



UNIVERSITY OF
SURREY

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

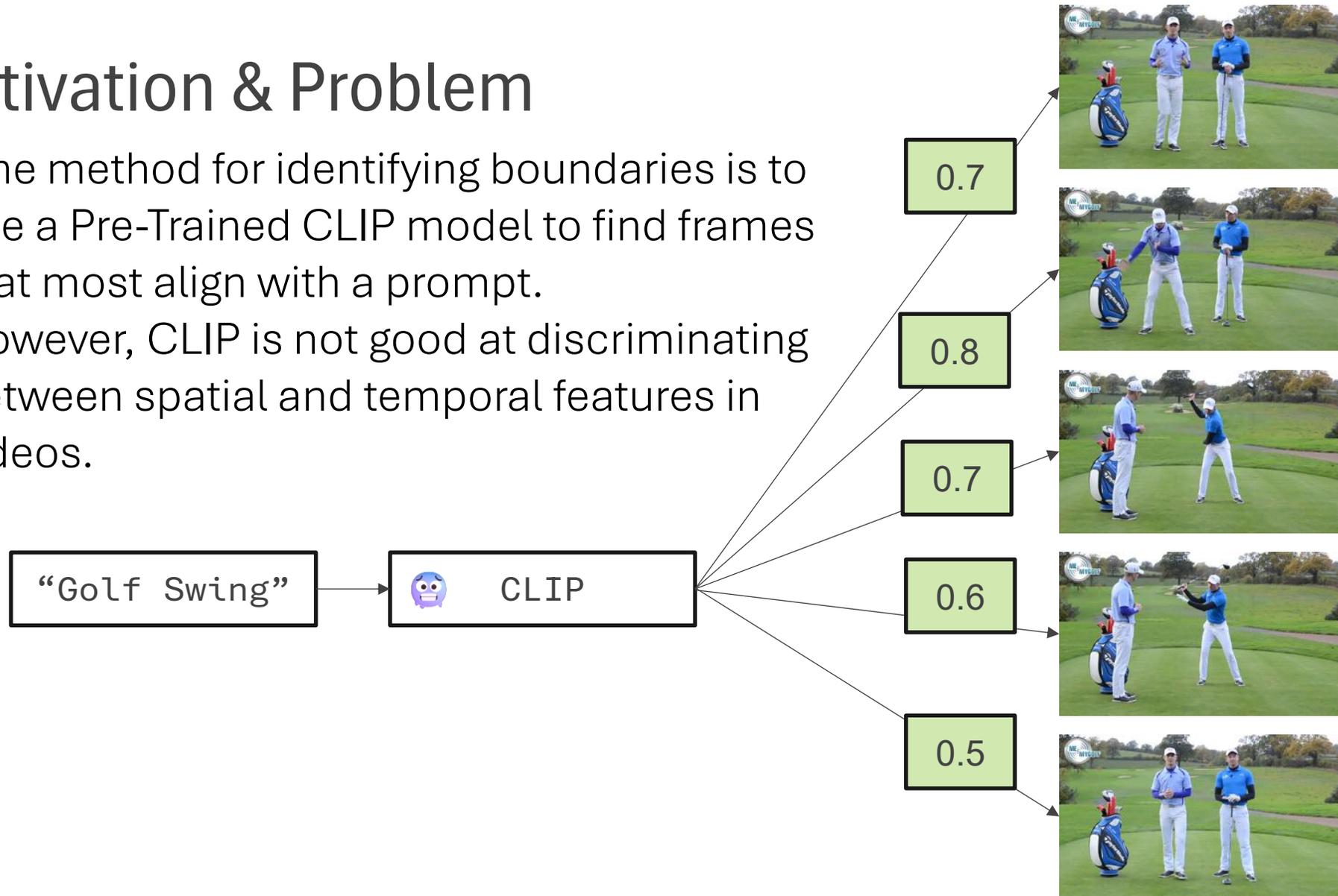


Motivation

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.

