



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**  
OCT 19-23, 2025  **HONOLULU HAWAII**

**CVSSP** | Centre for Vision,  
Speech and Signal  
Processing

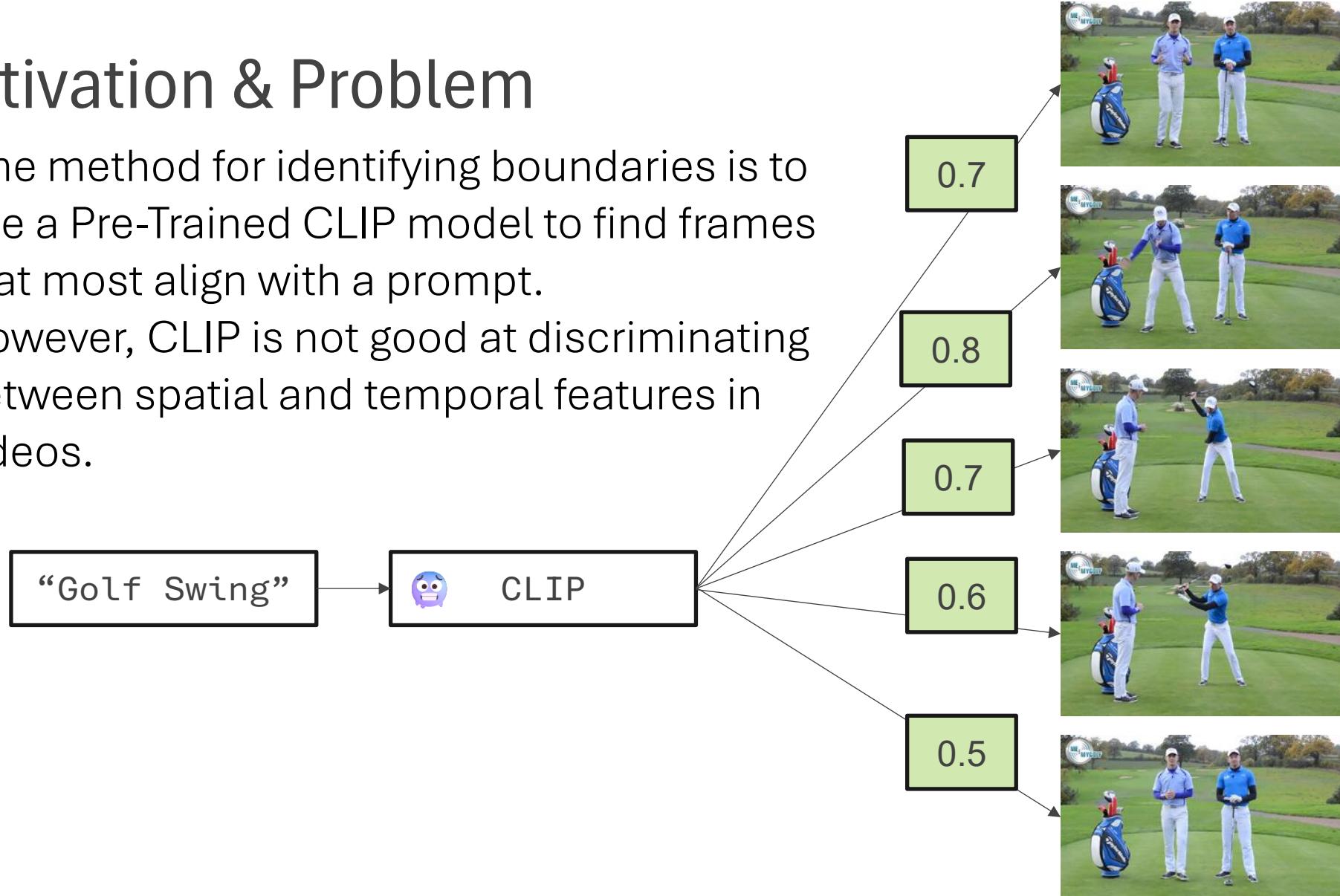


# Motivation

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.

