Checklist

1. For all authors...
   - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   - (b) Did you describe the limitations of your work? [Yes] Yes, the limitations are included in Section 5
   - (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work most likely does not have any negative societal impact. We have discussed the general societal impact in Section 1
   - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   - (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   - (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will include the links to the dataset in supplementary material.
   - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] As we are not training any models random seeds are not used in this study. Furthermore repeating experiments with proprietary models is costly.
   - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The are provided in Section 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   - (a) If your work uses existing assets, did you cite the creators? [Yes] The only assets we use are open-source models which we have cited.
   - (b) Did you mention the license of the assets? [Yes]
   - (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our dataset does not contain any PIIs other than what is shown in the video and is publicly available.
5. If you used crowdsourcing or conducted research with human subjects...
   - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix D. The instructions are available in the screen shots of the task HTML template we used.
   - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] Yes, the study is approved as STUDY00020473. The approval document will be shared upon request.
   - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Yes, we calibrated the price per task in a way that workers could earn $15 per hour which is the minumum wage.

A Open Model Details

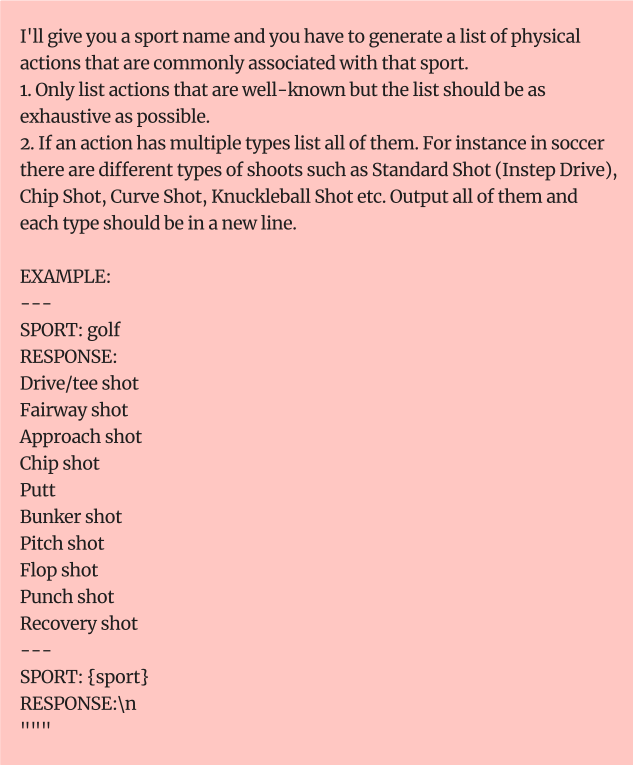
- mPLUG-Owl [58] is one of the first to align both image and video modalities to large language model. This is achieved with the Qformer-based abstractor module [23] that summarizes long and dense visual information with learnable tokens, which are then combined text queries as input to the language model.
- VideoChatGPT adapts LLaVA [29], an image-base visual instruction tuned model, to video understanding tasks by temporally pooling the sequence of frame embeddings to get the video-level features. These features are projected by linear layer as language embedding tokens and passed down to language model. The model is trained with 100,000 video-instruction pairs annotated by language models.
- Unlike the above work that does not integerate audio, VideoLLAMA [60] integrates two QFormers, one for video and audio branch, and aligns the output of both visual & audio encoders with LLM's embedding space.
- VideoLaVIT [17] efficiently captures the dense sequence of video by representing each video as keyframes and temporal motions. Specifically, the spatiotemporal motion encoder captures the time-varying contextual information contained in extracted motion vectors, thereby significantly enhancing LLMs' ability to comprehend the intricate actions in video. The key frame and motion tokens are then adapted to the LLMs.
- VideoChat2 [25] progressively trains the visual encoder and Qformer to LLMS, with the comprehensive instruction tuning dataset. Unlike prior work, the work adds multiple set of instruction tuning dataset curated from public dataset and newly instructions generated by ChatGPT, leading to huge boost in performance across diverse downstream task.
- LLaVA-Next-Video [62] efficiently adapts LLaVA [29] to efficiently pass in long sequence of videos with high resolution with their AnyRes algorithm. It also introduces DPO [38, 61] variant of the model trained by the preference data generated by LLM, where videos are represented with their detailed captions as supporting evidence.

Table 2: Architecture details of open source models and question prompts used in the input text.