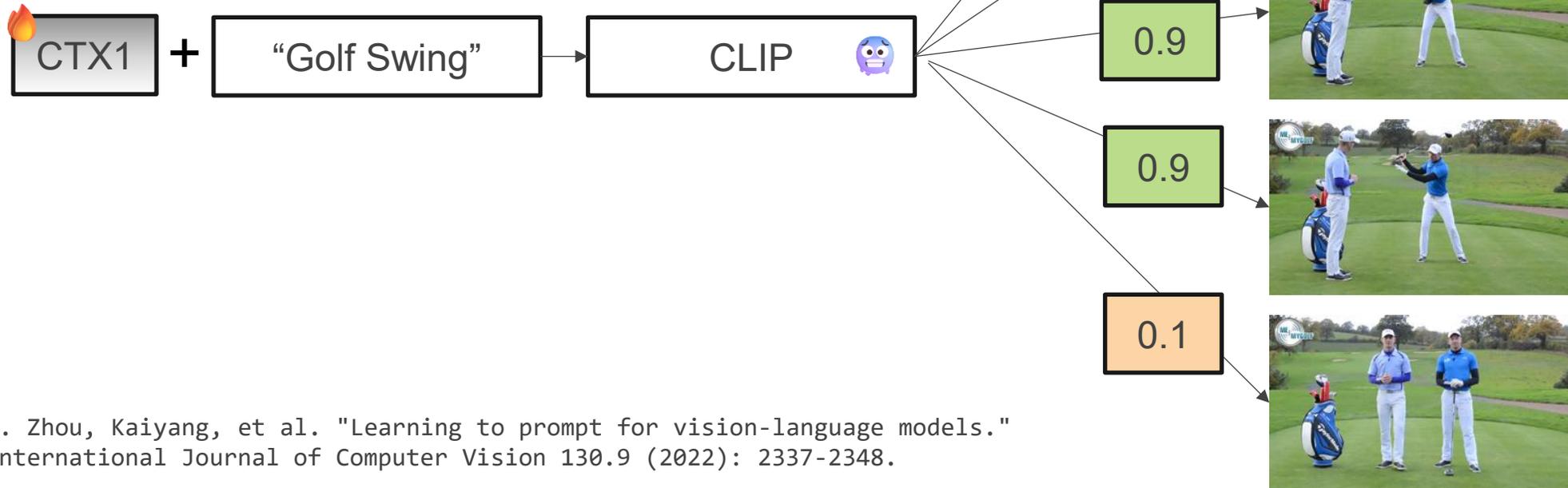
Motivation & Problem

- 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

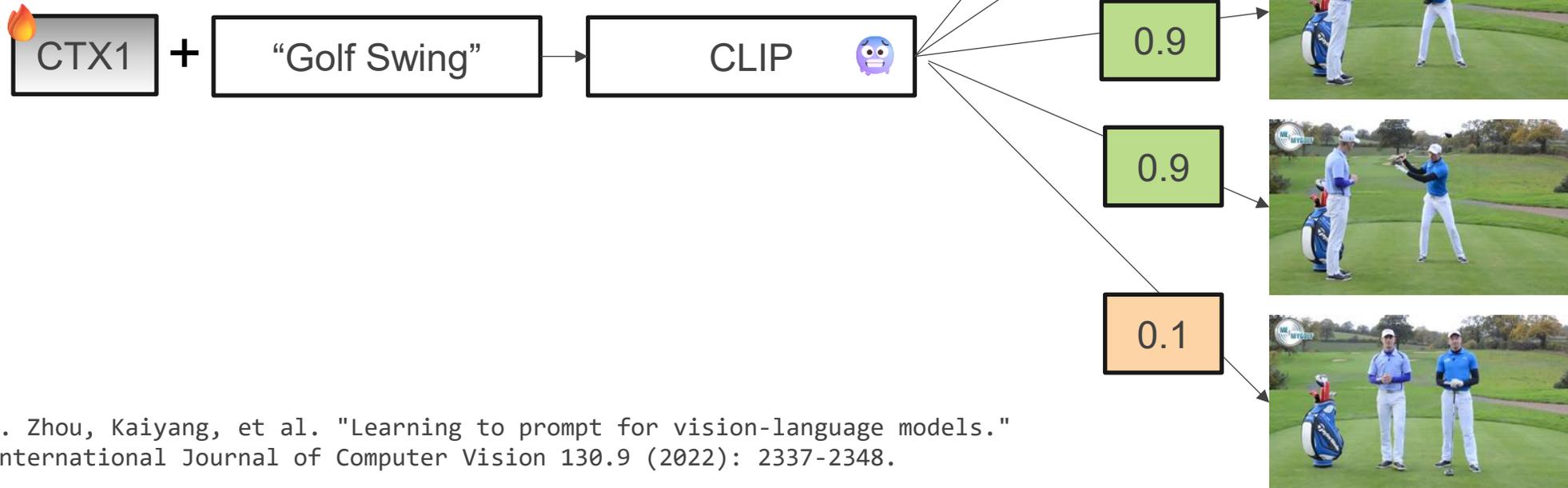
- Adding learnable contextual prompts is one way to efficiently adapt the text embedding to focus on salient actions for localization.



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.

