# AdaFrame: Adaptive Frame Selection for Fast Video Recognition

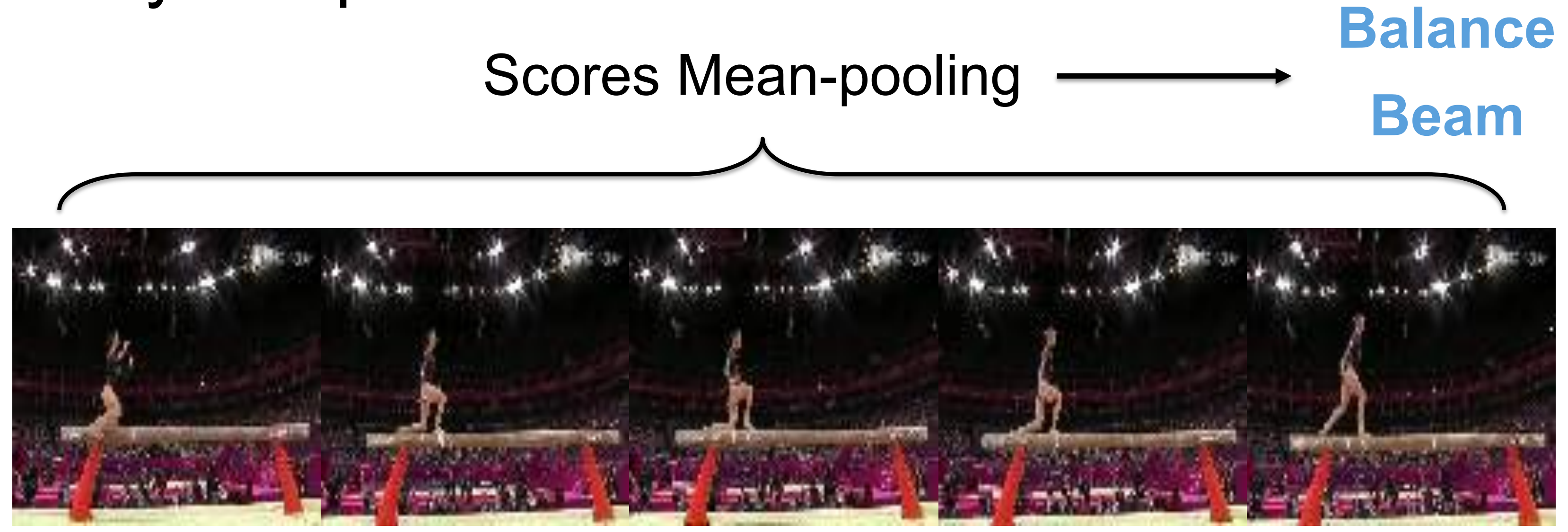
Zuxuan Wu[1], Caiming Xiong[2], Chih-Yao Ma[3], Richard Socher[2], Larry S. Davis[1]

