

# Multimodal Explanations by Predicting Counterfactuality in Videos (supplementary material)

## 1. Algorithm for finding maximum subpath in the 3D tensor

We show the algorithm to find the maximum subpath in the 3D tensor based on the dynamic programming proposed by [1] as below.

**Input:**

$M(u, t)$  : the local discriminative scores;  $\{u = (i, j) : \text{the 2D index of spatial coordinate}\} \{t : \text{the frame in the video}\}$

**Output:**

$S(u, t)$  : the accumulated scores of the best path leads to  $(u, t)$ ;

$P(u, t)$  : the best path record for tracing back;

$S^*$  : the accumulated score of the best path;

$l^*$  : the ending location of the best path;

```

 $S(u, 1) = M(u, 1), \forall u;$ 
 $P(u, t) = null, \forall (u, t);$ 
 $S^* = -\infty;$ 
 $l^* = null;$ 
for  $i \leftarrow 2$  to  $n$  do
    foreach  $u \in [1..w] \times [1..h]$  do
         $v_0 \leftarrow \arg \max_{v \in N(u)} S(v, i - 1);$ 
        if  $S(v_0, i - 1) > 0$  then
             $S(u, i) \leftarrow S(v_0, i - 1) + M(u, i);$ 
             $P(u, i) \leftarrow (v_0, i - 1);$ 
        else
             $S(u, i) \leftarrow M(u, i);$ 
        end if
        if  $S(u, i) > S^*$  then
             $S^* \leftarrow S(u, i);$ 
             $l^* \leftarrow (u, i);$ 
        end if
    end for
end for

