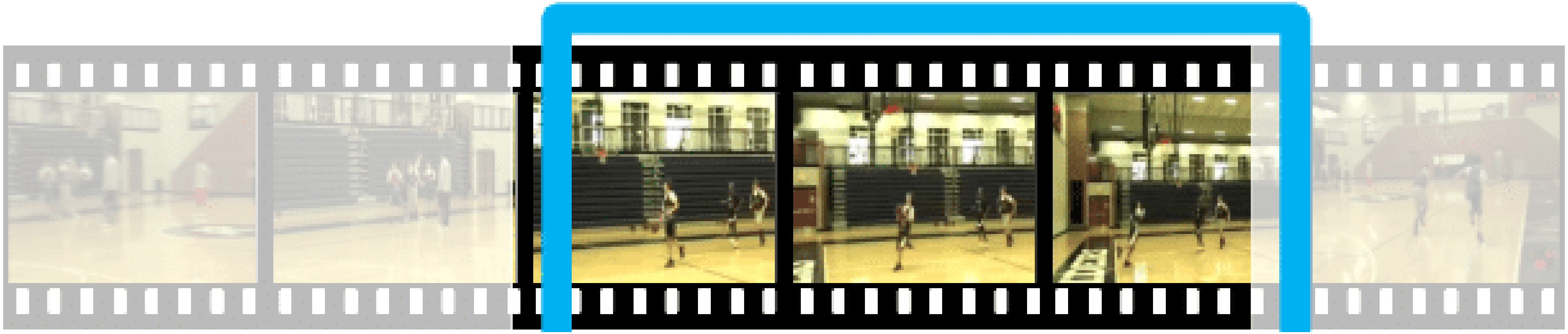
PLOT-TAL: Prompt Learning with Optimal Transport for Few Shot Temporal Action Localization

Edward Fish & Andrew Gilbert

University of Surrey, UK

ICCV HONOLULU HAWAII
OCT 19-23, 2025

CVSSP Centre for Vision, Speech and Signal Processing



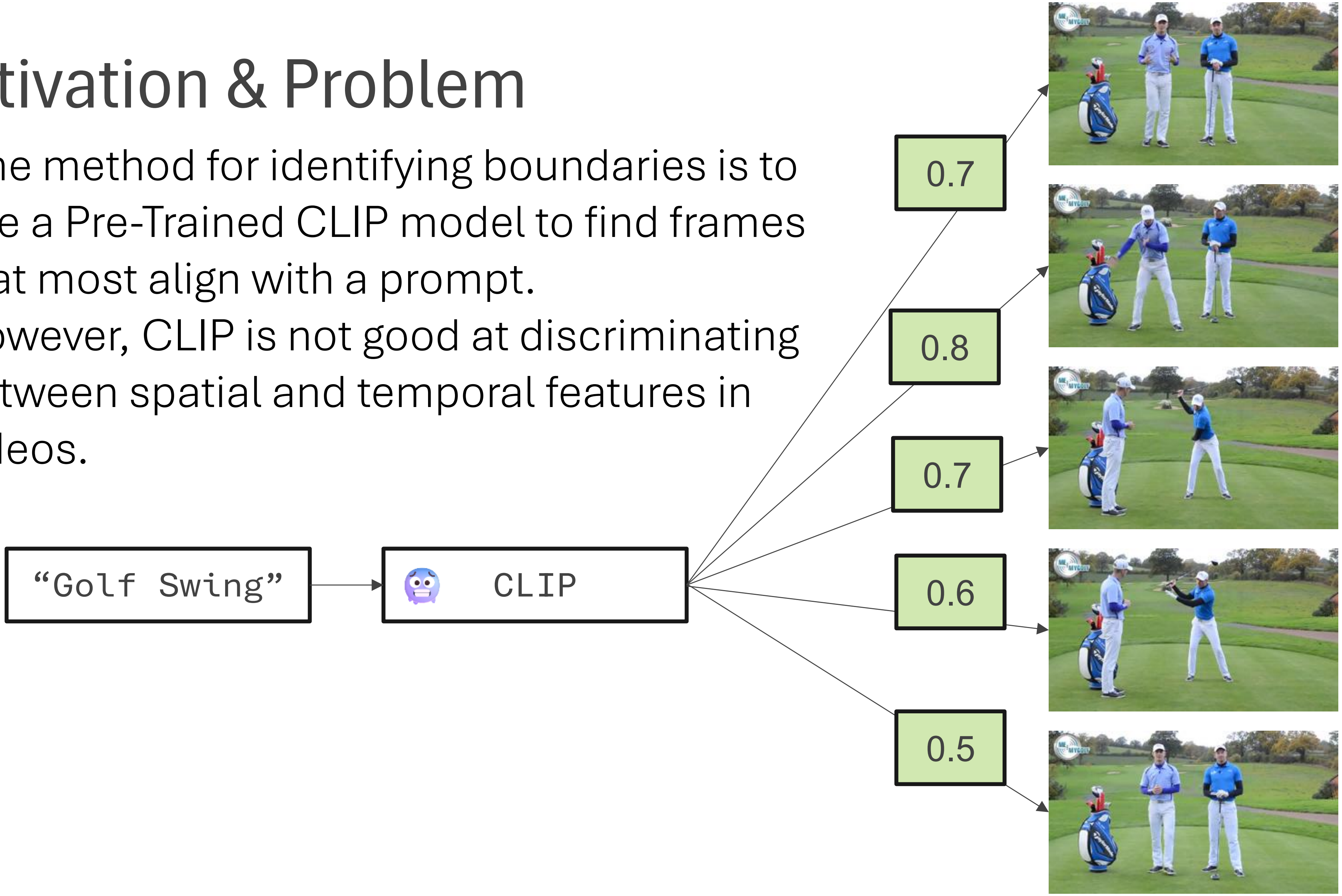
Motivation

In Few-Shot Temporal Action Localization we want to find the start and end of actions in a video given only a small number of training samples per class.