[1]UMD, [2]Salesforce Research, [3]Georgia Tech

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

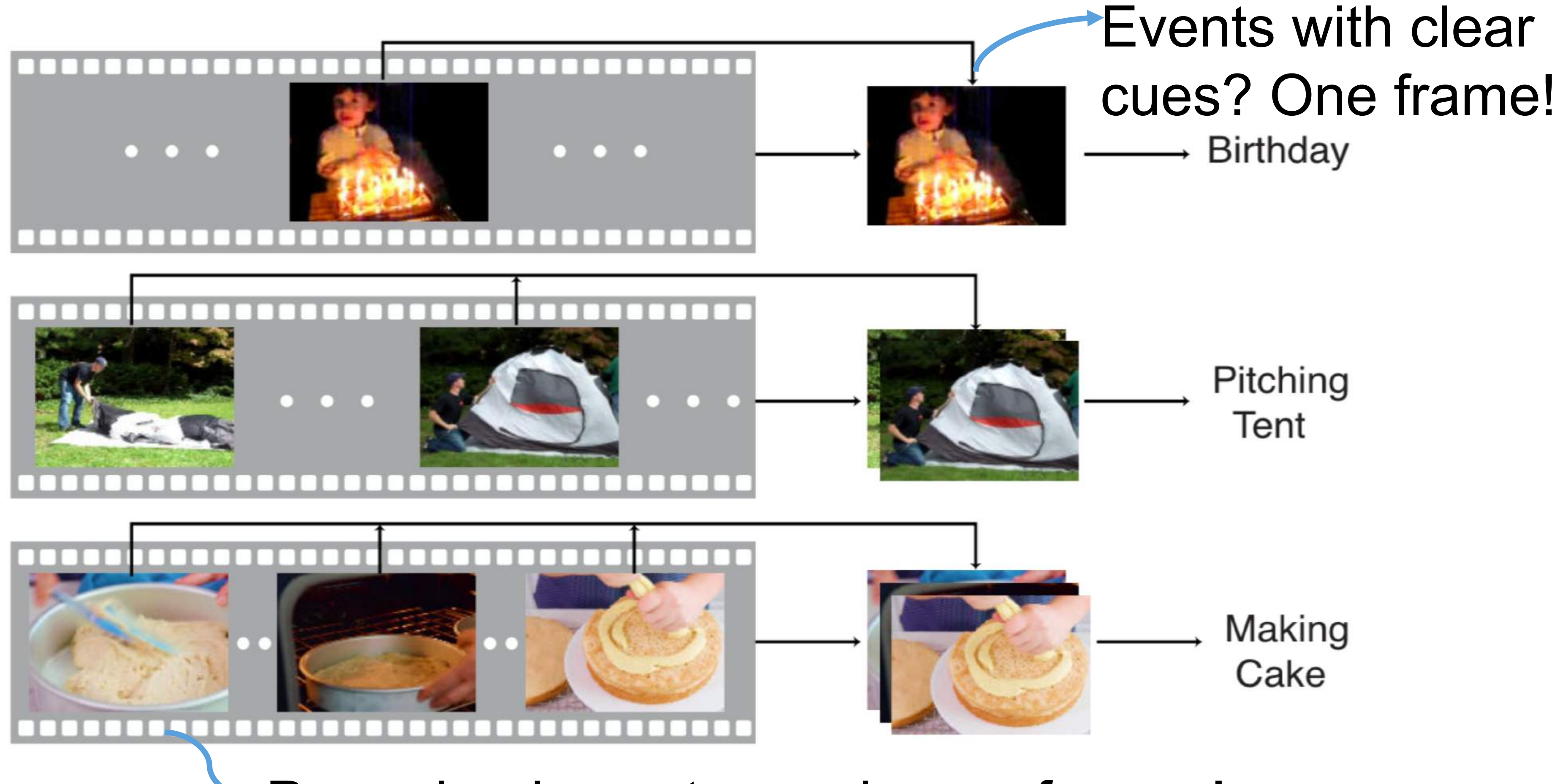
## Overview

Video classification framework: averages scores from 25 uniformly sampled frames.