```

**Algorithm 1:** Algorithm for finding maximum subpath in the 3D tensor [1]

## 2. The influence of the complexity of classification model on the negative class accuracy

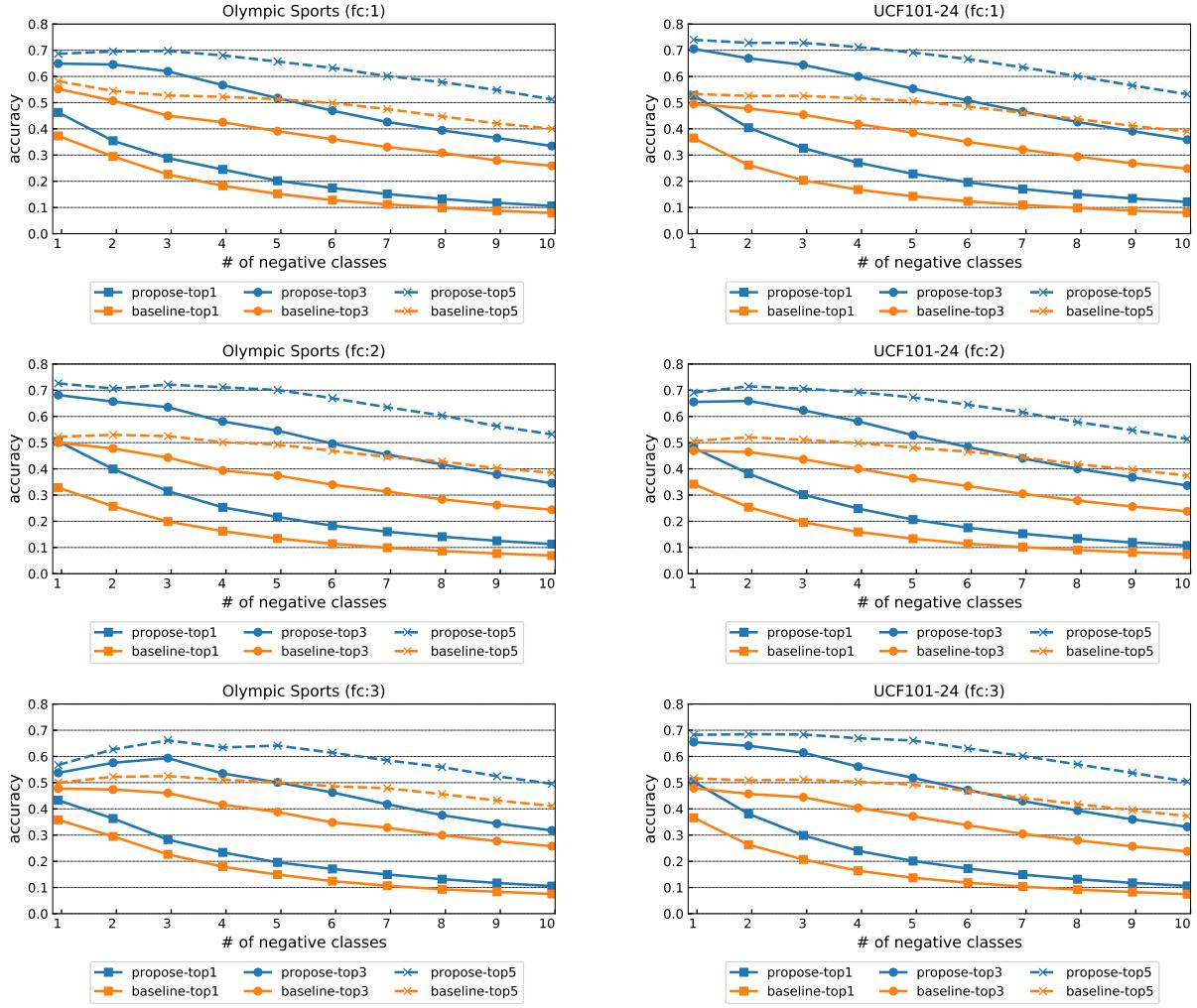


Figure 1. The negative class accuracy on Olympic Sports dataset (left) and UCF101-24 dataset (right). Each row corresponds to the number of fully-connected layer of the classification module. y-axis indicates the mean accuracy and x-axis means the number of the negative classes used for averaging whose prediction value is maximum.

### **3. List of attributes**

#### **3.1. Olympic Sports dataset**

'Run', 'Slow run', 'Fast run', 'Indoor', 'outdoor', 'Ball', 'small ball', 'big ball', 'Jump', 'Small local Jump', 'Local jump up', 'Jump Forward', 'Track', 'Bend', 'StandUp', 'Lift something', 'Raise Arms', 'One Arm Open', 'Turn Around', 'Throw Up', 'Throw away', 'water', 'Down Motion in Air', 'Up Motion in Air', 'Up Down Motion Local', 'Somersault in Air', 'With Pole', 'Two hand holding pole', 'One hand holding pole', 'Spring Platform', 'Motion in the air', 'one arm swing', 'Crouch', 'Two Arms Open', 'Two Arms Swing overhead', 'Turn around with two arms open', 'Run in Air', 'Big Step', 'Open Arm Lift', 'With Pat'

#### **3.2. UCF101-24 dataset**

'Body Motion is Flipping', 'Body Motion is Walking', 'Body Motion is Running', 'Body Motion is Riding', 'Body Motion is Up down', 'Body Motion is Pulling', 'Body Motion is Lifting', 'Body Motion is Pushing', 'Body Motion is Diving', 'Body Motion is Jumping Up', 'Body Motion is Jumping Forward', 'Body Motion is Jumping Over Obstacle', 'Body Motion is Spinning', 'Body Motion is Climbing Up', 'Body Motion is Horizontal', 'Body Motion is Vertical Up', 'Body Motion is Vertical Down', 'Body Motion is Bending', 'Object is Ball Like', 'Object is Big Ball Like', 'Object is Stick Like', 'Object is Rope Like', 'Object is Sharp', 'Object is Circular', 'Object is Cylindrical', 'Object is Musical Instrument', 'Object is Portable Musical Instrument', 'Object is Animal', 'Object is Boat Like', 'Posture is Sitting', 'Posture is Sitting In Front Of Table Like Object', 'Posture is Standing', 'Posture is Lying', 'Posture is Handstand', 'Body Parts Used is Head', 'Body Parts Used is Hands', 'Body Parts Used is Arms', 'Body Parts Used is Legs', 'Body Parts Used is Foot'