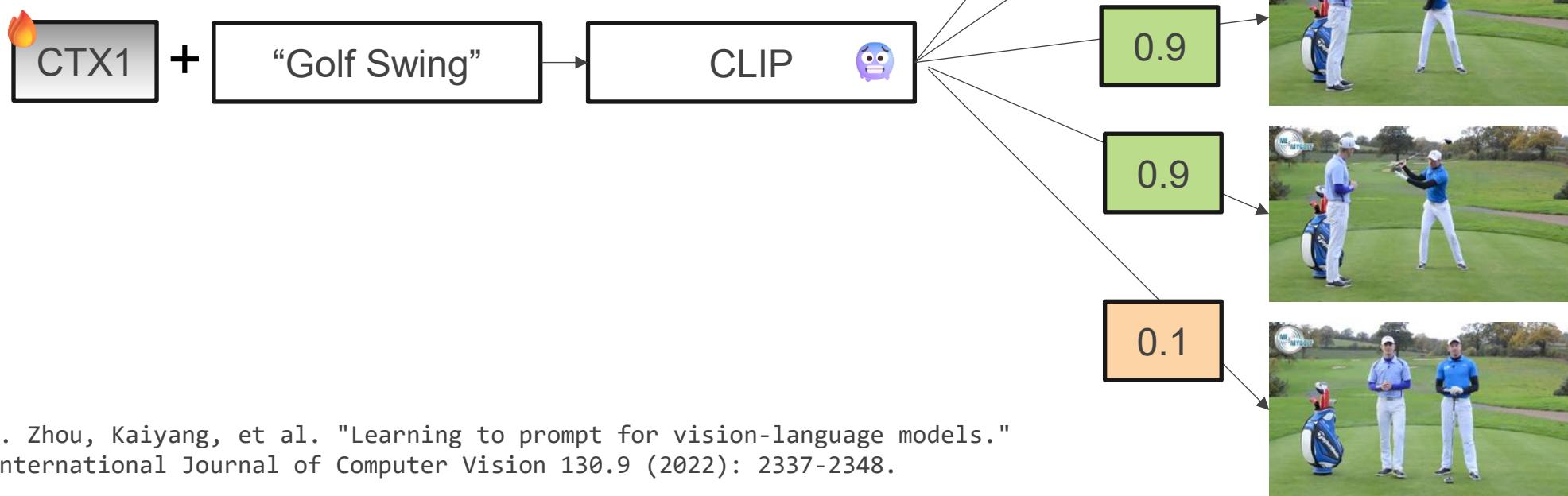
# Motivation & Problem

- 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.