Model	**LLM**	**Visual Encoder**	**Image Size**	**Question Prompt**
mPLUG-Owl [58]	LLAMA-7B [48]	CLIP ViT-L/14 [36]	224	Only give the best option.
VideoChatGPT [31]	Vicuna-7B-v1.1 [4]	CLIP ViT-L/14 [36]	224	*Answer with the option's letter from the given choices directly.*
VideoLLaMA [60]	LLAMA2-7B [49]	EVA ViT-G/14 [45]	224	Only give the best option.
Video-LaVIT [17]	LLAMA2-7B [49]	EVA ViT-G/14 [45]	224	Only give the best option.
VideoChat2 [25]	Vicuna-7B-v0 [4]	UMT-L [24]	224	*Only give the best option.*
LLaVA-Next-Video [62]	Vicuna-7B-v1.5 [4]	CLIP ViT-L/14-336 [36]	336	*Answer with the option's letter from the given choices directly.*

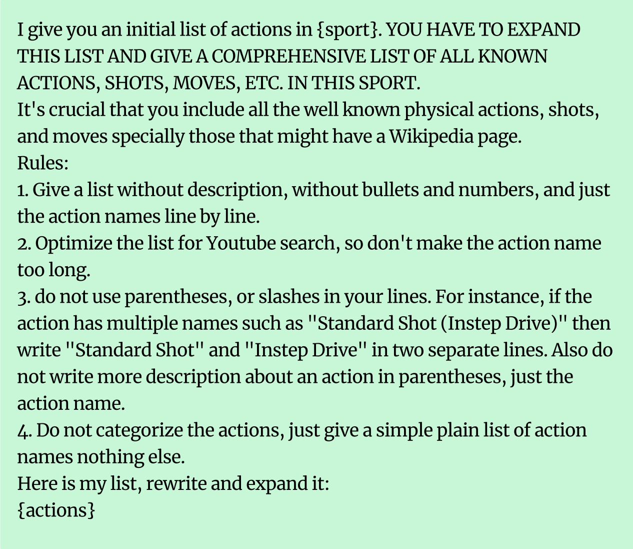
Table 2 further includes the architecture details and the input question prompt used for the open models during evaluation. We use the following system prompt for all the models: "*Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Based on your observations, select the best option that accurately addresses the question.*" The input question and multiple choice options are formulated as "Question: {`question`} Options: {`choices`}", and the output response is parsed to acquire the correct letter choice.

B Prompts



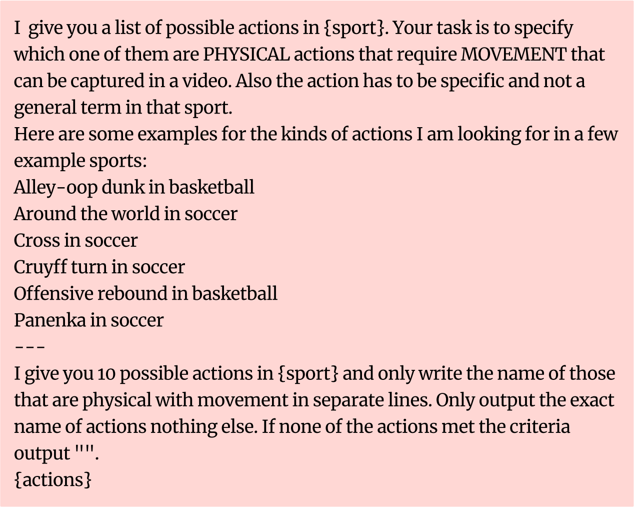
```
I'll give you a sport name and you have to generate a list of physical
actions that are commonly associated with that sport.
1. Only list actions that are well-known but the list should be as
exhaustive as possible.
2. If an action has multiple types list all of them. For instance in soccer
there are different types of shoots such as Standard Shot (Instep Drive),
Chip Shot, Curve Shot, Knuckleball Shot etc. Output all of them and
each type should be in a new line.

EXAMPLE:
---
SPORT: golf
RESPONSE:
Drive/tee shot
Fairway shot
Approach shot
Chip shot
Putt
Bunker shot
Pitch shot
Flop shot
Punch shot
Recovery shot
---
SPORT: {sport}
RESPONSE:\n
"""
```

Figure 6: GPT4 Prompt used for finding initial actions in different sports.



```
I give you an initial list of actions in {sport}. YOU HAVE TO EXPAND
THIS LIST AND GIVE A COMPREHENSIVE LIST OF ALL KNOWN
ACTIONS, SHOTS, MOVES, ETC. IN THIS SPORT.
It's crucial that you include all the well known physical actions, shots,
and moves specially those that might have a Wikipedia page.
Rules:
1. Give a list without description, without bullets and numbers, and just
the action names line by line.
2. Optimize the list for Youtube search, so don't make the action name
too long.
3. do not use parentheses, or slashes in your lines. For instance, if the
action has multiple names such as "Standard Shot (Instep Drive)" then
write "Standard Shot" and "Instep Drive" in two separate lines. Also do
not write more description about an action in parentheses, just the
action name.
4. Do not categorize the actions, just give a simple plain list of action
names nothing else.
Here is my list, rewrite and expand it:
{actions}
```

Figure 7: GPT4 Prompt used expanding the action list.