CVSSP Centre for Vision, Speech and Signal Processing

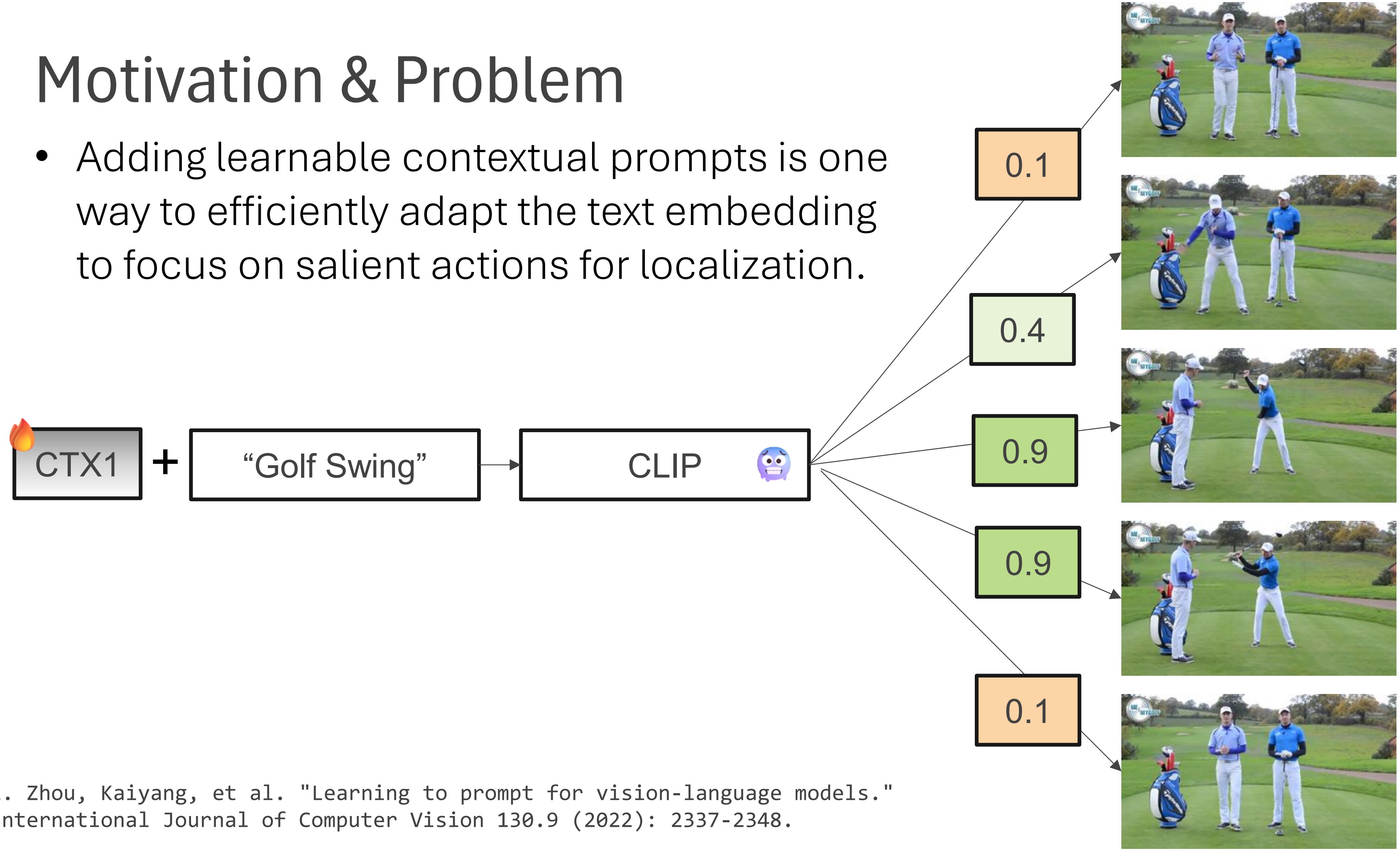
Motivation & Problem

- One method for identifying boundaries is to use a Pre-Trained CLIP model to find frames that most align with a prompt.
- However, CLIP is not good at discriminating between spatial and temporal features in videos.



Motivation & Problem

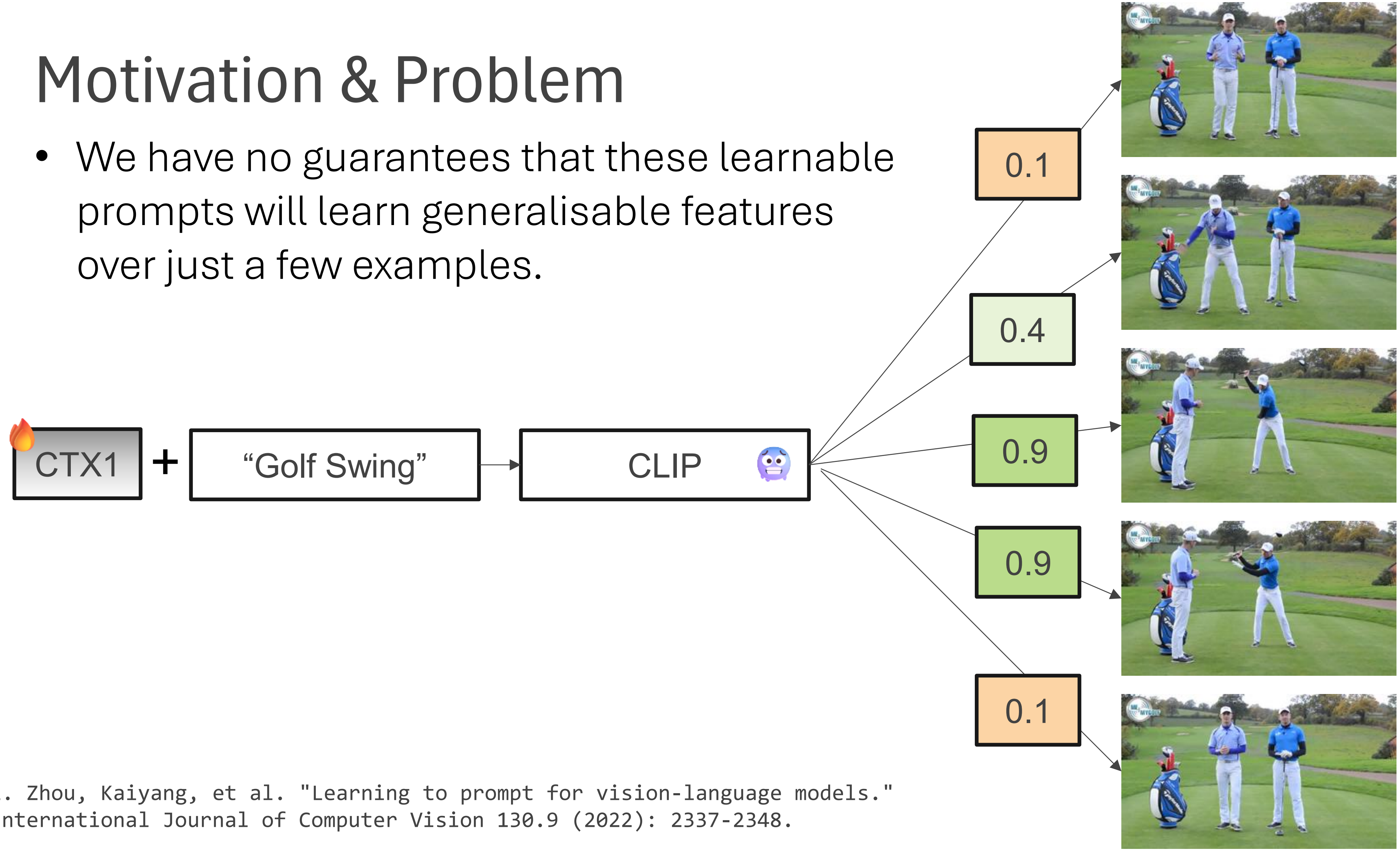
- Adding learnable contextual prompts is one way to efficiently adapt the text embedding to focus on salient actions for localization.



1. Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.

Motivation & Problem

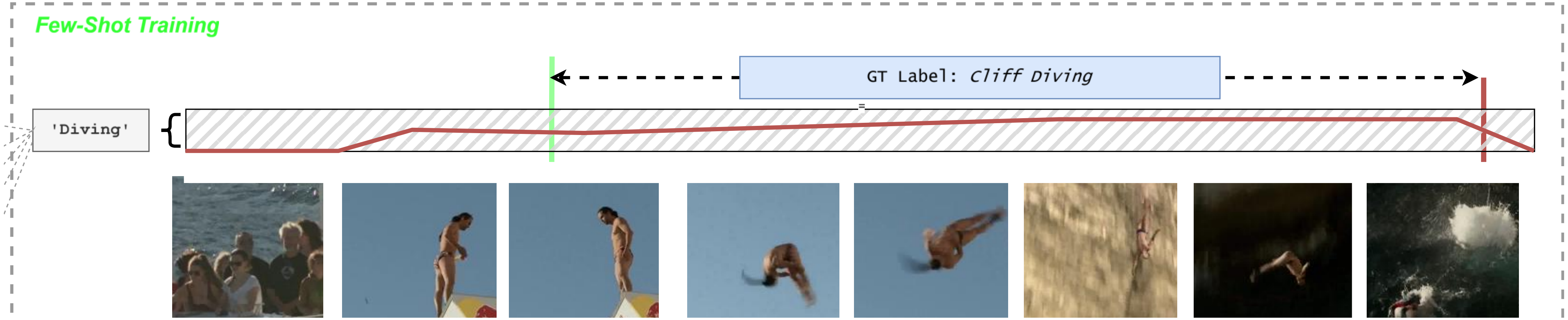
- We have no guarantees that these learnable prompts will learn generalisable features over just a few examples.



1. Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.

Motivation & Problem

- For example, in this case we might learn the boundaries for “`diving`” from cliff diving videos.

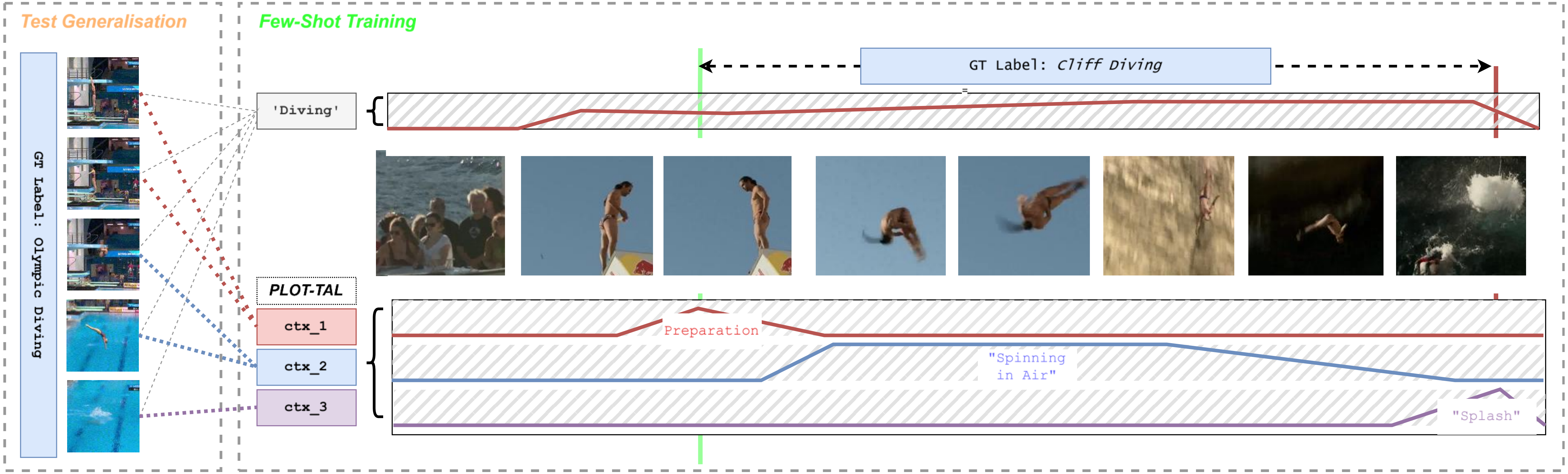


Motivation & Problem

- But if the diving action at test time is in a different context (Olympics), our method will fail to predict good boundaries.