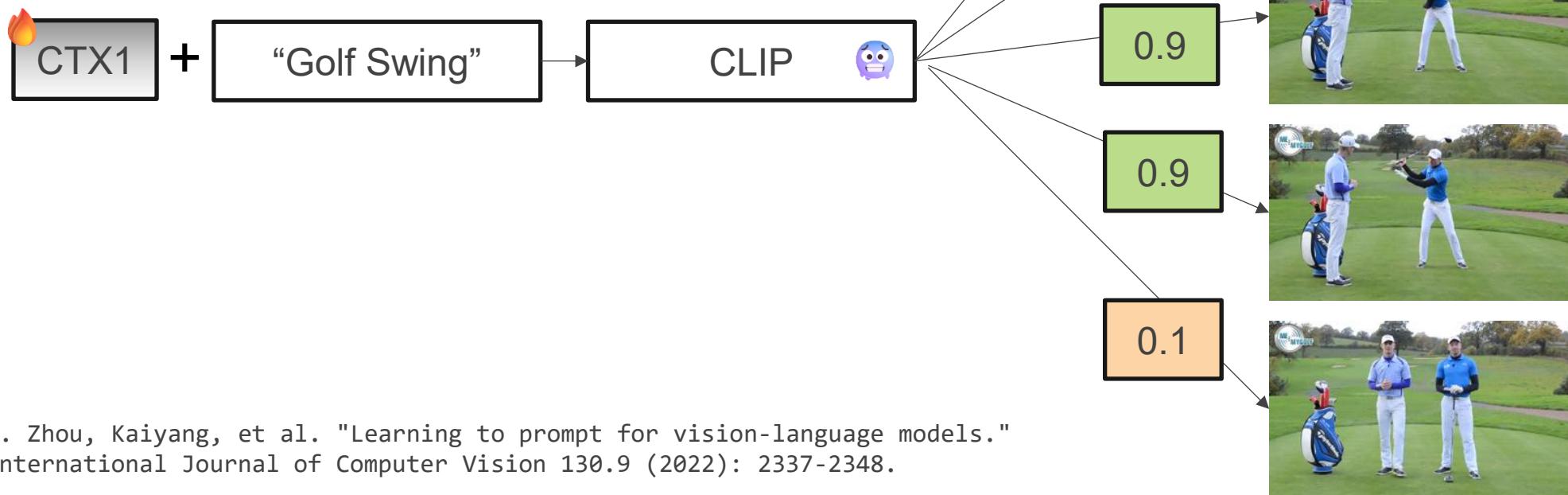
# Motivation & Problem

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



# Motivation & Problem

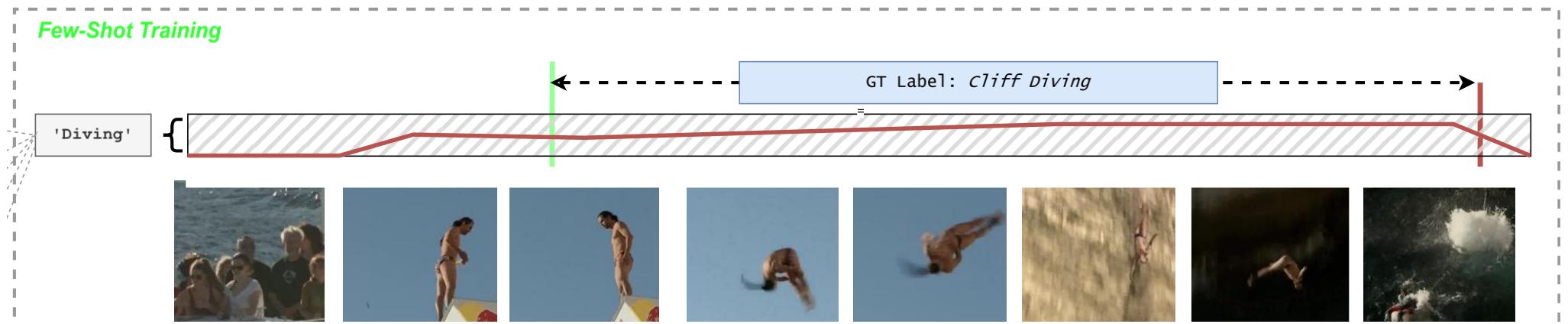
- We have no guarantees that these learnable prompts will learn generalisable features over just a few examples.



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

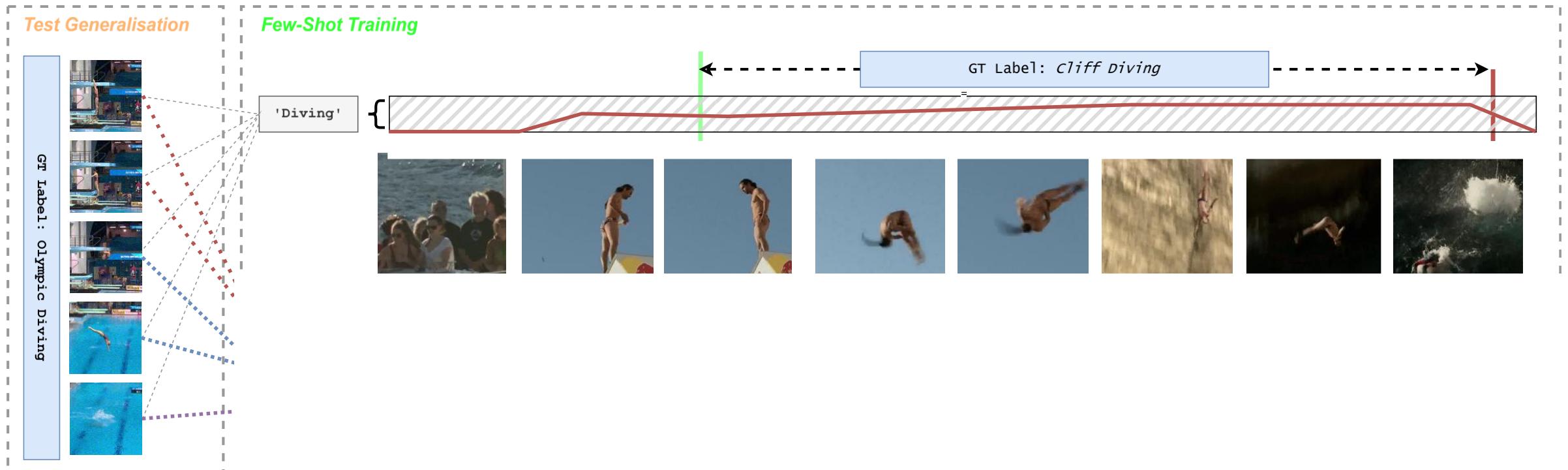
# Motivation & Problem

- 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.