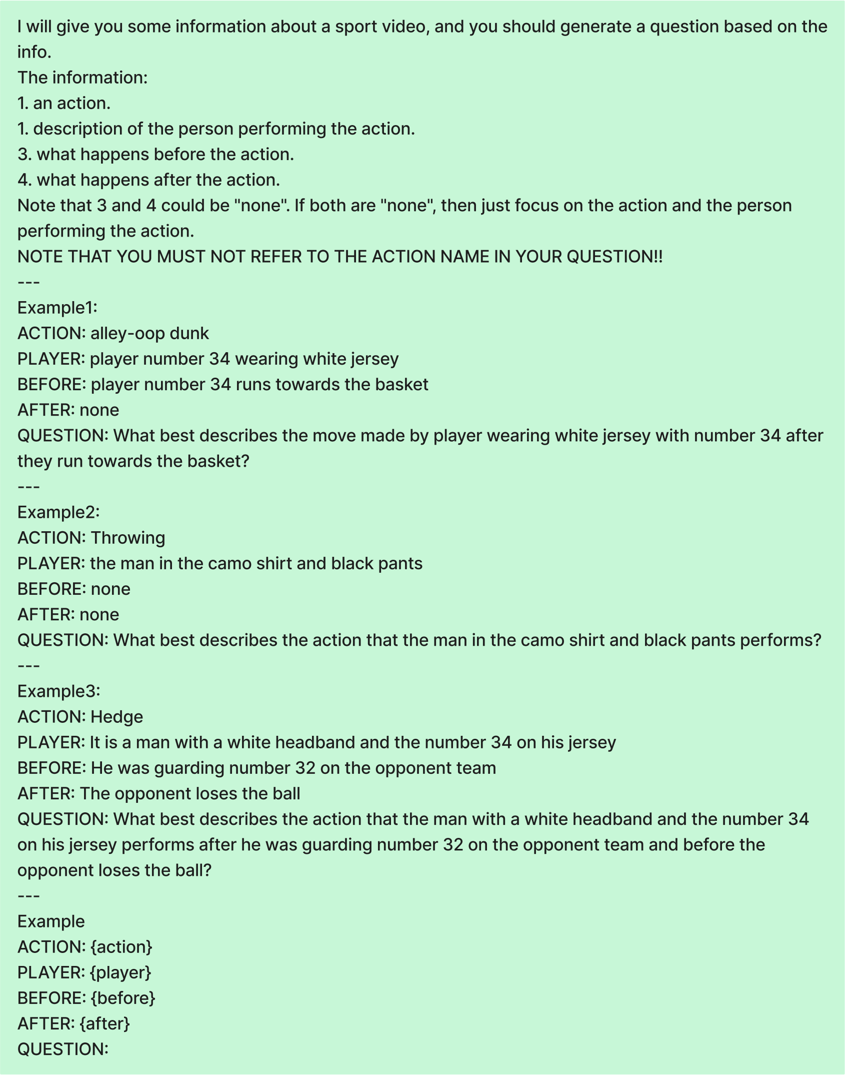
```
I  give you a list of possible actions in {sport}. Your task is to specify
which one of them are PHYSICAL actions that require MOVEMENT that
can be captured in a video. Also the action has to be specific and not a
general term in that sport.
Here are some examples for the kinds of actions I am looking for in a few
example sports:
Alley-oop dunk in basketball
Around the world in soccer
Cross in soccer
Cruyff turn in soccer
Offensive rebound in basketball
Panenka in soccer
---
I give you 10 possible actions in {sport} and only write the name of those
that are physical with movement in separate lines. Only output the exact
name of actions nothing else. If none of the actions met the criteria
output "".
{actions}
```

Figure 8: GPT4 Prompt used for shrinking the list and removing non-physical actions.

C Jail-breaking Multi-modal Gemini

When investigating Gemini models on the Vertex AI web app, we noticed that it might leak some information about how Gemini processes multi-modal inputs:

1. Figure 12 shows a screenshot of Google's Vertex web app. When feeding an image the token count is always 258, regardless of resolution. Therefore, if the number of tokens shown is accurate (which might not be) this could imply all images are resized to a certain size before feeding to the model. One hypothesis could be that there are 16×16 patches that are fed to the model with two indicator tokens such as "<IMAGE>" and "</IMAGE>".
2. With videos, we noticed that the only factor that seemed to matter in token count was the video length in time. If a video had N frames, the token count shown was always $\lfloor N/FPS \rfloor \times 265$. Therefore, according to the web app, each still image takes 258 tokens and each video frame takes 265 tokens. Those extra tokens in videos might be the timestamp tokens accompanying each frame.
3. Another unusual observation was that when we uploaded a video with fewer frames than the FPS, the token count shown is zero. Yet, the model still processes and describes what's in the video somewhat correctly. This could potentially indicate that the web app calculates number of tokens offline using a predetermined formula without counting the actual tokens that are fed to the model.
4. One potential implication of the above observations is that the video model always sample one frame per second when processing videos. We investigated further and were able to recover the exact frames that model samples from videos. If the frame rate of the video is N, then Gemini samples middle frame from each second. Therefore the indices of sampled frame numbers will be $N/2$, $N/2 + N$, $N/2 + 2N$, $N/2 + 3N$ and so forth.
5. The way to test the above claim is to inject some random images inside a regular video at those positions. When you feed such inputs to the model and ask the model to describe it with a prompt such as "Exactly describe what's happening in this video. Don't leave out any details" the model only describes still images and nothing about the video; or outputs a response such as "“The provided video is a still image and does not contain any video or movement to describe". We could reproduce this behavior every time we fed the input.
6. Even if the video is an unsafe content (e.g. NSFW), by changing those specific frames, the model describes only those injected images. However, if one the frames at those positions is changed to an unsafe image the model does not output anything.



```
I will give you some information about a sport video, and you should generate a question based on the
info.
The information:
1. an action.
1. description of the person performing the action.
3. what happens before the action.
4. what happens after the action.
Note that 3 and 4 could be "none". If both are "none", then just focus on the action and the person
performing the action.
NOTE THAT YOU MUST NOT REFER TO THE ACTION NAME IN YOUR QUESTION!!
---
Example1:
ACTION: alley-oop dunk
PLAYER: player number 34 wearing white jersey
BEFORE: player number 34 runs towards the basket
AFTER: none
QUESTION: What best describes the move made by player wearing white jersey with number 34 after
they run towards the basket?
---
Example2:
ACTION: Throwing
PLAYER: the man in the camo shirt and black pants
BEFORE: none
AFTER: none
QUESTION: What best describes the action that the man in the camo shirt and black pants performs?
---
Example3:
ACTION: Hedge
PLAYER: It is a man with a white headband and the number 34 on his jersey
BEFORE: He was guarding number 32 on the opponent team
AFTER: The opponent loses the ball
QUESTION: What best describes the action that the man with a white headband and the number 34
on his jersey performs after he was guarding number 32 on the opponent team and before the
opponent loses the ball?
---
Example
ACTION: {action}
PLAYER: {player}
BEFORE: {before}
AFTER: {after}
QUESTION:
```

Figure 9: GPT4 Prompt used for writing questions about the video segments.

D Mechanical Turk HITs

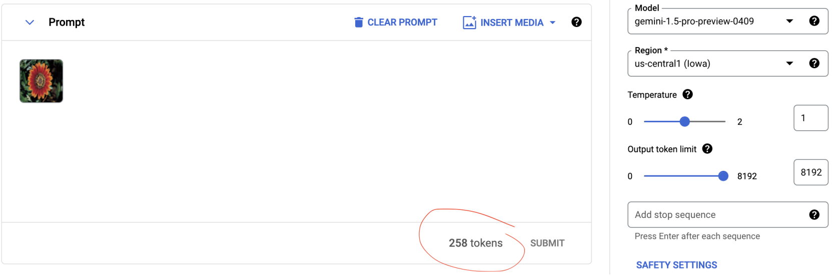
Figures 14 and 15 shows the templates and the instructions used for verification and localization of actions with the help of crowd-workers on Amazon Mehcanical Turk. For both tasks we calibrated the price per hit so that the workers could earn $15 per hour which is the minimum wage.

```
Write some hard negatives for move {action} in sport
{domain}.
The negatives should be plausible and EXTREMELY hard to
distinguish from the correct answer. However, THEY MUST
BE WRONG AND DIFFERENT from the correct one. Also, the
hard negatives must be well-known {domain} moves.
for each action, write 9 hard negatives. and write one hard
negative in a line without any bullet points or numbers.
----
EXAMPLE:
ACTION: windshield wiper forehand
VERY HARD NEGATIVES:
Inside-out forehand
Topspin lob
Slice backhand
Flat serve
Kick serve
Reverse forehand
Volley at the net
Drop shot
Two-handed backhand
----
ACTION: {action}
VERY HARD NEGATIVES:
```

Figure 10: GPT4 Prompt used for writing hard negatives for an action.

```
Answer the given question according to the video. Only
output the choice number and nothing else. When
answering the question consider all legal and illegal moves
and drills.\n{question}\n{options}
```

Figure 11: Final prompt used to evaluate proprietary models.