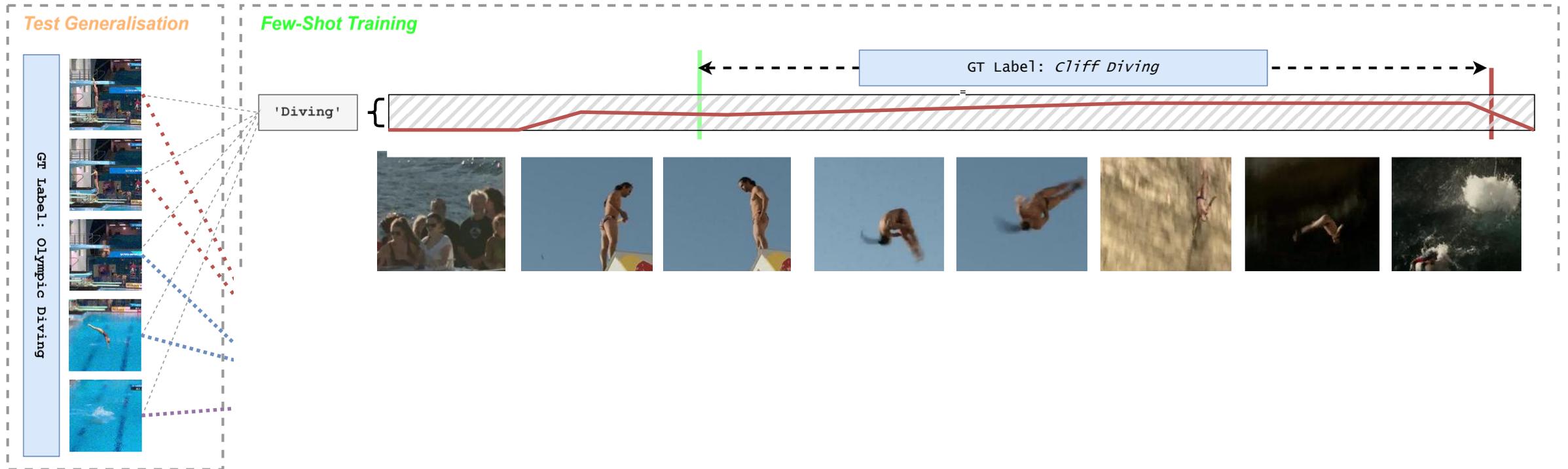
Motivation & Problem

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



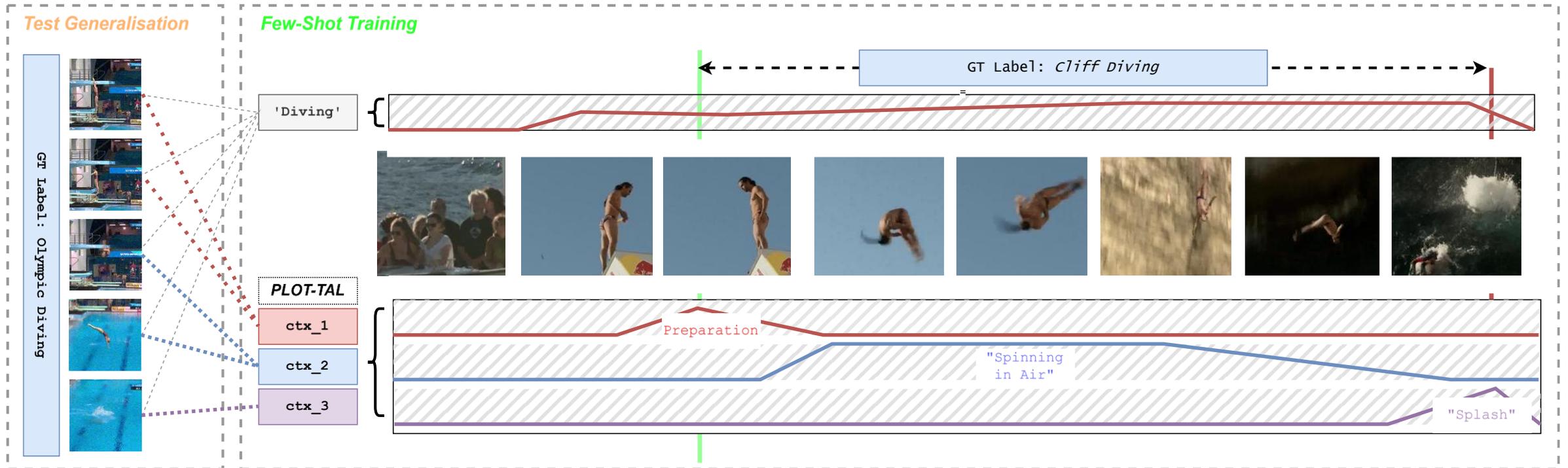
Motivation & Problem

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



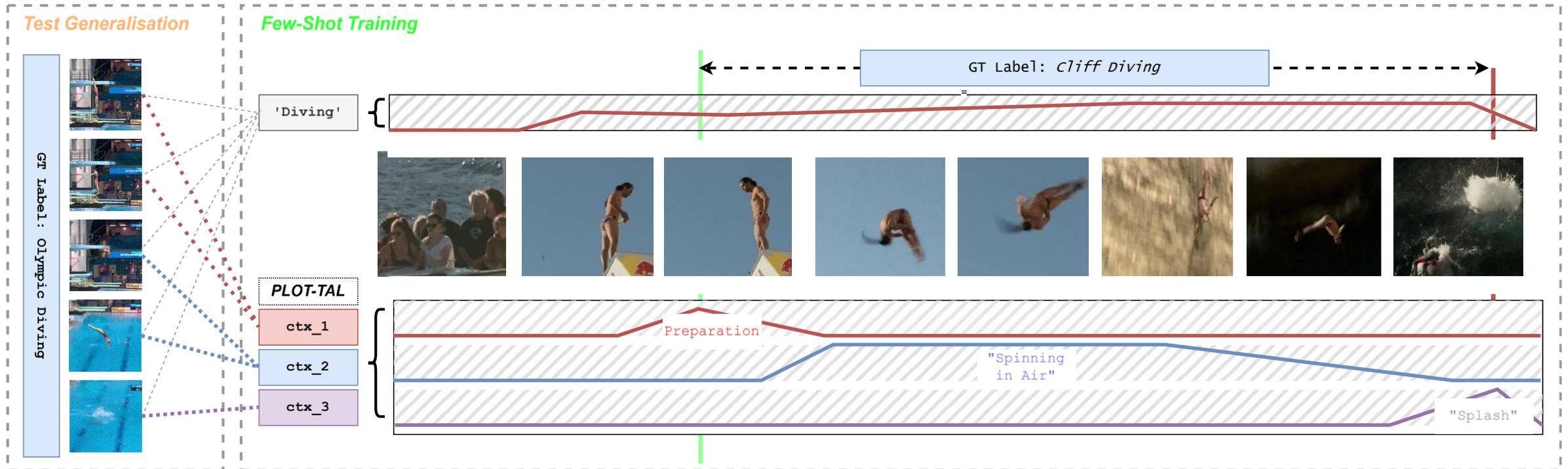
Motivation & Problem

- 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.



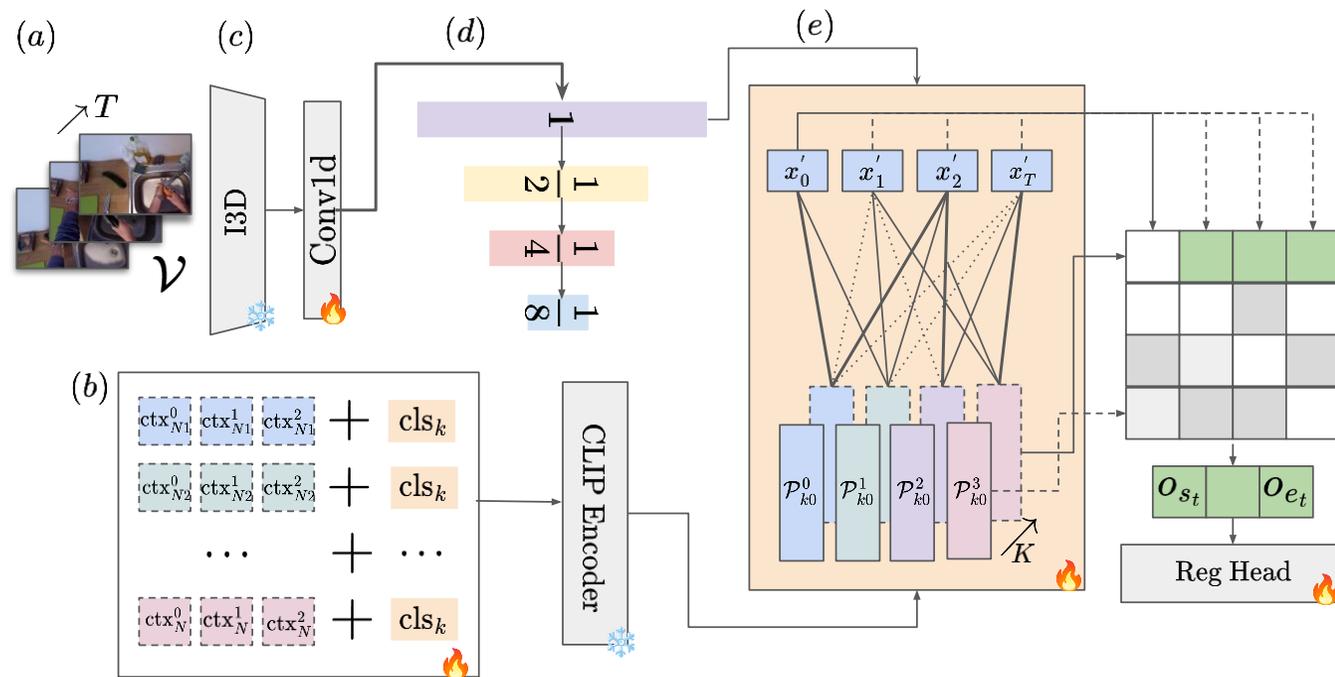
Motivation & Problem

- 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.