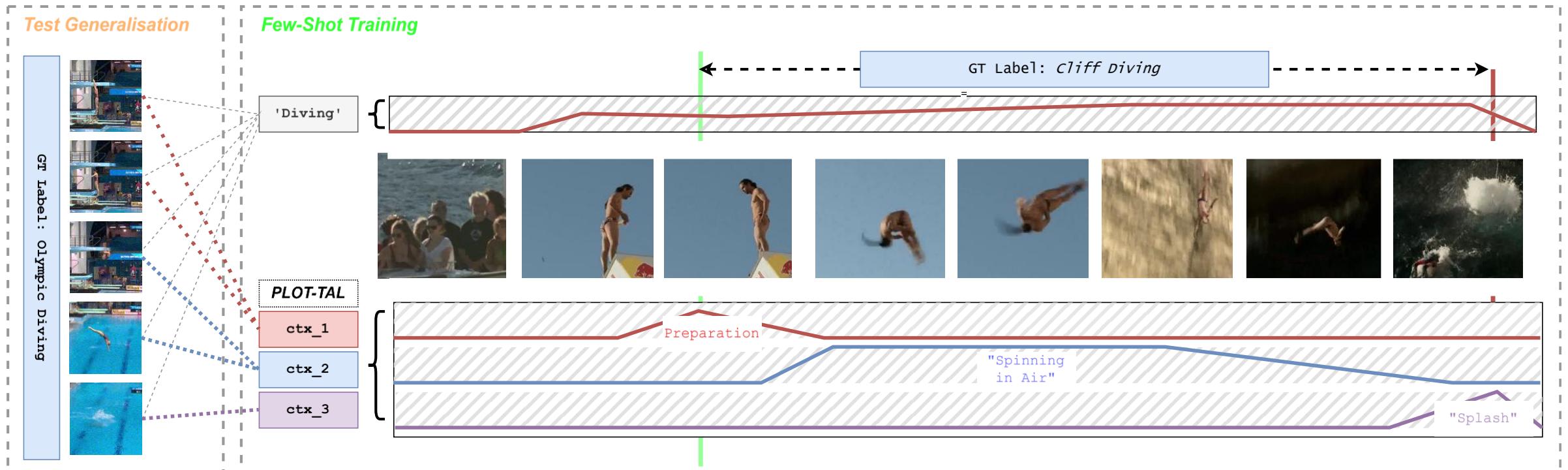
# Motivation & Problem

- 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.



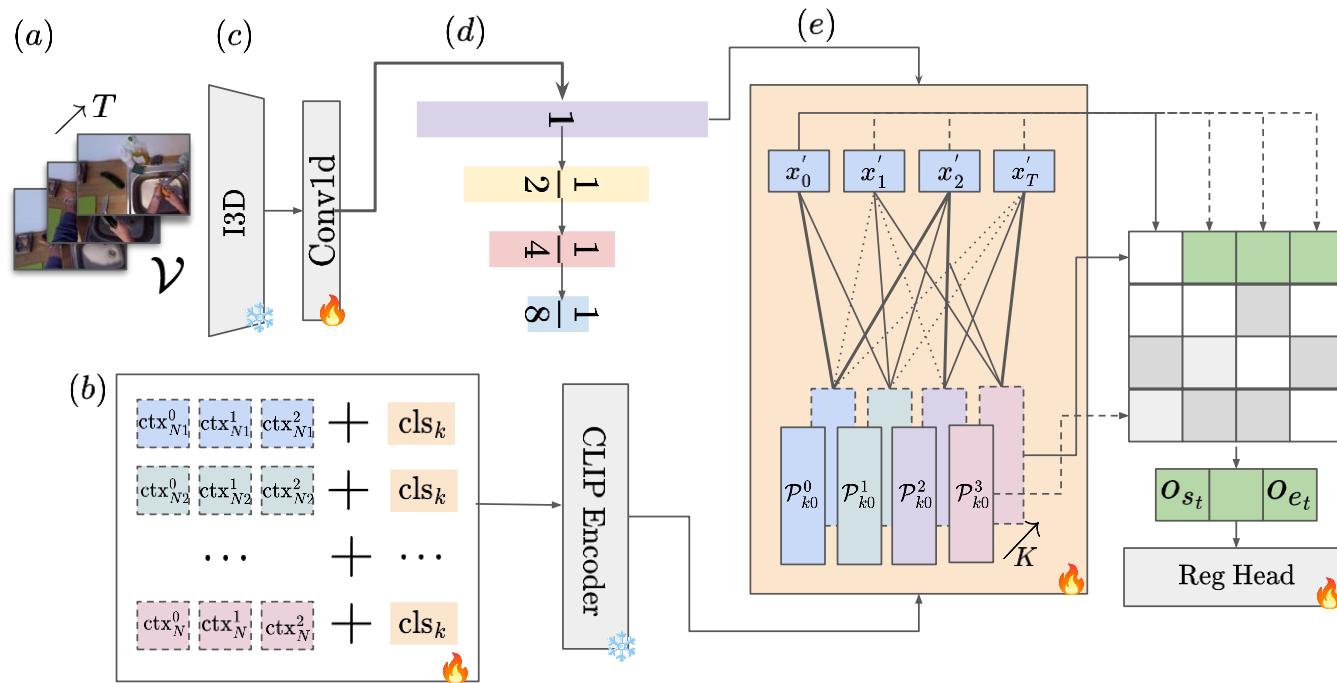
# Motivation & Problem

- 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.