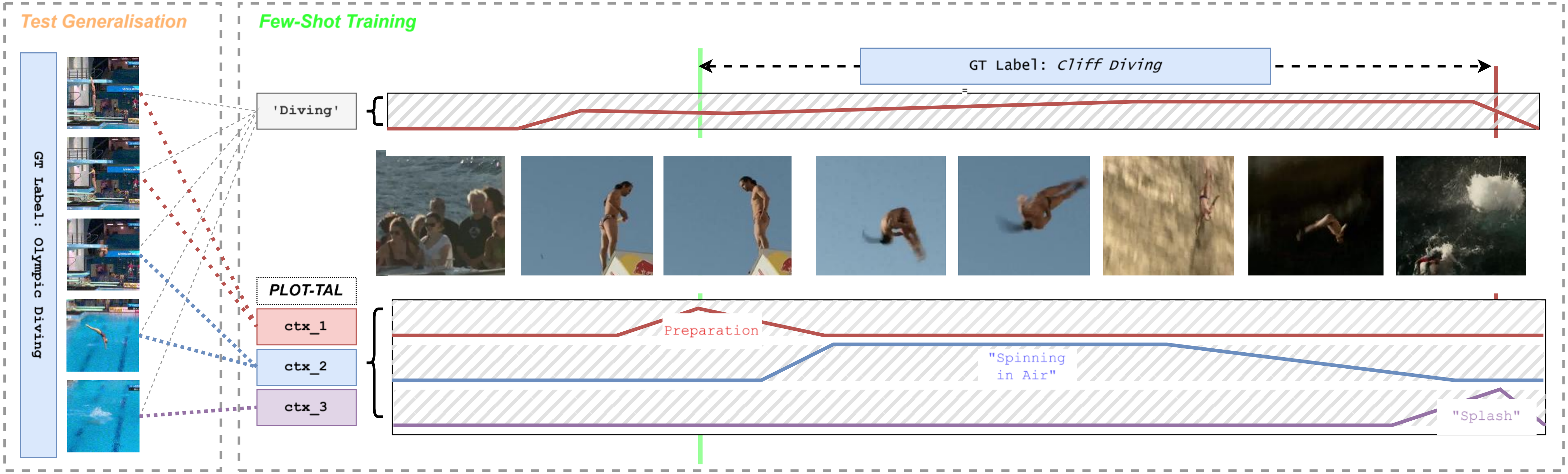
Motivation & Problem

- Our solution is to learn multiple learnable prompts for each class which learn sub-actions which will generalise to new contexts with just a few examples.



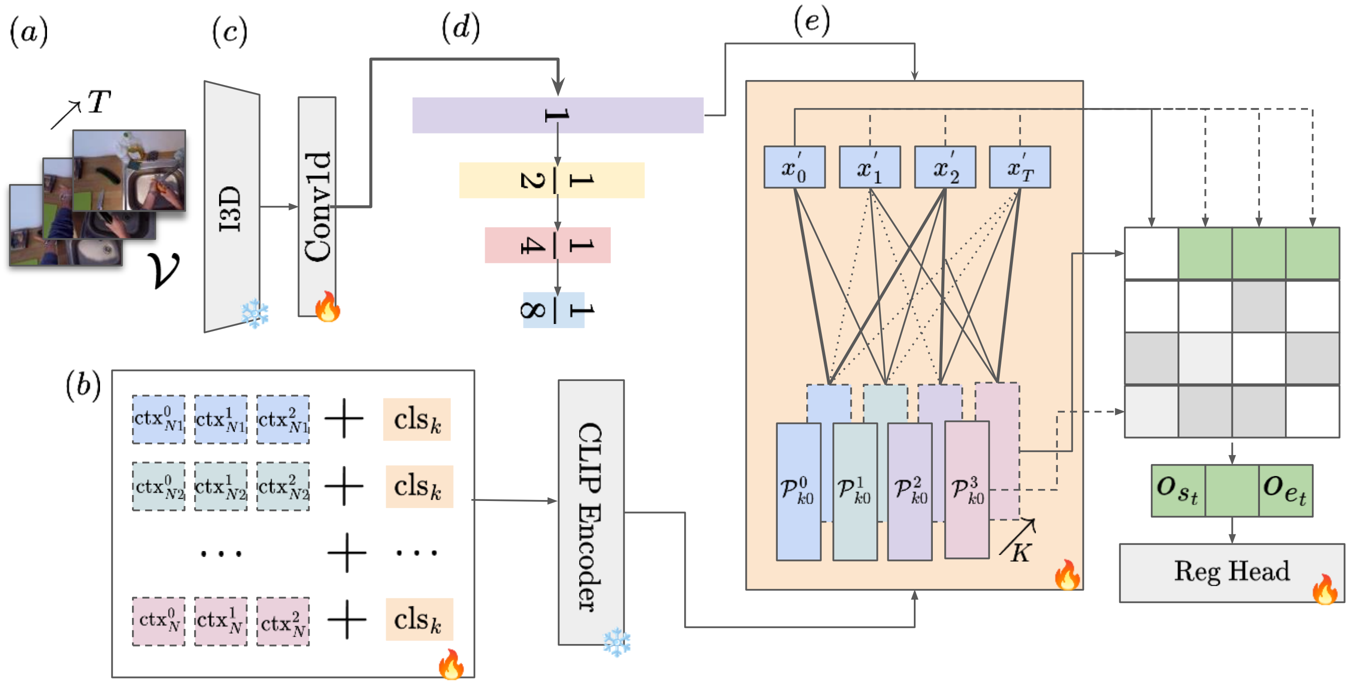
Motivation & Problem

- However, we need to learn these sub actions with only a few examples. How do we ensure they do not also overfit to the mean of image features.