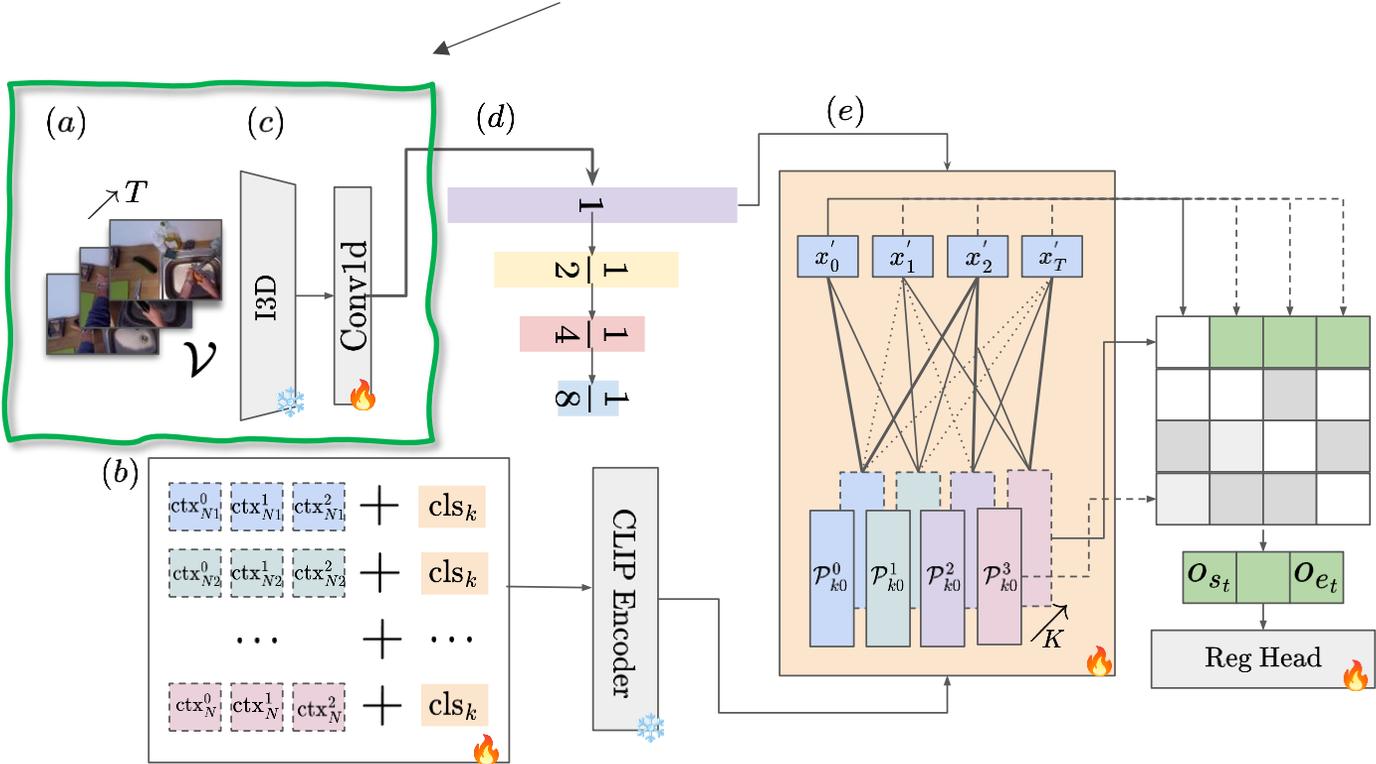
Methodology

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



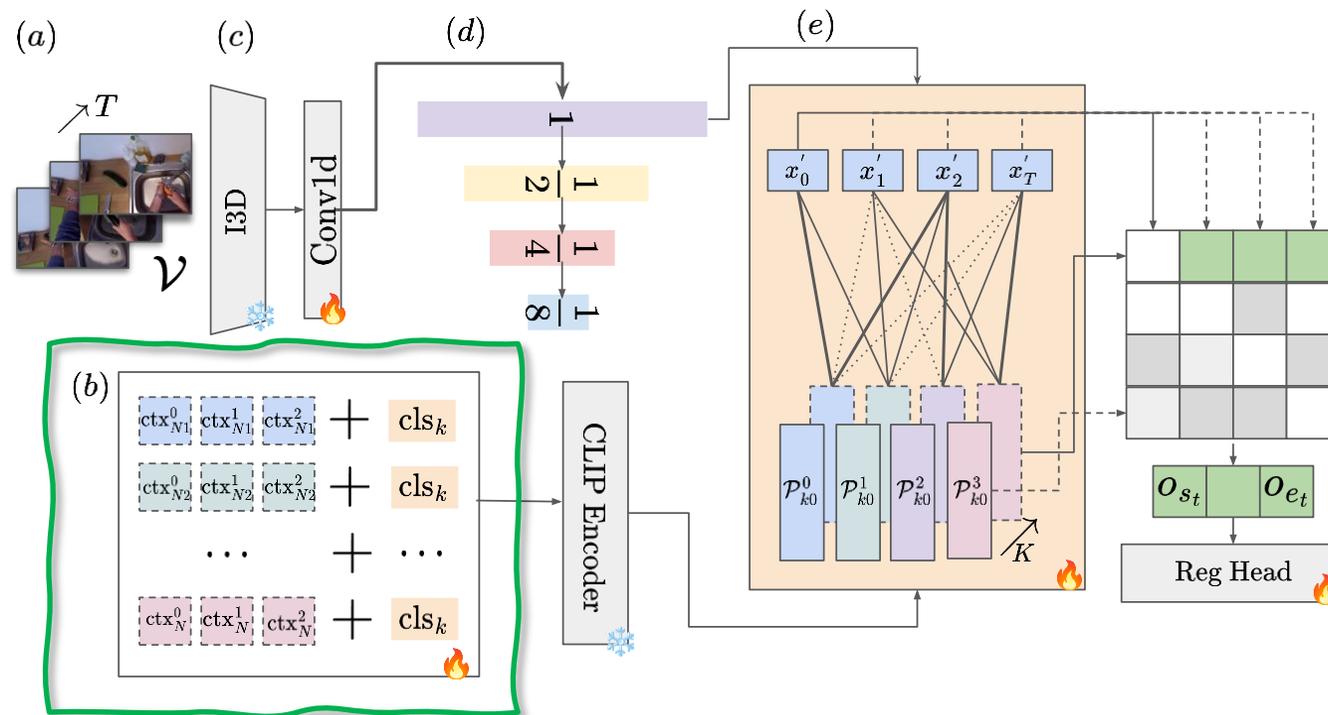
Methodology

(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.