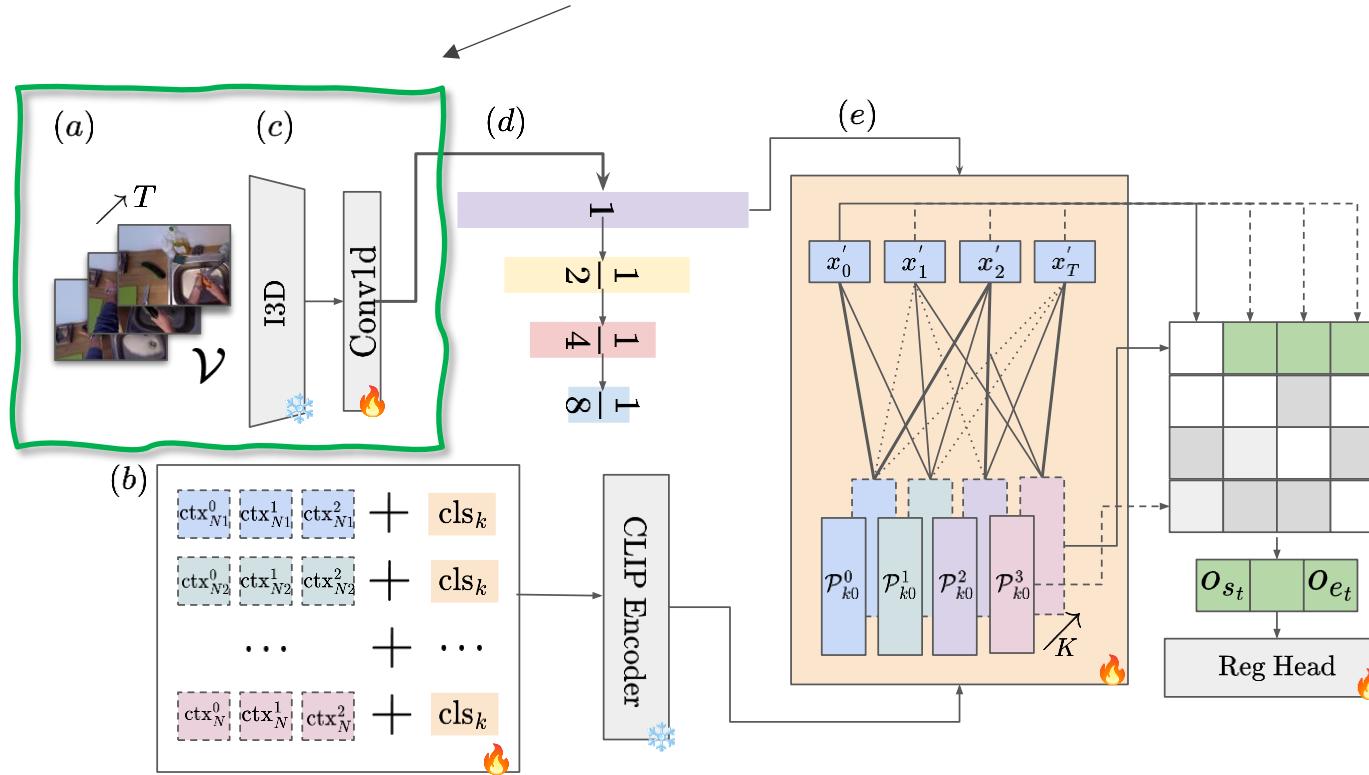
# Methodology

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



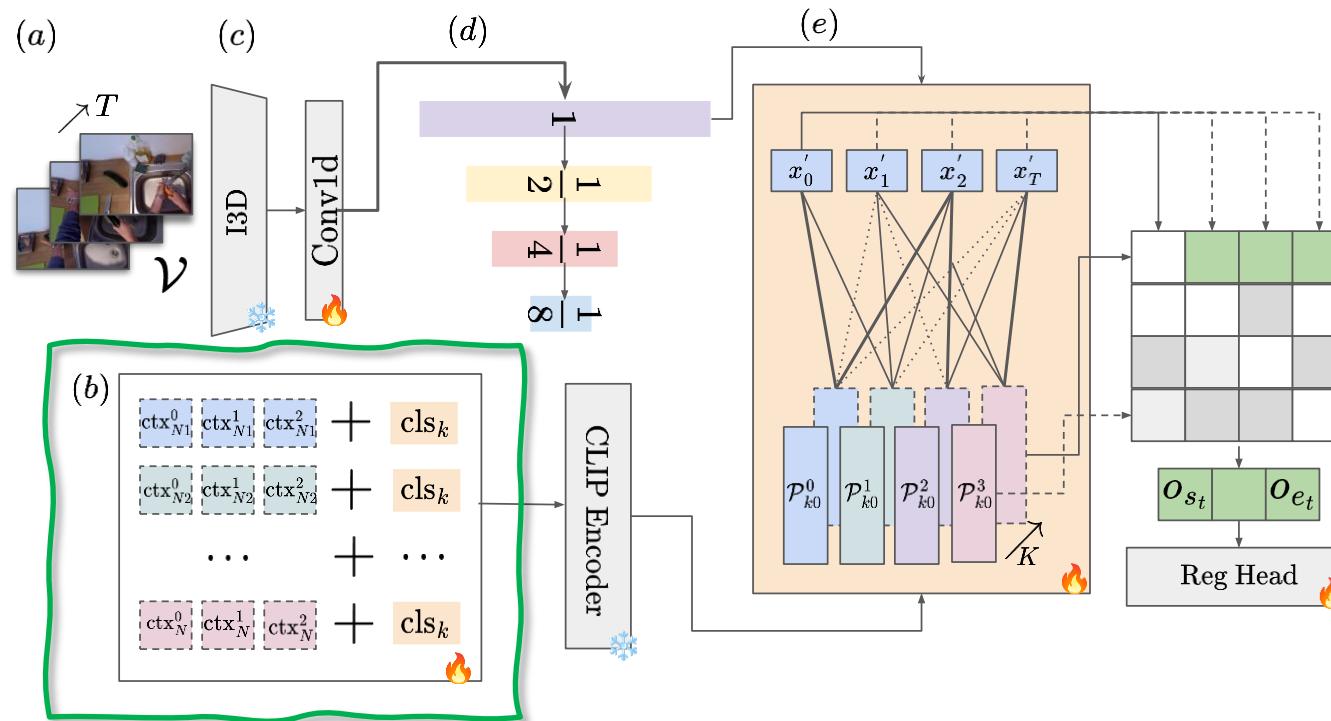
# Methodology

(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.