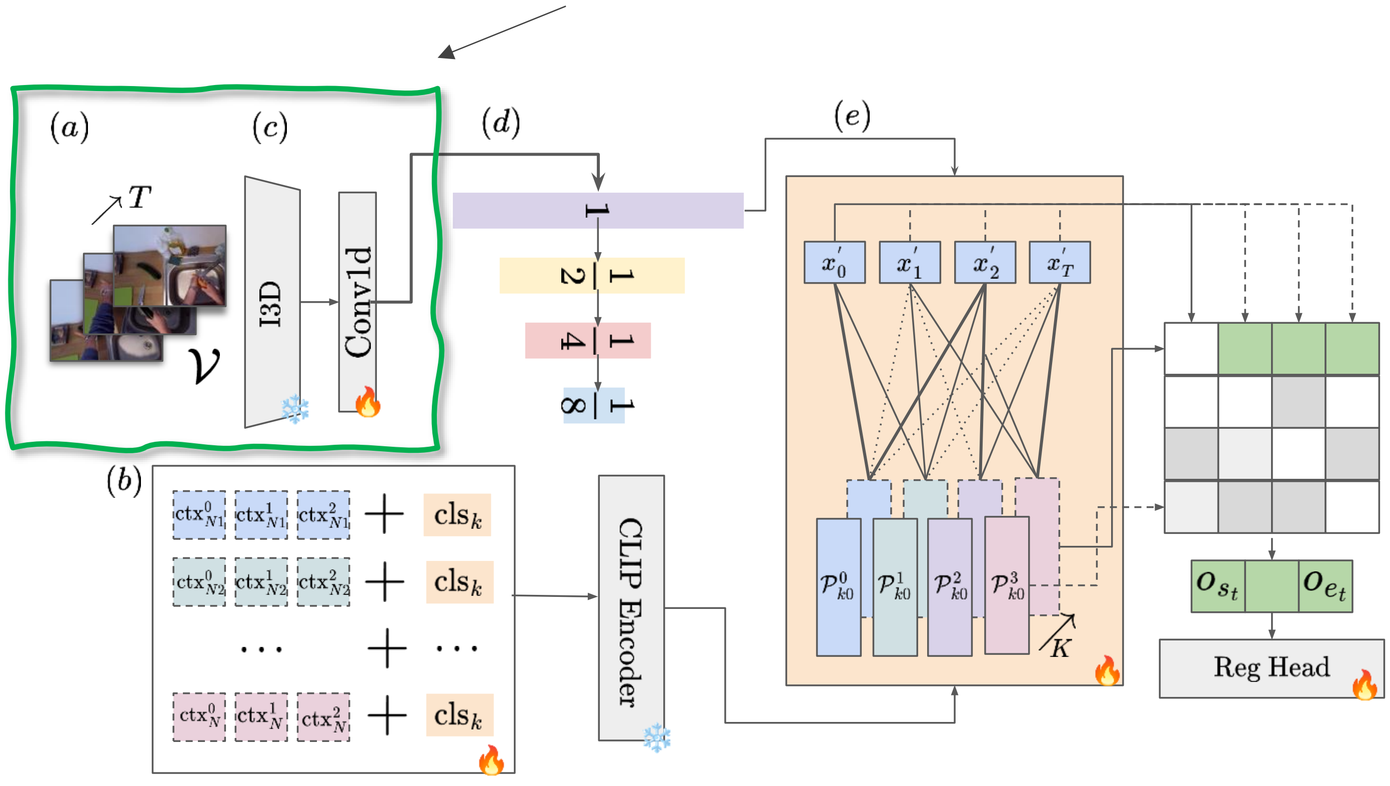
Methodology

- To do so, we use Optimal Transport as a method for distributing learnable prompts among all visual features over multiple temporal resolutions.



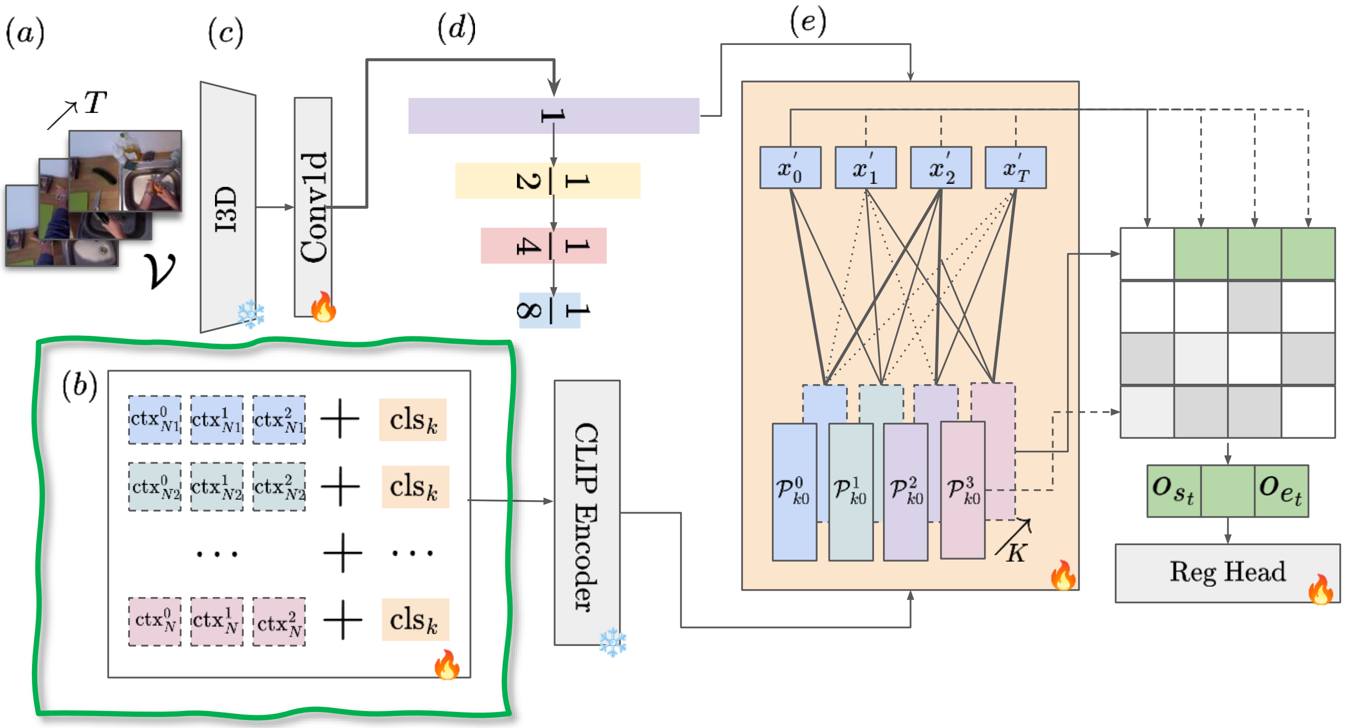
Methodology

(A–C) We first extract T frames from a video *V* using a frozen I3D encoder pretrained on Kinetics and train a Conv1D layer adapter to align these features with our CLIP text embeddings.