**Do we really need 25 frames to recognize all videos?**

**Key Observation:** Different video clips have different computational requirements.

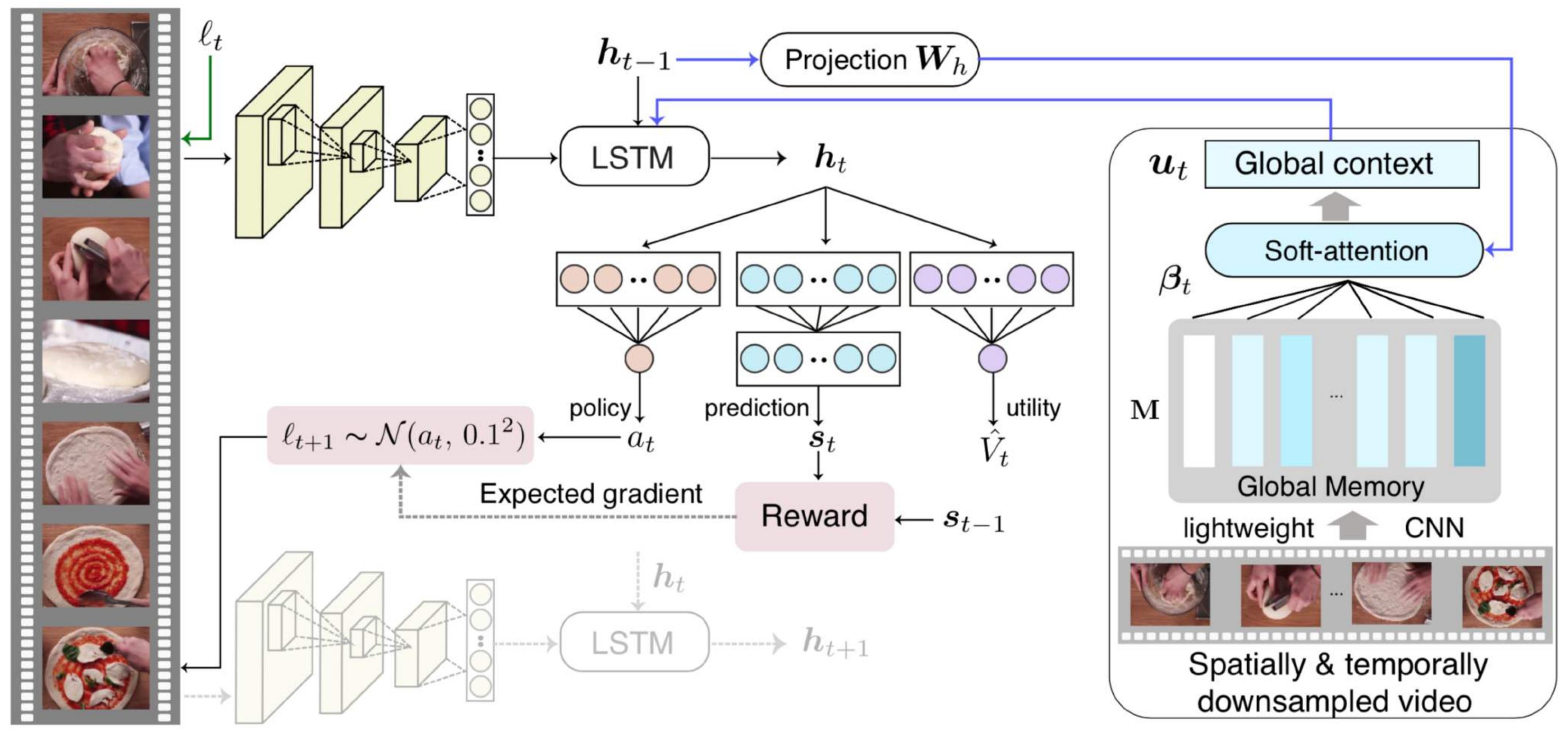


Procedural events need more frames!