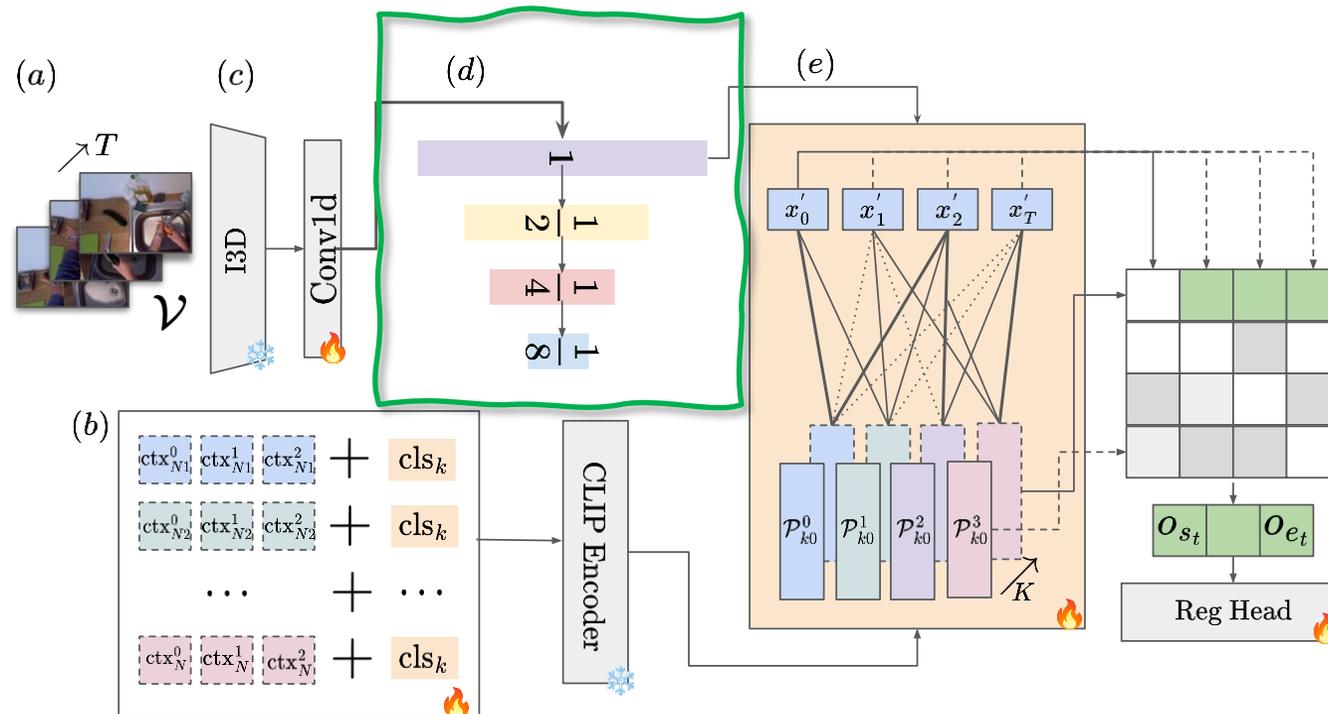
Methodology

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



Methodology

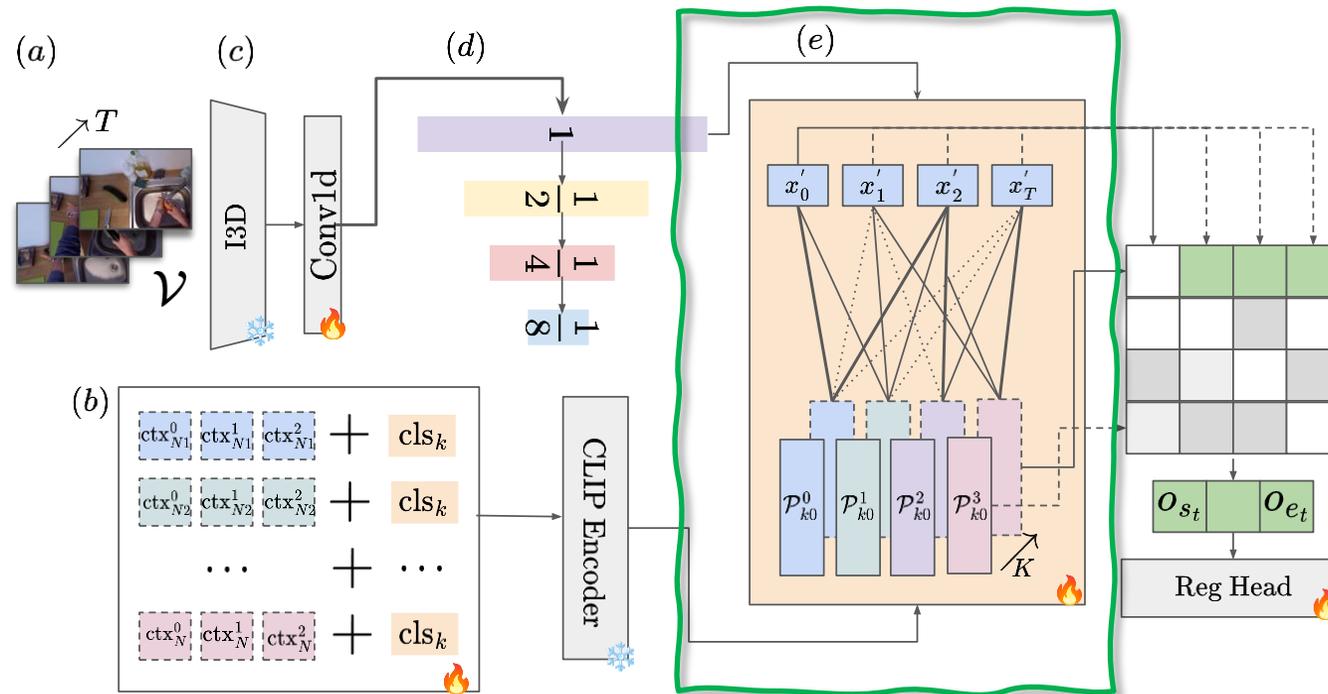
(D) We use average pooling to down sample visual features via a temporal feature pyramid creating an array of features for each temporal resolution.