Model: gemini-1.5-pro-preview-0409

Region *: us-central1 (Iowa)

Temperature 0 2 1

Output token limit 0 8192 8192

Add stop sequence

Press Enter after each sequence

258 tokens

Figure 12: Screenshot of Google's Vertex AI web app.

E Link to Dataset

Google Drive The link to the main dataset jsonl file: https://drive.google.com/file/d/1TtJ2hu6etf8js7RzWaRGBWiTSKDLSTRP/view?usp=sharing
The metadata file containing information about the keys of each object in the jsonl file: https://drive.google.com/file/d/1zONJO-Xdhp9A23U-7gm9vZaXNZQs3p4R/view?usp=sharing

Note: The data is in JSONL format, which is a widely recognized format. The metadata corresponding to our dataset is simple and only contains description of the keys of objects. Because of simplicity of the data, we chose not to use tools such as ML Croissant to create the data and metadata files. We will host the data on Huggingface and our GitHub repository.

F Dataset Statistics

Number of actions: 284
Number of videos: 557
Number of sports: 43
Average length of videos: 5.55 seconds
Average frame rate of videos: 32.7 FPS
Distribution of actions per sport:

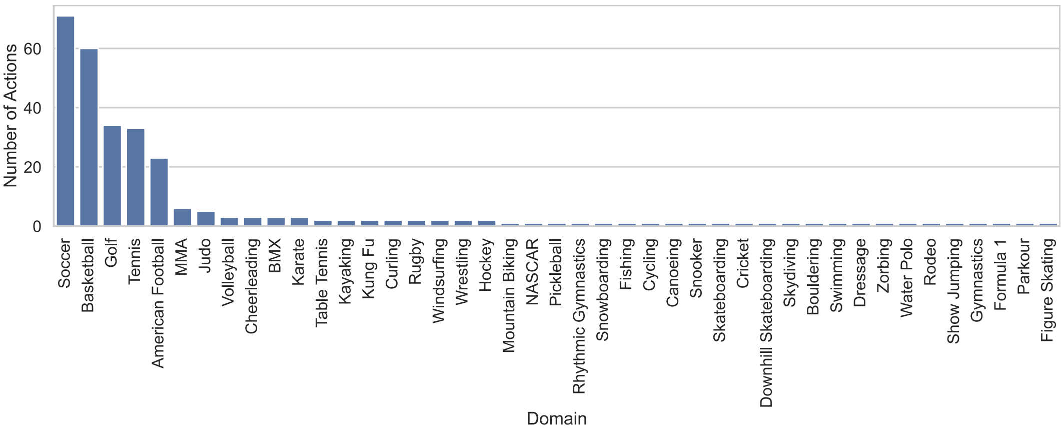


Figure 13: **Distribution of actions across sports.**

G Datasheet

G.1 Motivation

- **For what purpose was the dataset created?** The main purpose of creating `ActionAtlas` was to evaluate state-of-the-art VLMs on identifying fine-grained actions that are not easily recognizable by a single frame. Correctly recognizing such actions necessitates the following capabilities which we believe were missing in previous video datasets, especially action recognition datasets: 1. High frame sample rate to catch fine motions in the action. 2. Correctly tracking the action actor in both time and space across the frames.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The dataset is created by RAIVN lab at the University of Washington.
- **Who funded the creation of the dataset?** The project was funded by Microsoft Accelerate Foundation Models Research program, University of Washington, and Allen Institute for Artificiall Intelligence.

G.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Each instance represents a fine-grained action in some sports which consists of a video, a question, and five multiple choice options from which only one is correct.

- **How many instances are there in total (of each type, if appropriate)?** There are 557 video-MCQ pairs in the dataset.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** No, the dataset is not a sample of a larger datset.
- **What data does each instance consist of?** Each instance consists of a video, a question, five multiple choice options, and a ground truth answer which is the option number of the ground truth action.
- **Is there a label or target associated with each instance?** Yes, the label for each instance is the correct choice for the question.
- **Is any information missing from individual instances?** No.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** No, the videos are sourced from different authors and creators on YouTube.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** The dataset only consists of a test set.
- **Are there any errors, sources of noise, or redundancies in the dataset?** We employed extensive filtering mechanisms including automatic and AI tools and filtering by crowd-workers and authors to eliminate any potential errors and noise in the data. Some videos in the dataset might be different segments from the same original YouTube video.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** Yes, the data is self-contained.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** No.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No, all the videos are segments of already available and public YouTube videos and they are already filtered by YouTube to remove harmful content.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** No.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** As the videos are sport videos sourced from YouTube, there is a possibility of recognizing famous athletes in the videos. However, when writing questions, we did not use the name of individuals in the videos; instead, we refer to them by general attributes, such as color or number of the jersey. For more details refer to Section 3 of the paper.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** No.