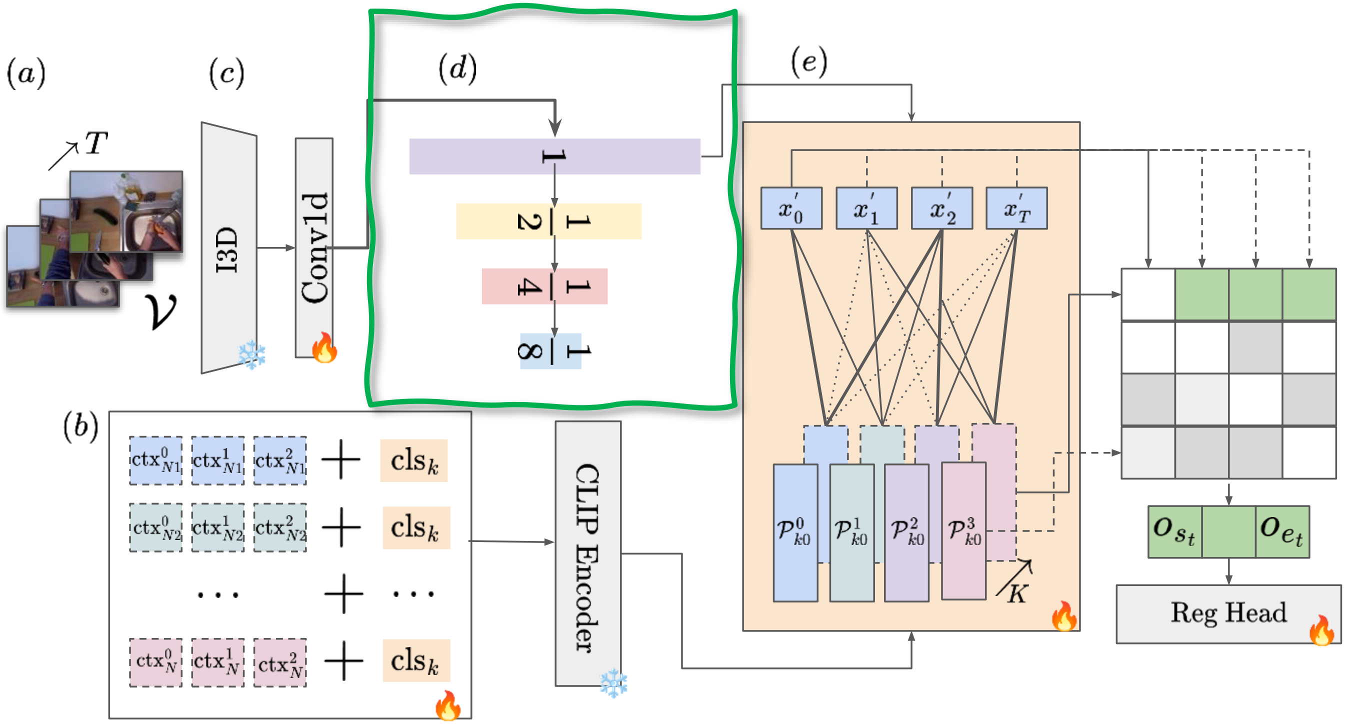
Methodology

(B) We randomly initialise N learnable prompts for each class K in the data.



Methodology

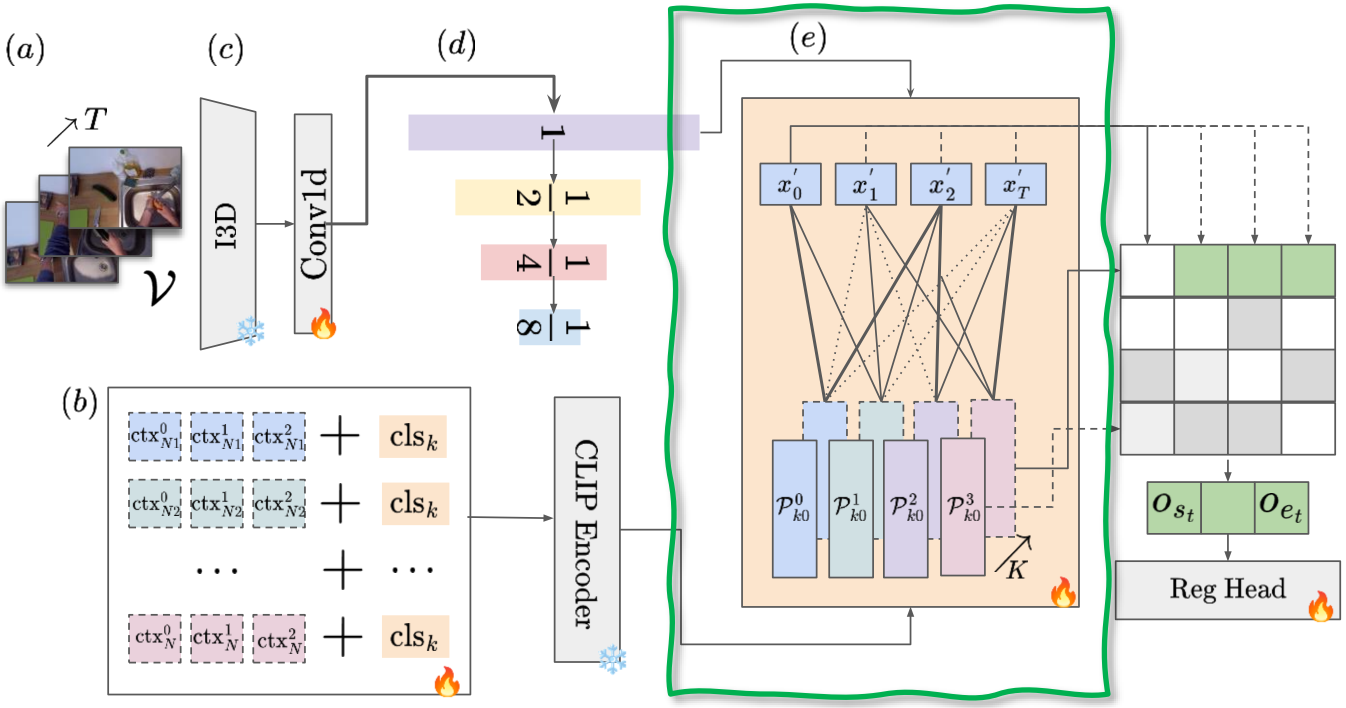
(D) We use average pooling to down sample visual features via a temporal feature pyramid creating an array of features for each temporal resolution.