Methodology

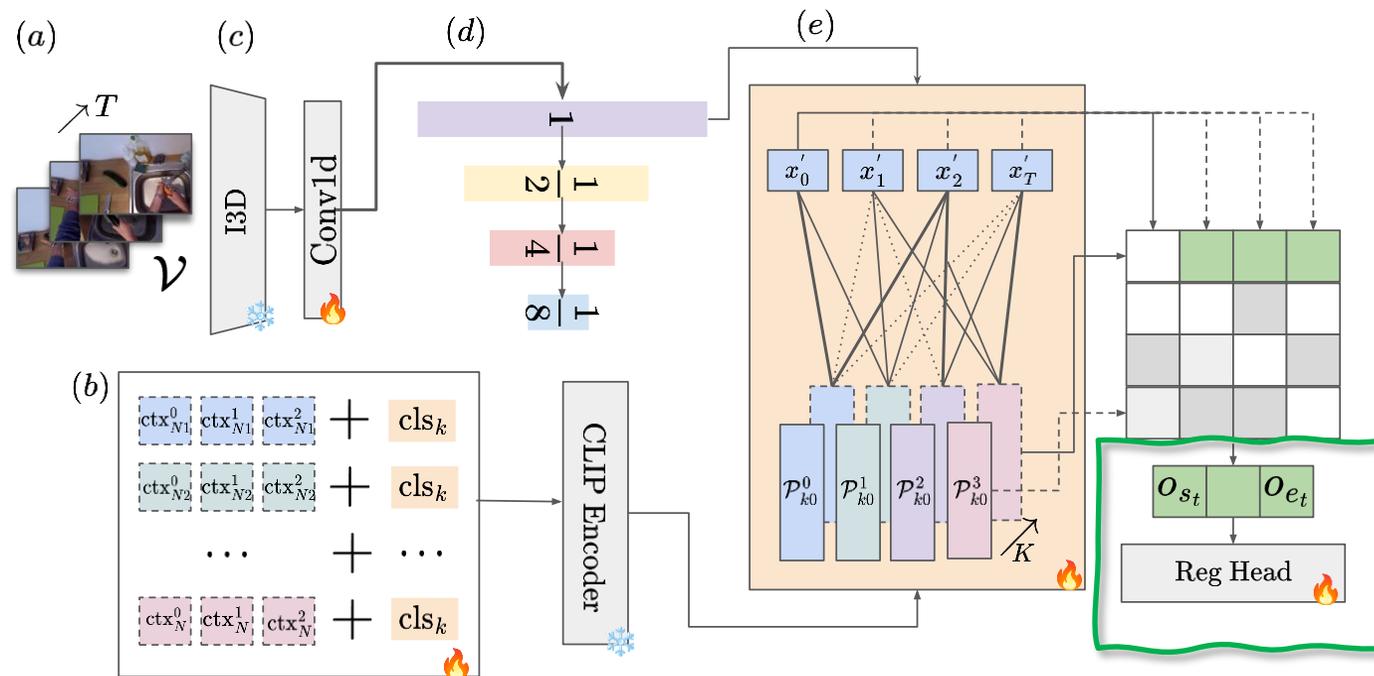
The key here is that Optimal Transport algorithm includes entropic regularisation (Sinkhorn-Knopp algorithm) which ensures that the assignments between prompts and visual features are well distributed preventing collapse of prompts to one single visual feature.

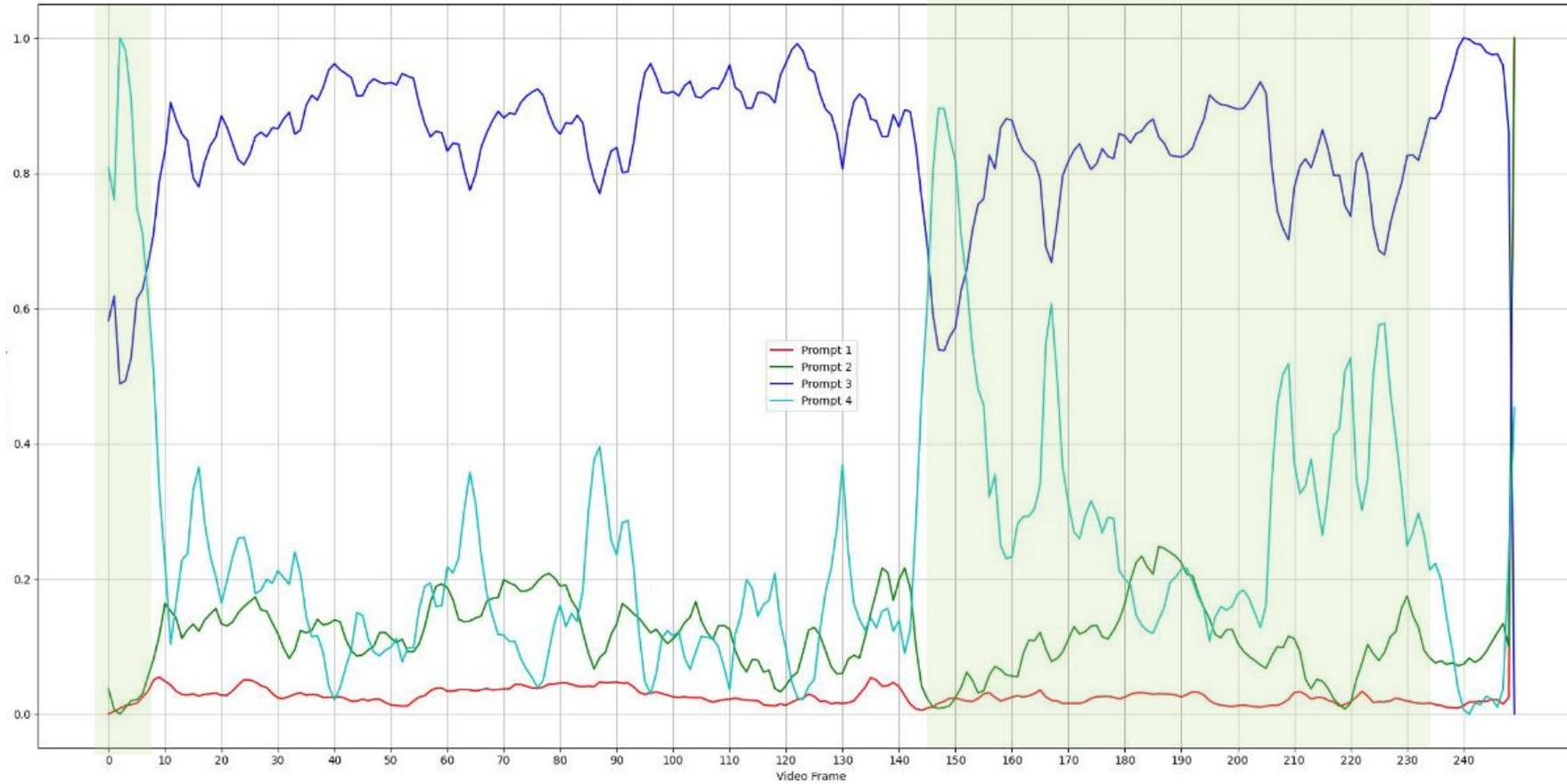
(E) We then compute the optimal transport plan between each temporal feature and prompt, across all pyramid levels.



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.