G.3 Collection Process

- **How was the data associated with each instance acquired?** The data was sourced from YouTube.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** We used softwares such as Elasticsearch, GPT4, Whisper, Amazon Mechanical Turk to collect the data.

- **Who was involved in the data collection process (e.g., students, crowd-workers, contractors) and how were they compensated (e.g., how much were crowd-workers paid)?** The student authors and crows-workers. We adjusted the price per task so that the workers could make $15 per hour as the minumum wage.
- **Over what timeframe was the data collected?** The data was collected mainly between January 2024 and June 2024.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** Yes, we got IRB approval for crowd-sourcing on Amazon Mechanical Turk from University of Washington.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** We requested crowd-workers to write questions about the given videos and we do not collect any personal data from them.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** The dataset is unlikely to affect the crowd-workers. Moreover, for the individuals featured in the videos, we refrained from using any personally identifiable information (PII) like names in the questions. Instead, we referred to them using general attributes such as jersey numbers and clothing colors.

G.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** We did many rounds of filtering and cleaning which are discussed in Section 3 of the paper to make sure the data is of high quality. The final videos used in the dataset are raw mp4 videos.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** The raw videos are available on YouTube as an external source.
- **Is the software that was used to preprocess/clean/label the data available?** Yes. For a thorough description of software used refer to section 3 of the paper.

G.5 Uses

- **Has the dataset been used for any tasks already?** No.
- **Is there a repository that links to any or all papers or systems that use the dataset?** No.
- **What (other) tasks could the dataset be used for?** The dataset could be used for video tasks such as Video Understanding, Video Question Answering, and Video Compression.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** No.
- **Are there tasks for which the dataset should not be used?** No.

G.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** No.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** On the dataset's website, Huggingface datasets, and Github.
- **When will the dataset be distributed?** We plan to release the dataset publicly by the end of June 2024.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** The current version of the dataset is licensed under Creative Commons Attribution 4.0.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

G.7 Maintenance

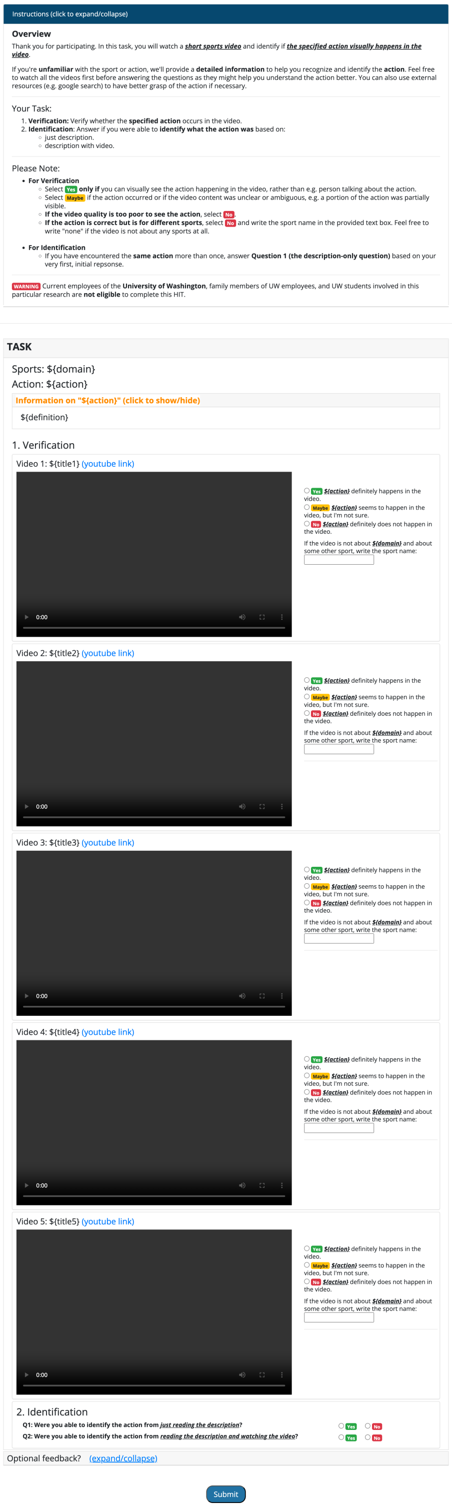
- **Who will be supporting/hosting/maintaining the dataset?** The dataset will be hosted on our website, GitHub repository, Huggingface, and Google drive.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** Email address.
- **Is there an erratum?** No.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Yes, we plan to update the data for any potential errors that will be discovered in the future.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** No.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** Most likely yes.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Yes, we plan to implement such mechanisms on the website of our dataset.