**Our Idea:** Learn which frames to use to recognize events/actions on a per-video basis

## Method

**Goal:** Learn video-specific frame usage policies



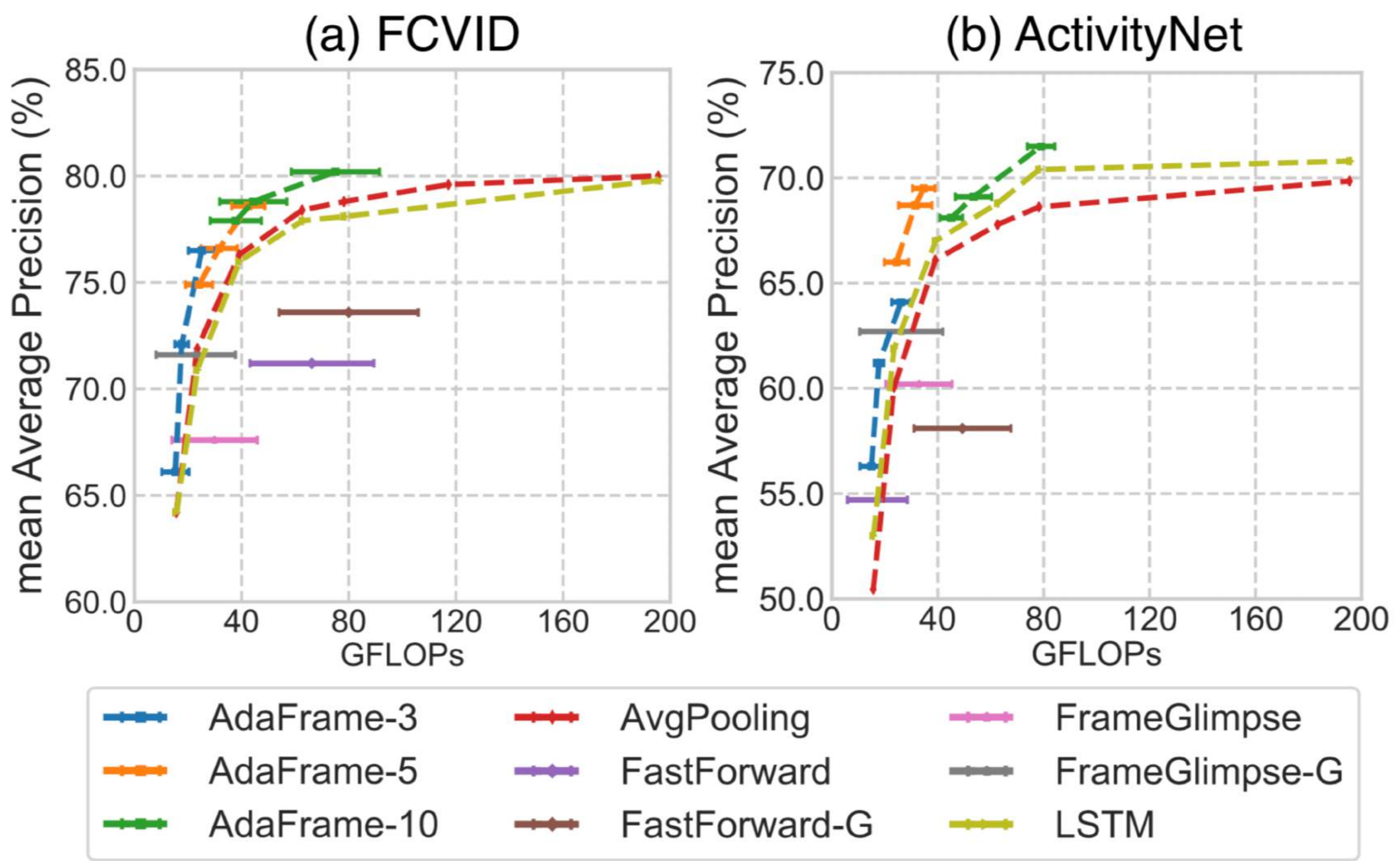
We use a memory-augmented LSTM serving as an agent to interact with the video, consisting of:

- **a memory**, generated with lightweight CNNs, to provide context information
- **a policy net**, sampling from a Gaussian distribution, to decide where to go next
- **a utility net**, measuring advantages of seeing more frames in the future
- **a prediction net**, producing class probabilities

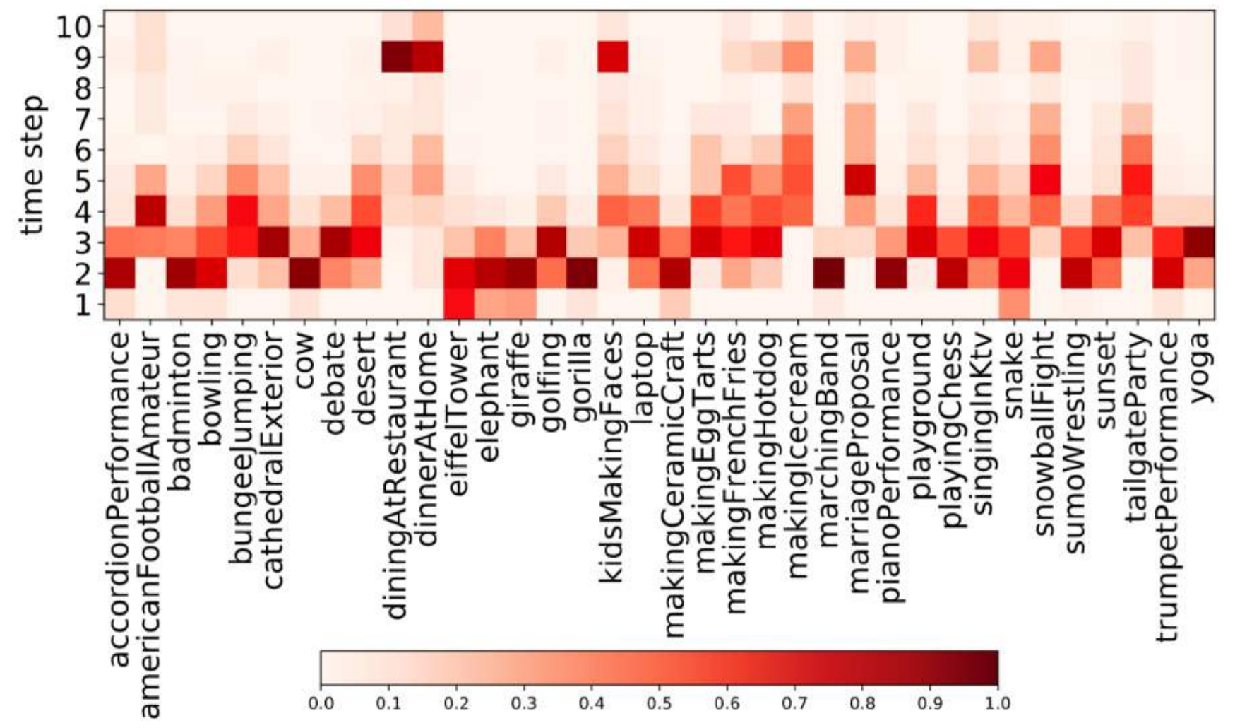
Trained with RL with a reward function forcing more accurate predictions when seeing more frames