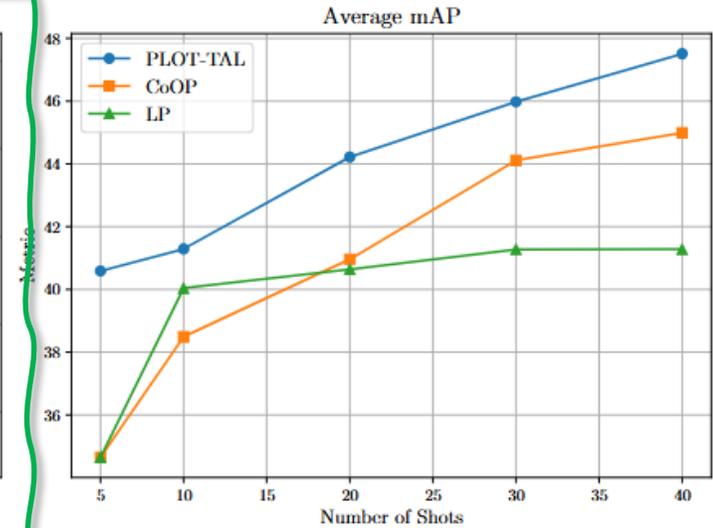
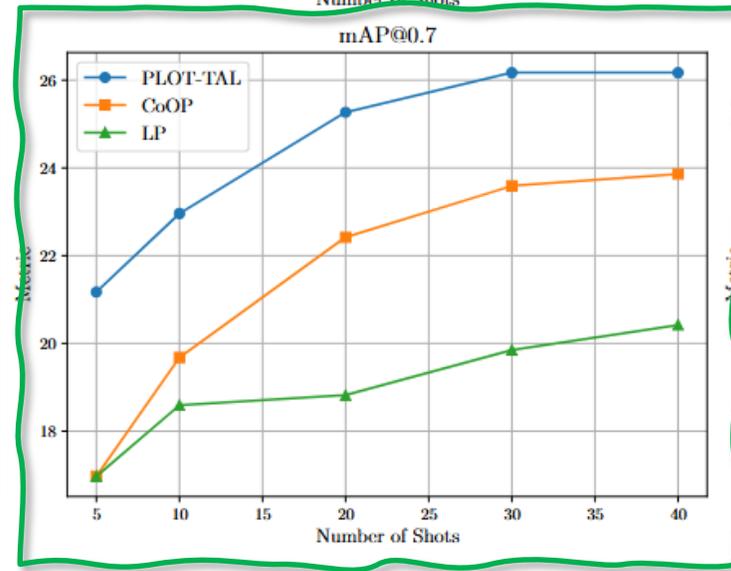
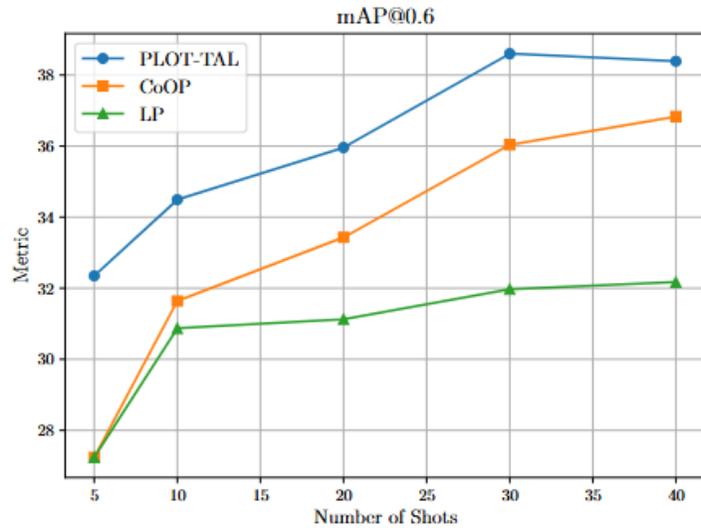
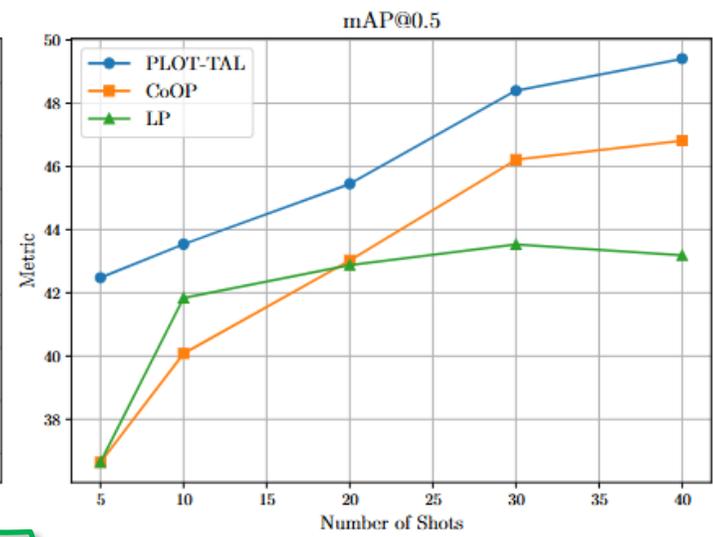
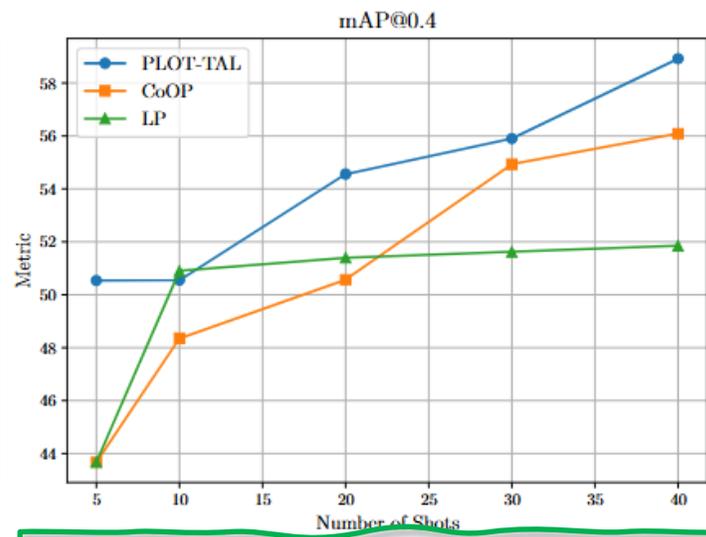
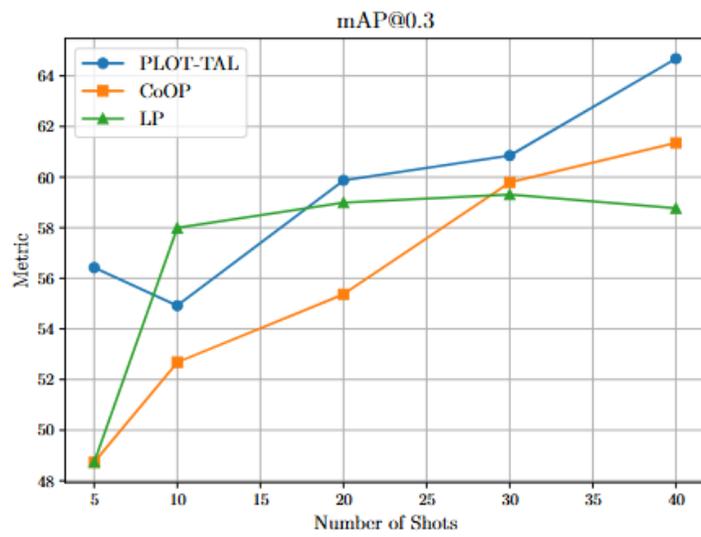


Here we can see the alignment for each learnable prompt using the transport policy. Prompts are aligned with different features across the video including the stadium and players. Cricket shot prediction is shown in green.