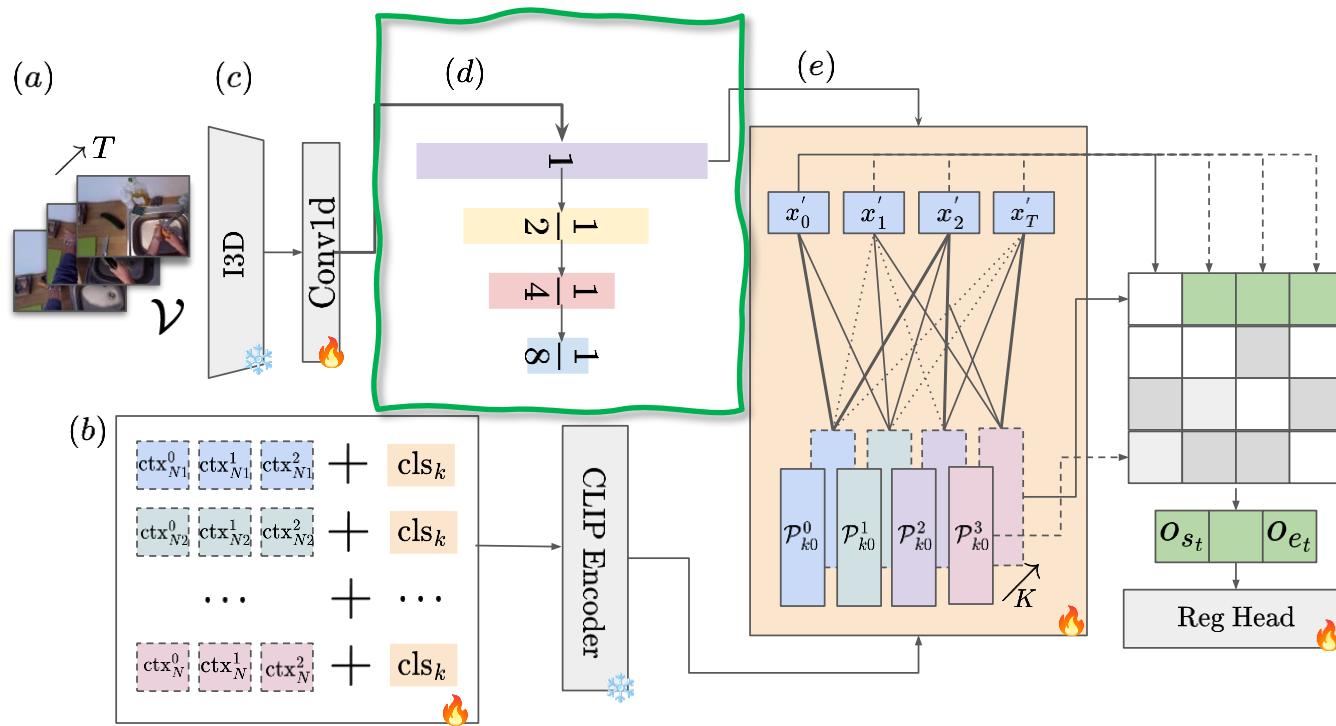
# Methodology

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



# Methodology

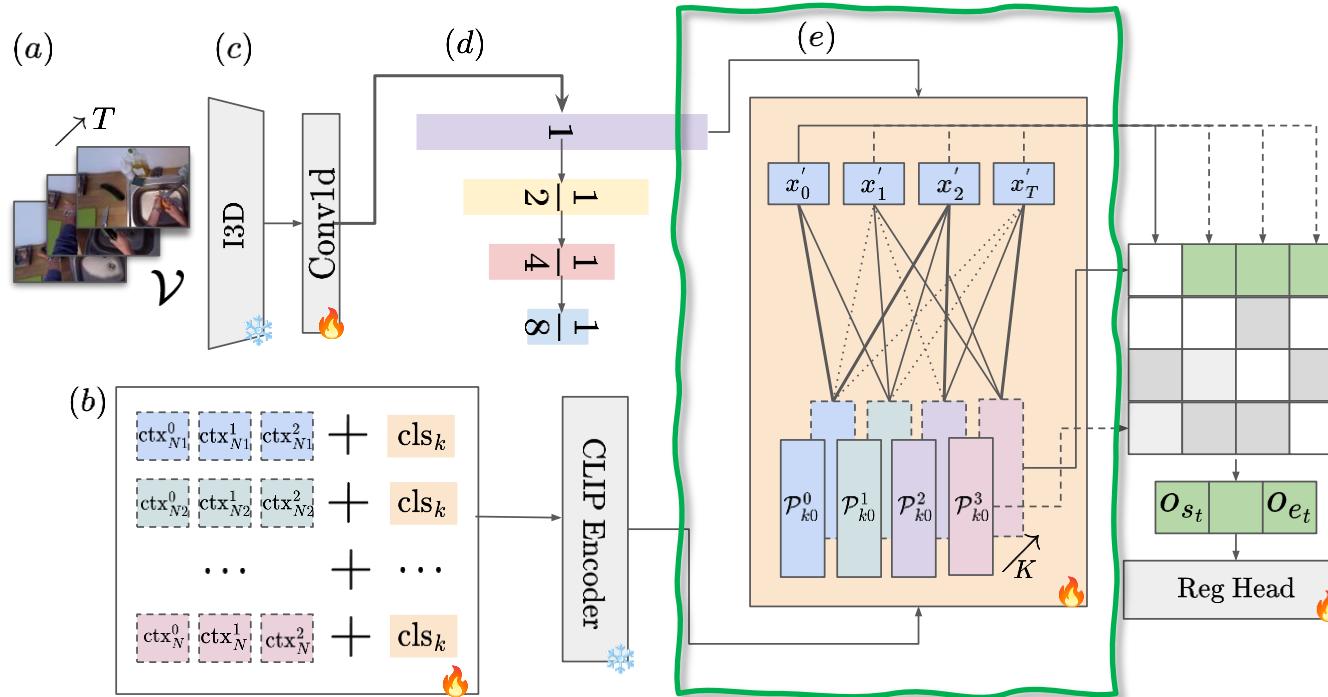
(D) We use average pooling to down sample visual features via a temporal feature pyramid creating an array of features for each temporal resolution.