## 4. Output Examples



SalsaSpin not SoccerJuggling because Body\_Motion is Pulling



RopeClimbing not SalsaSpin because Body\_Motion is Vertical\_Up



Biking not SkateBoarding because Posture is Sitting



IceDancing not Skiing because Body\_Motion is Spinning



Skijet not Surfing because Body\_Parts\_Used is Arms



SkateBoarding not Surfing because Body\_Parts\_Used is Legs



Biking not Basketball because Posture is Sitting



TennisSwing not SoccerJuggling because Object is Circular



Fencing not Basketball because Object is Sharp



LongJump not CliffDiving because Body\_Motion is Running



LongJump not SkateBoarding because Body\_Motion is Running



Biking not PoleVault because Body\_Parts\_Used is Legs



SkateBoarding not HorseRiding because Body\_Motion is Bending

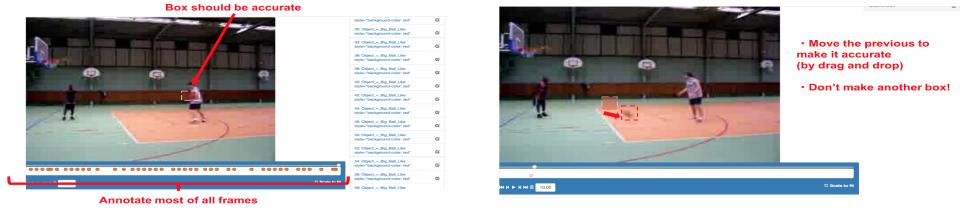
## 5. Dataset collection

Please read the instruction carefully. The submission considered not to read the instruction will be automatically rejected.

Your task is to accurately draw rectangles for the `concept` (such as motion or object) written in red color in the top of the task page.

### Quick Instructions

1. Mark the checkbox `scale to fit`.
2. Watch the whole video (Push the `>` button), and go back to the first frame (Push the `<<` button). After watching, shortcut key will be activated.
3. Draw a box accurately around the `concept`.
4. Hit `s` to go step through frames.
5. Adjust the box to keep it accurate. Please do not make a new box! Just move the previous (dotted) box! Some workers made this mistake, but submissions having more than two boxes for the same 'concept' will be rejected automatically.



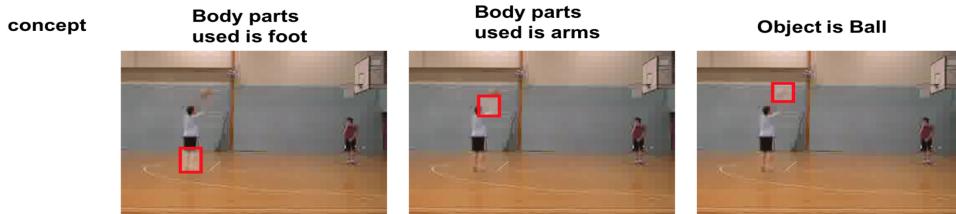
### Keyboard Shortcuts (activated after watching video)

- `a` step back 1 frame
- `s` step forward 1 frame
- `d` or 'delete' to delete selected rectangle

### Important Details

1. Box should be as small as possible, and it does not need to cover the area of the person. For example, if the `concept` is 'Body\_Parts\_Used is Arms', box should cover only arms, not legs or other parts.
2. If you do not see the `concept` in particular frames, you do not need to draw box in those frames. If the target `concept` disappears then delete the rectangle. (press 'delete' with the box selected) For example, if the `concept` is 'Motion is Flipping', start to draw box when person start to flip, and delete the box when he stops flipping.
3. You can skip frames when the location of box does not change. Dotted box will be used for those frames.
4. Draw box by yourself at least once every 5 frames.
5. After finishing annotation, check if your annotation is correct by watching video again (Push the `>` button).

### Good Examples



### Bad Examples

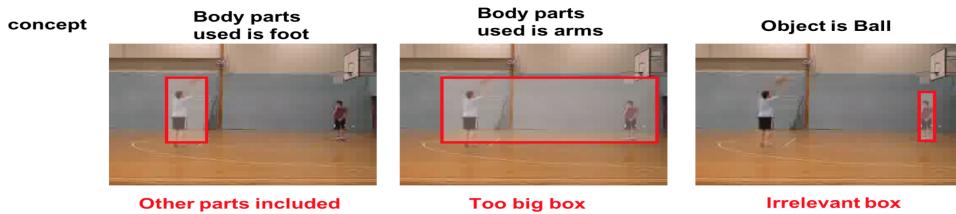


Figure 2. Screen shot of the instruction for collecting bounding box annotation on AWS.

