0 <input checked="" type="checkbox"/> Wide? <input type="checkbox"/> Out? <input type="checkbox"/> 206.0
5. Runs? <input type="checkbox"/> 1 <input checked="" type="checkbox"/> Wide? <input type="checkbox"/> Out? <input type="checkbox"/> 247.8
6. Runs? <input type="checkbox"/> 0 <input checked="" type="checkbox"/> Wide? <input type="checkbox"/> Out? <input type="checkbox"/> 283.0



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 8: 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 9: 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.

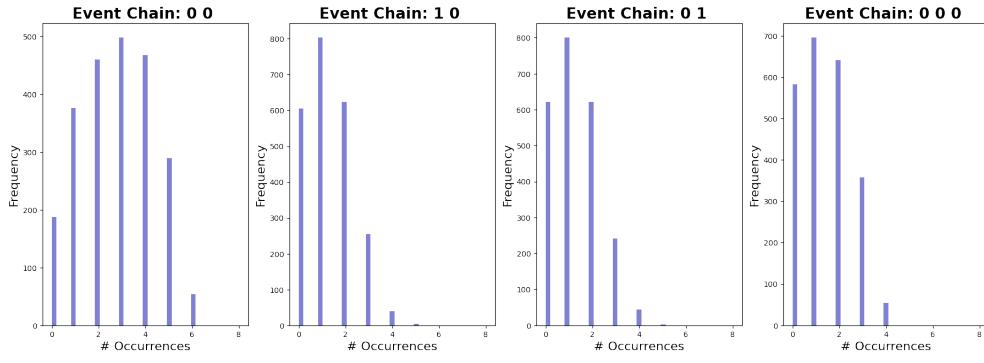
6. **Wide ball**



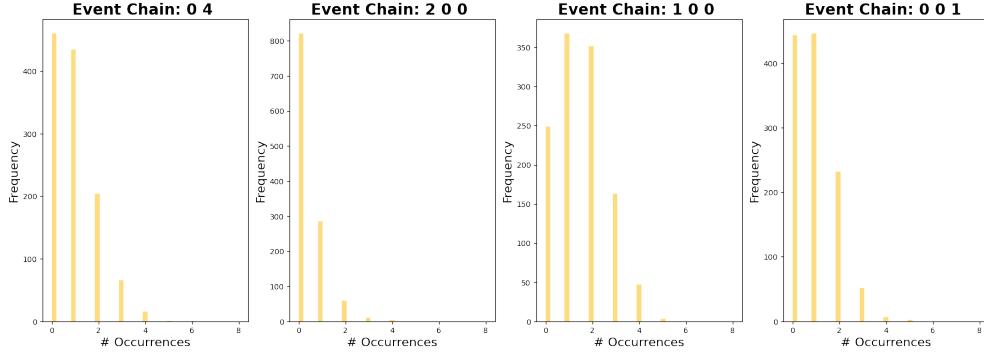
As can be seen from the clip, the ball is a wide one. By definition, a ball is considered wide if it is bowled too wide to be played by a batsman. Also, a wide ball leads to another ball being played on the same ball number and 1 run also being awarded.

Figure 10: Instructions page for AMT interface for Cricket. Each of the 12 events is described in gif format.

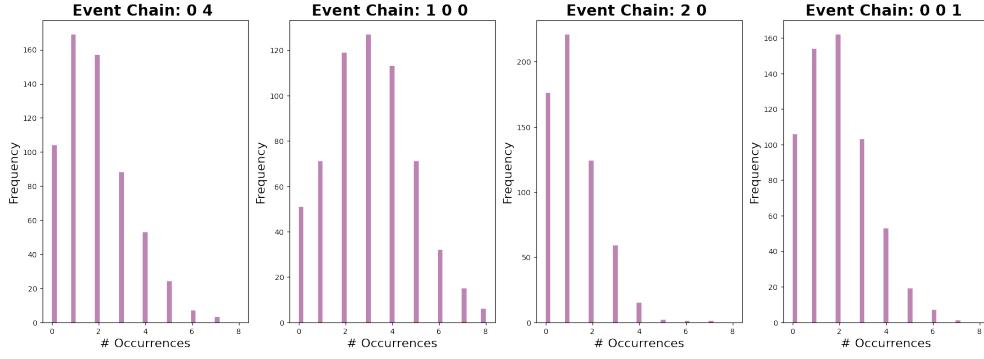
Frequencies of Events on 2-over Clips



Frequencies of Events on 4-over Clips



Frequencies of Events on 8-over Clips



Frequencies of Events on 10-over Clips

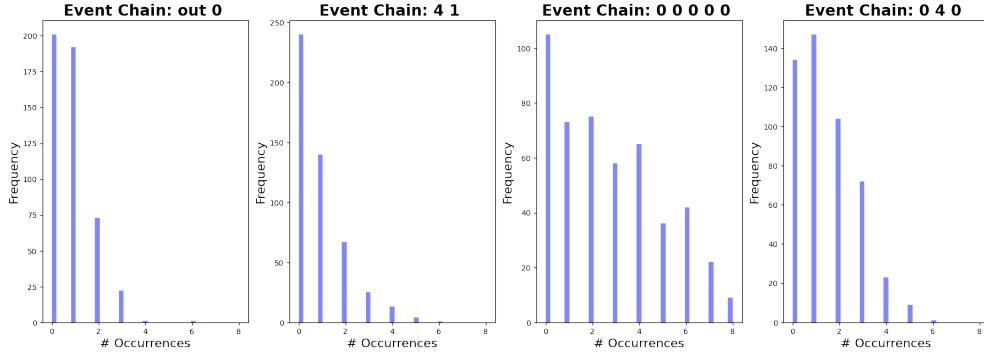


Figure 11: Ground truth output frequencies to some of the queries used in multi-choice query type questions in the train set of LCric for 2-over, 4-over, 8-over and 10-over clips.