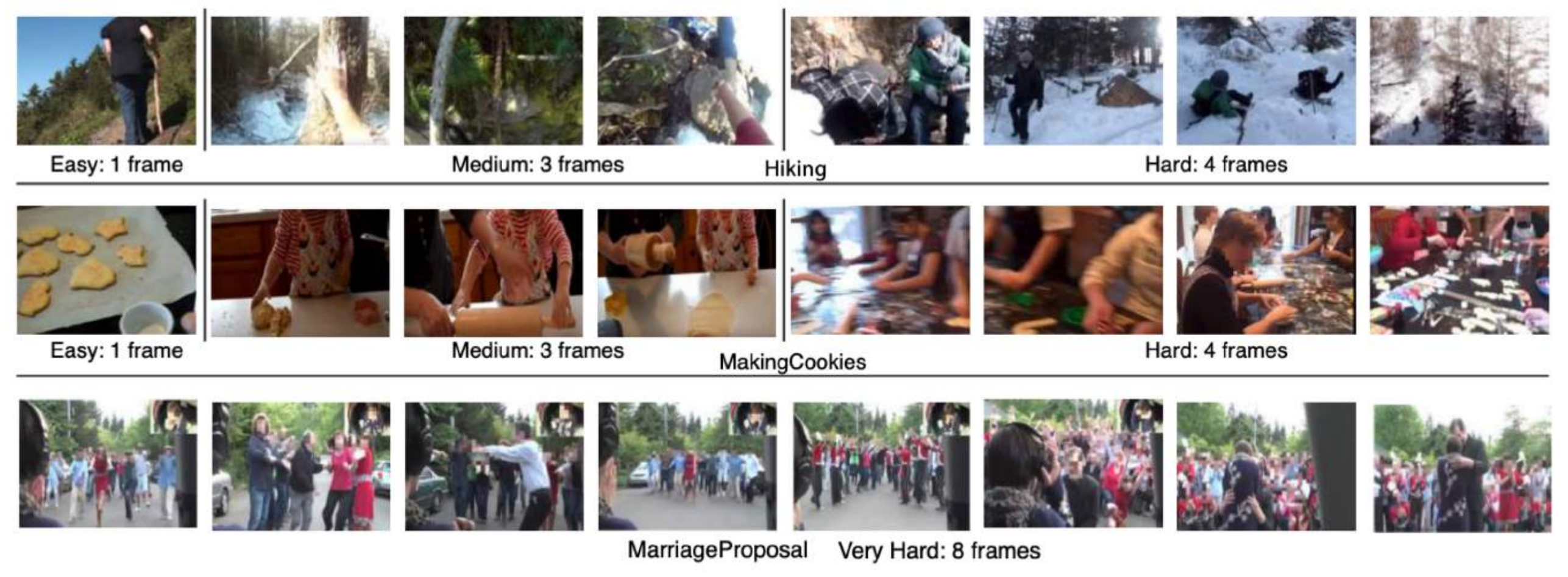
Using predicted future utilities (advantage of seeing more frames) for **adaptive inference**!

## Results



58.9% and 63.3% fewer computations on average (going *as high as 90.6%)* ***without*** **degradation** in accuracy on FCVID (~8.21 frame) and ActivityNet (~8.65 frames), respectively.





**Frame usage indicates the difficulty for prediction**, easier samples need fewer frames while harder ones require more not only **within the same category** but also **among different classes**.