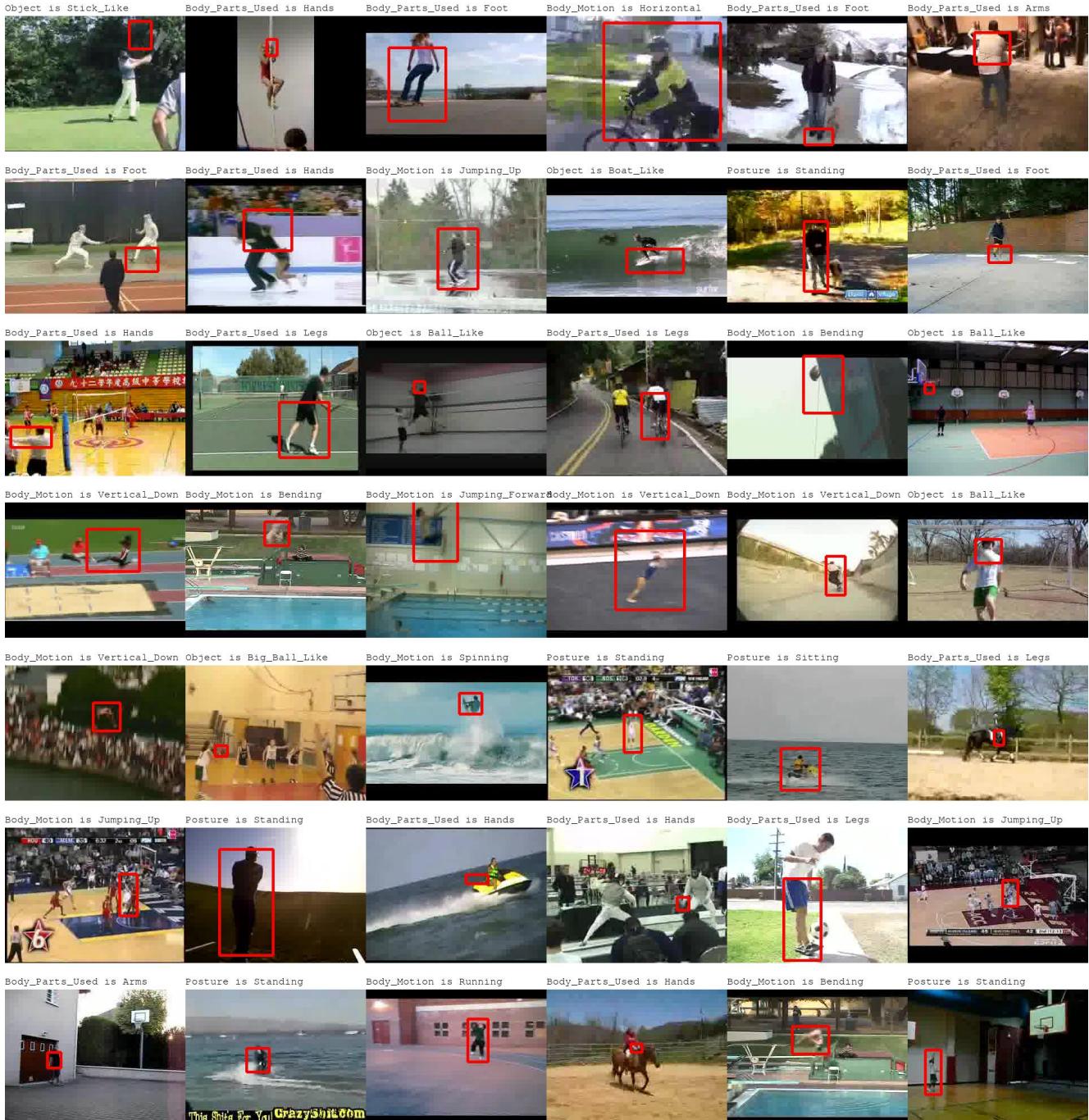
## 6. Dataset Examples











## References

- [1] D. Tran, J. Yuan, and D. Forsyth. Video event detection: From subvolume localization to spatio-temporal path search. 2014. [1](#)