H License

The current version of the dataset is licensed under Creative Commons Attribution 4.0.

I Author Statement

The authors bear all responsibility in case of violation of rights and confirmation of the data license.



Instructions (click to expand/collapse)

Overview

Thank you for participating. In this task, you will watch a ***short sports video*** and identify if ***the specified action visually happens in the video***.

If you're **unfamiliar** with the sport or action, we'll provide a **detailed information** to help you recognize and identify the **action**. Feel free to watch all the videos first before answering the questions as they might help you understand the action better. You can also use external resources (e.g. google search) to have better grasp of the action if necessary.

Your Task:

1. **Verification:** Verify whether the **specified action** occurs in the video.
2. **Identification**: Answer if you were able to **identify what the action was** based on:
   - just description.
   - description with video.

Please Note:

- **For Verification**
  - Select Yes **only if** you can visually see the action happening in the video, rather than e.g. person talking about the action.
  - Select Maybe if the action occurred or if the video content was unclear or ambiguous, e.g. a portion of the action was partially visible.
  - **If the video quality is too poor to see the action**, select No.
  - **If the action is correct but is for different sports**, select No and write the sport name in the provided text box. Feel free to write "none" if the video is not about any sports at all.
- **For Identification**
  - If you have encountered the **same action** more than once, answer **Question 1 (the description-only question)** based on your very first, initial repsonse.

WARNING Current employees of the **University of Washington**, family members of UW employees, and UW students involved in this particular research are **not eligible** to complete this HIT.

TASK

Sports: ${domain}

Action: ${action}

Information on "${action}" (click to show/hide)

${definition}

1. Verification

Video 1: ${title1} (youtube link)

○ Yes ***${action}*** definitely happens in the video.
○ Maybe ***${action}*** seems to happen in the video, but I'm not sure.
○ No ***${action}*** definitely does not happen in the video.

If the video is not about ***${domain}*** and about some other sport, write the sport name:

Video 2: ${title2} (youtube link)

○ Yes ***${action}*** definitely happens in the video.
○ Maybe ***${action}*** seems to happen in the video, but I'm not sure.
○ No ***${action}*** definitely does not happen in the video.

If the video is not about ***${domain}*** and about some other sport, write the sport name:

Video 3: ${title3} (youtube link)

○ Yes ***${action}*** definitely happens in the video.
○ Maybe ***${action}*** seems to happen in the video, but I'm not sure.
○ No ***${action}*** definitely does not happen in the video.

If the video is not about ***${domain}*** and about some other sport, write the sport name:

Video 4: ${title4} (youtube link)

○ Yes ***${action}*** definitely happens in the video.
○ Maybe ***${action}*** seems to happen in the video, but I'm not sure.
○ No ***${action}*** definitely does not happen in the video.

If the video is not about ***${domain}*** and about some other sport, write the sport name:

Video 5: ${title5} (youtube link)

○ Yes ***${action}*** definitely happens in the video.
○ Maybe ***${action}*** seems to happen in the video, but I'm not sure.
○ No ***${action}*** definitely does not happen in the video.

If the video is not about ***${domain}*** and about some other sport, write the sport name:

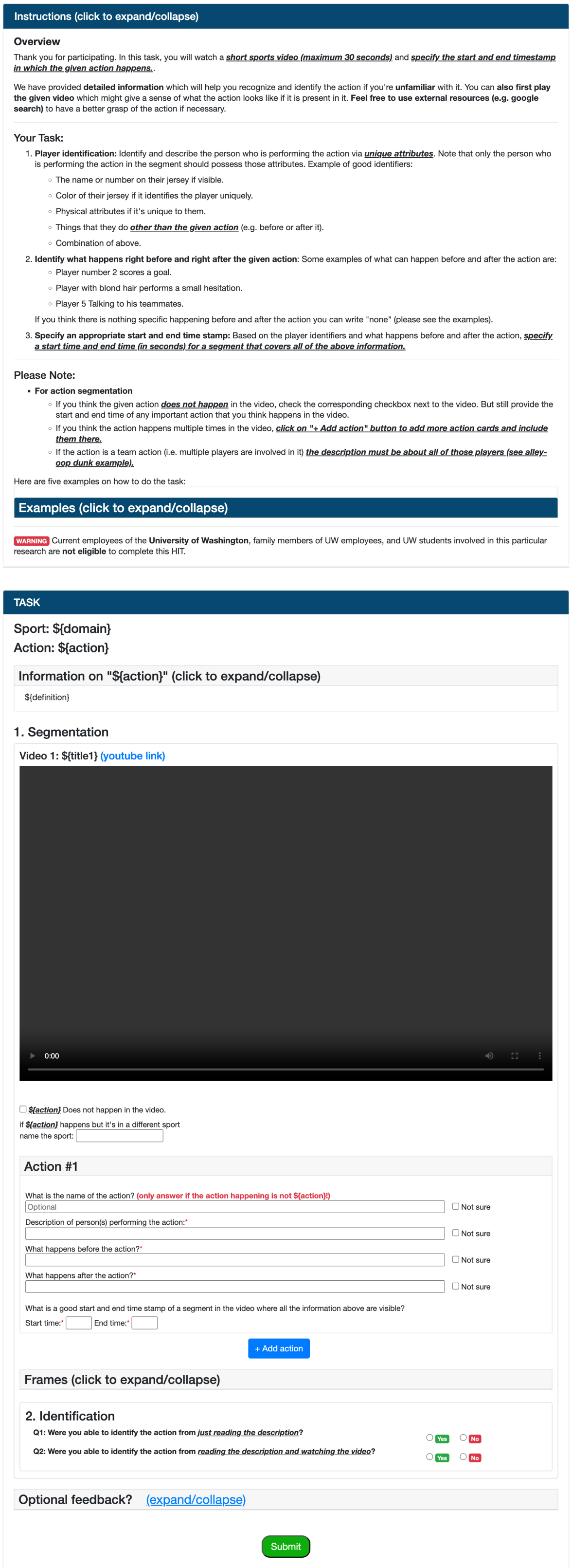
2. Identification

Q1: Were you able to identify the action from *just reading the description*? ○ Yes ○ No

Q2: Were you able to identify the action from *reading the description and watching the video*? ○ Yes ○ No

Optional feedback? (expand/collapse)

Figure 14: **Template used for Verifying presence of actions by crowd-workers.**



Instructions (click to expand/collapse)

Overview

Thank you for participating. In this task, you will watch a ***short sports video (maximum 30 seconds)*** and ***specify the start and end timestamp in which the given action happens.***.

We have provided **detailed information** which will help you recognize and identify the action if you're **unfamiliar** with it. You can **also first play the given video** which might give a sense of what the action looks like if it is present in it. **Feel free to use external resources (e.g. google search)** to have a better grasp of the action if necessary.

Your Task:

1. **Player identification:** Identify and describe the person who is performing the action via ***unique attributes***. Note that only the person who is performing the action in the segment should possess those attributes. Example of good identifiers:
   - The name or number on their jersey if visible.
   - Color of their jersey if it identifies the player uniquely.
   - Physical attributes if it's unique to them.
   - Things that they do ***other than the given action*** (e.g. before or after it).
   - Combination of above.
2. **Identify what happens right before and right after the given action**: Some examples of what can happen before and after the action are:
   - Player number 2 scores a goal.
   - Player with blond hair performs a small hesitation.
   - Player 5 Talking to his teammates.

   If you think there is nothing specific happening before and after the action you can write "none" (please see the examples).
3. **Specify an appropriate start and end time stamp:** Based on the player identifiers and what happens before and after the action, ***specify a start time and end time (in seconds) for a segment that covers all of the above information.***

Please Note:

- **For action segmentation**
  - If you think the given action ***does not happen*** in the video, check the corresponding checkbox next to the video. But still provide the start and end time of any important action that you think happens in the video.
  - If you think the action happens multiple times in the video, ***click on "+ Add action" button to add more action cards and include them there.***
  - If the action is a team action (i.e. multiple players are involved in it) ***the description must be about all of those players (see alley-oop dunk example).***

Here are five examples on how to do the task:

Examples (click to expand/collapse)

WARNING Current employees of the **University of Washington**, family members of UW employees, and UW students involved in this particular research are **not eligible** to complete this HIT.

TASK

Sport: ${domain}
Action: ${action}

Information on "${action}" (click to expand/collapse)

${definition}

1. Segmentation

Video 1: ${title1} (youtube link)

□ ***${action}*** Does not happen in the video.

if ***${action}*** happens but it's in a different sport
name the sport: ______

Action #1

What is the name of the action? **(only answer if the action happening is not ${action}!)**
Optional ______ □ Not sure

Description of person(s) performing the action:*
______ □ Not sure

What happens before the action?*
______ □ Not sure

What happens after the action?*
______ □ Not sure

What is a good start and end time stamp of a segment in the video where all the information above are visible?

Start time:* ______ End time:* ______

+ Add action

Frames (click to expand/collapse)

2. Identification

Q1: Were you able to identify the action from *just reading the description*? ○ Yes ○ No

Q2: Were you able to identify the action from *reading the description and watching the video*? ○ Yes ○ No

Optional feedback? (expand/collapse)

Submit

Figure 15: **Template used for localizing actions in 30 second segments.**