Methodology

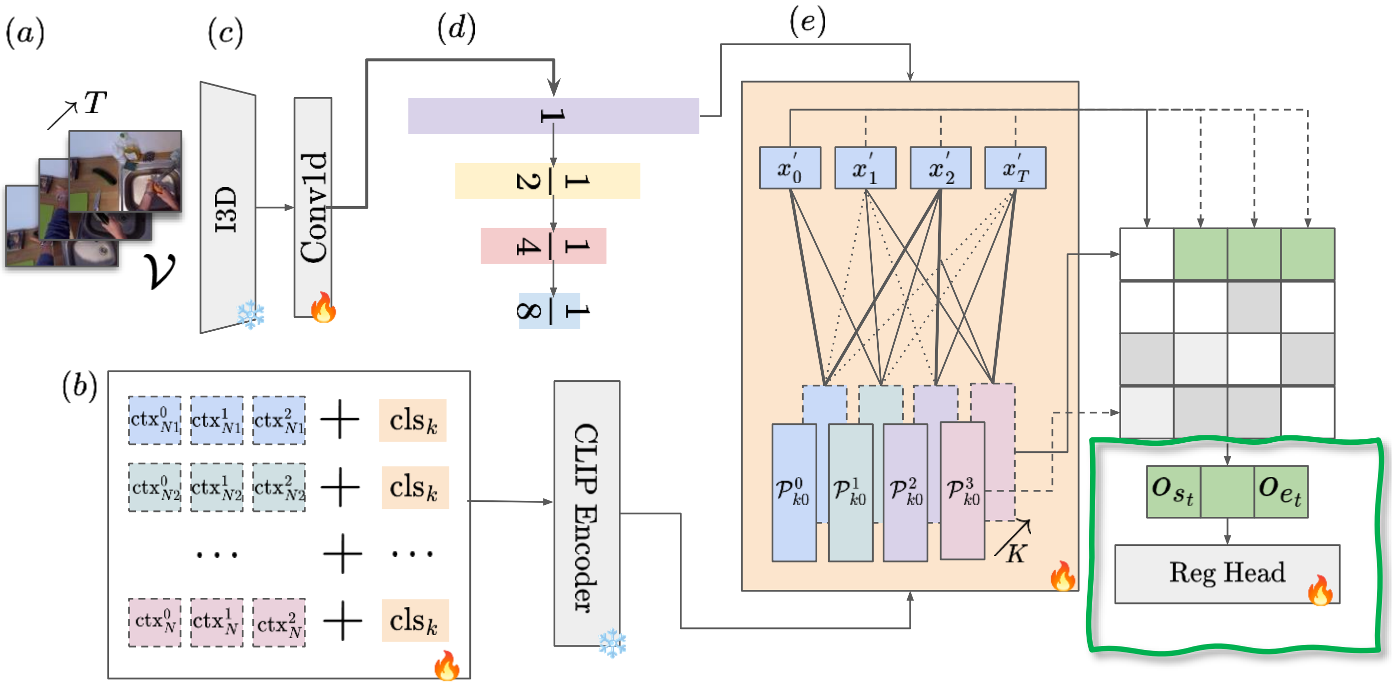
(E) We then compute the optimal transport plan between each temporal feature and prompt, across all pyramid levels.

The key here is that Optimal Transport algorithm includes entropic regularisation (Sinkhorn-Knopp algorithm) which ensures that the assignments between prompts and visual features are well distributed preventing collapse of prompts to one single visual feature.



Methodology

Finally, a regression and classification head are used to predict start and end times and class labels. The transport plan is fixed, and gradients back-propagate to the visual encoder and learnable prompts.

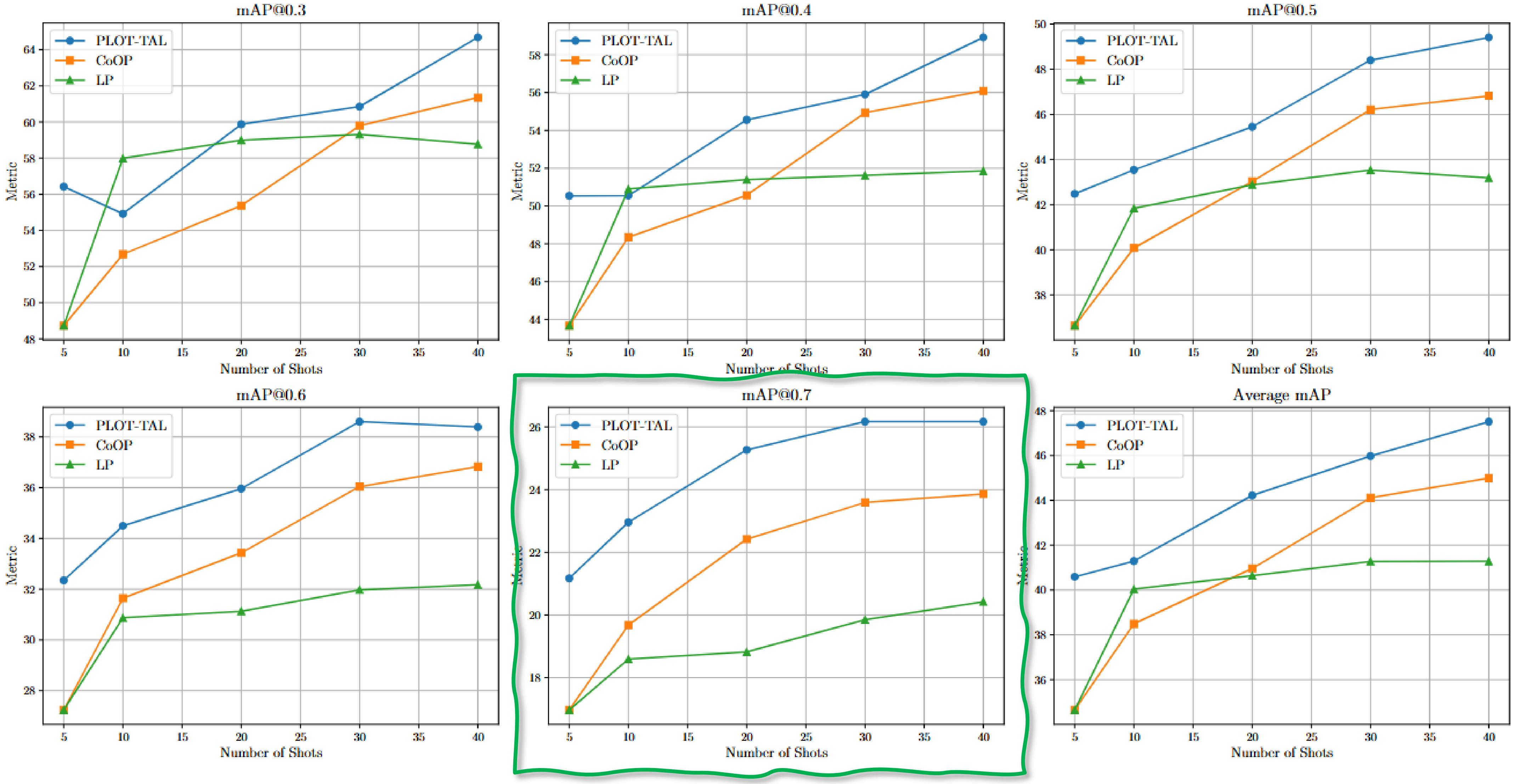


Here we can see the alignment for each learnable prompt using the transport policy. Prompts are aligned with different features across the video including the stadium and players. Cricket shot prediction is shown in green.