# Methodology

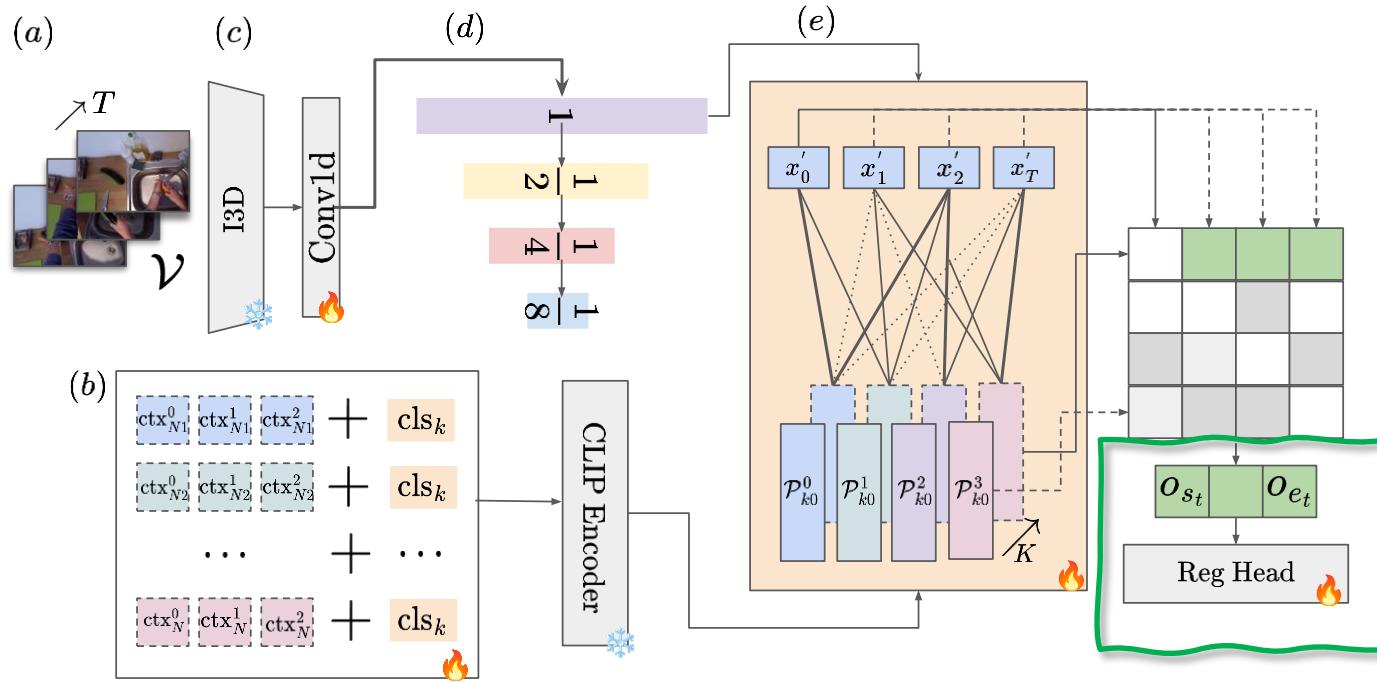
The key here is that Optimal Transport algorithm includes entropic regularisation (Sinkhorn-Knopp algorithm) which ensures that the assignments between prompts and visual features are well distributed preventing collapse of prompts to one single visual feature.