Results

Method	Approach	Avg. mAP (%)
<i>Meta-Learning Approaches (5-shot, 5-way)</i>		
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
<i>End-to-End Prompt Learning (5-shot, 20-way)</i>		
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.



Thank You!

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



CVSSP
Centre for Vision,
Speech and Signal
Processing



UNIVERSITY OF SURREY

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

Edward Fish, Andrew Gilbert



ICCV HONOLULU HAWAII
OCT 19-23, 2025

Motivation

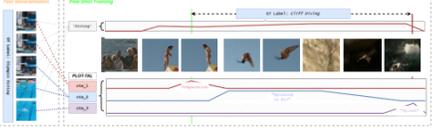
In few-shot temporal action localisation we need to generalise from just 5 instances of an action to the same action in different environments and contexts.

The Problem: Standard few-shot methods use a single prompt to learn an action. From sparse data, this prompt learns a blurry, non-discriminative "average" of the action, leading to imprecise standard errors and poor generalization.

Our Hypotheses: Actions are not monolithic; they are composed of smaller sub-events (e.g., a "high jump" is a run, a leap, and an arch). Learning these compositional parts from a few examples is a more robust and generalizable approach.

Our Solution: In PLOT-TAL, we represent each action class with a number of learnable prompts. Each prompt is encouraged to become a "specialist" on a distinct sub-event of the action. We use Optimal Transport (OT) as a structural regularizer to find the most efficient alignment between the prompts and the video's features, forcing the prompts to specialize and remain diverse, thus preventing them from all learning the same redundant information which may not generalise to new contexts.

Ensembles of prompts can learn unique discriminative features when aligned with actions via Optimal Transport



A single prompt trained on a few examples of "running" in a specific context (top) tends to overfit to environmental cues like the cliffs and sea. This holistic representation fails to generalise to a novel environment. Our method learns an ensemble of prompts that specialise on the compositional, environment-agnostic sub-events of the action which can generalise to new contexts. Optimal Transport is the key mechanism that enforces this specialization, ensuring the prompts remain diverse and discriminative.

Results

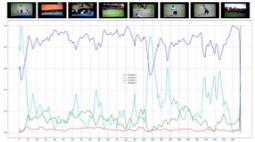
Method	Approach	Arg. mAP (%)
Video Learning, Supervised (Full, Train)		
Common Action Loc. (17)	ML	22.8
MCNN (11)	ML	24.9
Multi-Level Mags. (10)	ML	31.9
U-Net (Environment (1))	ML	52.7
Zero- and Few-Shot Learning (1-Awk, 20 ways)		
CLIP (1)	CLIP + PL	34.0
PLOT-TAL (Prom)	OT + PL	38.24
PLOT-TAL (Network-Share)	OT + PL	40.09

Results on 14 AVCSs compared to existing few-shot approaches. Performance (see increasing number of training samples per class).

Method	EPC-Kinetics Train				EPC-Kinetics Test			
	0#1	0#2	0#4	0#8	0#1	0#2	0#4	0#8
Open (Arg)	14.3	13.3	13.1	10.3	9.3	12.1	21.2	29.9
Linear Probe (LP)	18.0	15.4	14.4	12.2	9.5	13.9	22.8	28.9
CLIP (1)	16.1	13.0	13.8	11.4	9.5	13.3	18.5	17.6
PLOT-TAL (Prom)	17.9	16.7	18.1	12.7	10.0	14.8	21.8	29.9
PLOT-TAL (Net)	19.4	17.6	19.4	14.1	11.4	15.8	23.4	31.4

Results on open Kinetics-splitter single epoch prompt learning (CLIP) with varying prompt features.

Qualitative Results



In this example, we visualise the transport cost for each learnable prompt and feature. We can observe how some prompts are aligned with specific actions in the videos such as the cricketer shot, while others align with contextual visual features such as the field.

Methodology

(A-C) We first extract T frames from a video using a frozen DD encoder pretrained on Kinetics.



(D) A temporal feature pyramid pools features at multiple temporal lengths.

(E) The transport plan is fixed and multiplied by the features. We concatenate all temporal layers and perform regressors and classification via TD-CrossEntropy and MLP heads.

(B) We initialise M learnable prompts for each class C. They are prepended to the prompt and embedded via CLIP.

(E) Optimal Transport ensures each prompt corresponds to a unique visual feature at varying temporal resolutions via the transport plan.

Ablations

Prompts (N)	mAP @ 1s				EPC Levels	mAP @ 0.5s	Arg. mAP (%)
	K2	K3	K4	K5			
4	55.48	56.21	55.96	55.97	21.38	46.46	55.81
6	56.40	56.54	56.96	57.25	21.35	46.89	56.30
8	55.80	56.72	57.19	57.60	20.70	50.29	56.03
10	56.25	56.57	56.87	56.99	21.46	46.61	56.46
12	55.74	56.25	56.62	56.57	20.96	50.75	56.15
14	54.23	56.04	56.00	56.76	20.46	50.75	56.15
16	53.90	56.26	56.46	56.84	20.12	50.75	56.12

Effect of increasing number of learnable context prompts per class on 1-Awk PACE dataset.

Embedding Type	mAP @ 1s				EPC Levels	mAP @ 0.5s	Arg. mAP (%)
	K2	K3	K4	K5			
CLIP-Vision (ViT-B/16)	46.99	42.09	36.28	25.24	19.82	32.90	46.46
RGB (EDS)	43.13	38.76	31.74	23.15	14.46	30.24	46.46
Optical Flow (EDS)	26.63	21.89	18.54	14.07	4.93	18.53	46.46

Effect of changing the number of temporal down-sample steps in the Feature Pyramid Network.

RGB + Flow (EDS)	mAP @ 1s				EPC Levels	mAP @ 0.5s	Arg. mAP (%)
	K2	K3	K4	K5			
55.88	56.21	55.86	55.97	21.38	46.46	55.81	

Results with alternative video encoders. We test CLIP (including both Optical Flow and RGB) to investigate their effect on visual embeddings.

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

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