Results

Method	Approach	Avg. mAP (%)
Meta-Learning Approaches (5-shot, 5-way)		
Common Action Loc. [30]	ML	22.8
MUPPET [17]	ML + PL	24.9
Multi-Level Align. [10]	ML	31.8
Q. A. Transformer [16]	ML	32.7
End-to-End Prompt Learning (5-shot, 20-way)		
CoOp [35]	E2E + PL	34.65
PLOT-TAL (Ours)	E2E + PL	**38.24**
PLOT-TAL (Verbose) (Ours)	E2E + PL	**40.59**

Performance on THUMOS dataset. Note that we can train our model end to end (E2E) with Prompt Learning (PL) over all classes and perform better than Meta Learning Approaches.



PLOT-TAL is particularly effective at high IoU with very few samples and scales well with more examples.

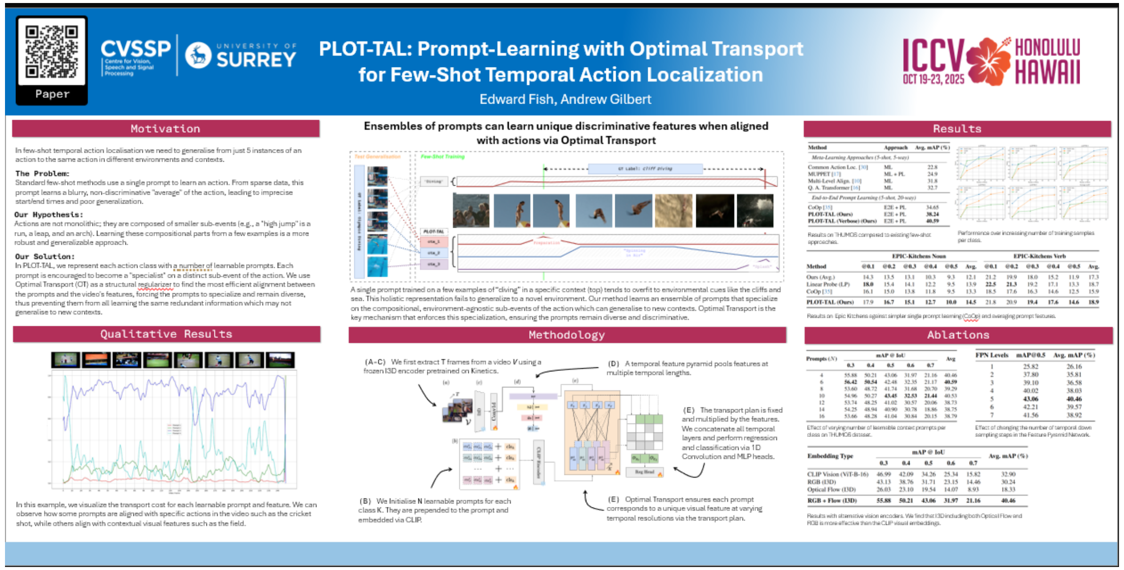
To summarise..

- Optimal transport is an effective way to align features in a few-shot setup for challenging tasks such as TAL.
- Enforcing the assignment to be smooth helps with generalisation (via entropic regularisation).
- We use small and efficient networks (I3D & CLIP) as a proof of concept – adapting this method to larger VLLM models would likely show even better performance.

SCAN ME

Thank You!

Edward Fish : Edward.fish@surrey.ac.uk
Andrew Gilbert: A.gilbert@surrey.ac.uk



Methodology

- Within the inner loop, entropic regularization acts as a crucial 'softening' factor, preventing the model from making overly rigid assignments and guiding the Sinkhorn algorithm to find a more stable, distributed transport plan.
- The transport plan is computed in an internal optimization loop during training during for each forward pass.

Algorithm 1 PLOT-TAL Optimization Loop

Input: Video features $\{\mathbf{F}_l\}_{l=1}^{L}$, class labels $\{c\}$
Output: Optimized context vectors $\{\text{ctx}\}$
Initialize learnable context vectors $\{\text{ctx}\}$
for each training iteration **do**
  for each class c and pyramid level l **do**
    Generate prompt embeddings $\mathbf{G}_c \in \mathbb{R}^{N \times D}$
    Calculate cost matrix $\mathbf{C}_{l,c} = 1 - \mathbf{F}_l \mathbf{G}_c^\top$
    //— Inner Loop: Sinkhorn Algorithm —
    Initialize $\mathbf{v} \leftarrow \mathbf{1}/N$
    for $t_{in} = 1$ to T_{in} **do**
      $\mathbf{u} \leftarrow \mathbf{1}/(\exp(-\mathbf{C}_{l,c}/\lambda)\mathbf{v})$
      $\mathbf{v} \leftarrow \mathbf{1}/(\exp(-\mathbf{C}_{l,c}/\lambda)^\top \mathbf{u})$
    end for
    Compute transport plan $\mathbf{T}^*_{l,c}$ from $\mathbf{u}, \mathbf{v}$
    Compute OT distance $d_{\text{OT}}(l,c) = \langle \mathbf{T}^*_{l,c}, \mathbf{C}_{l,c} \rangle$
  end for
  //— Outer Loop —
  Compute final predictions using aligned features
  Compute total loss $\mathcal{L}_{\text{total}}$ (Eq. 4)
  Backpropagate gradients from $\mathcal{L}_{\text{total}}$ to update $\{\text{ctx}\}$
end for
return Optimized context vectors $\{\text{ctx}\}$