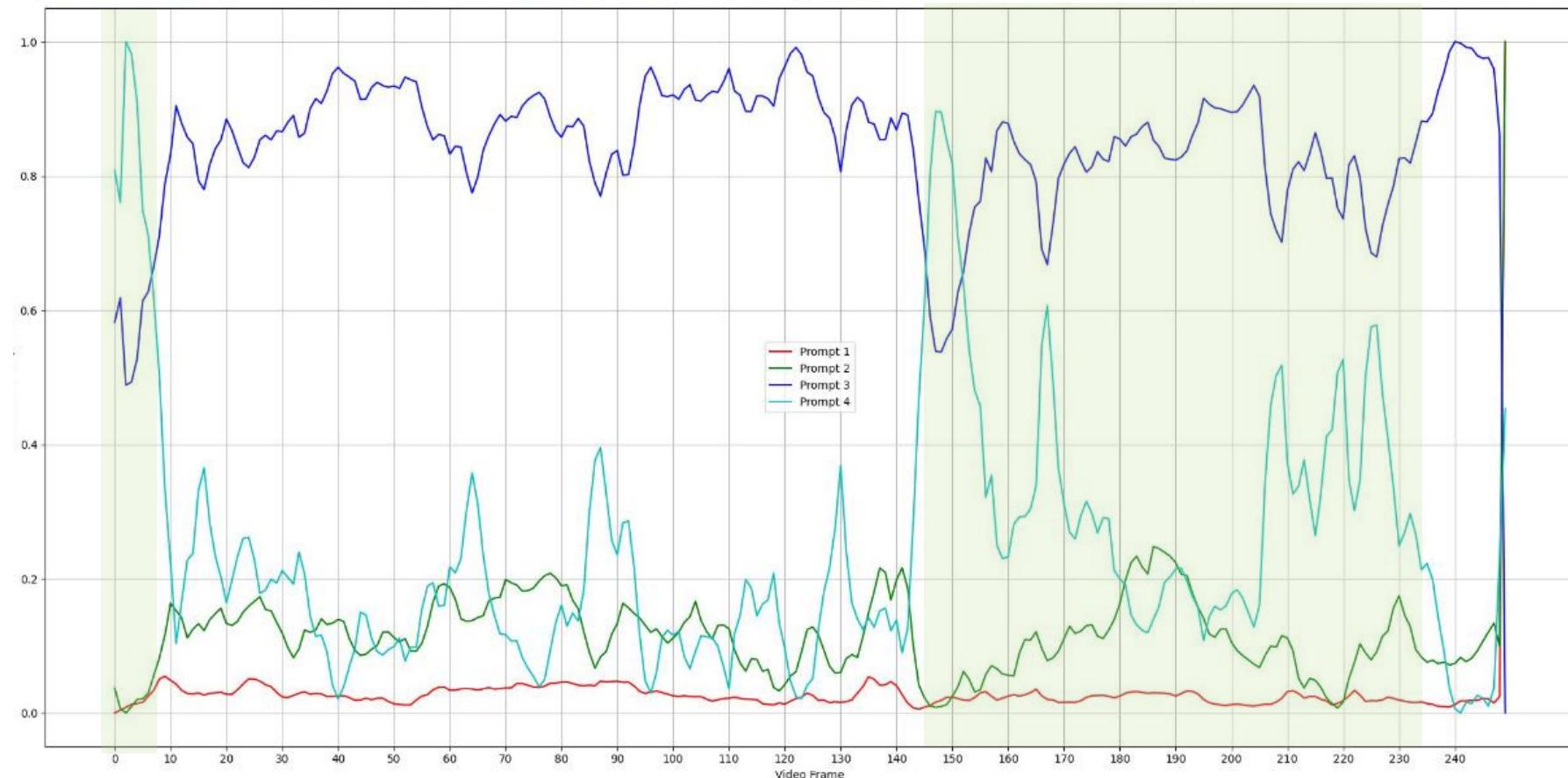
(E) We then compute the optimal transport plan between each temporal feature and prompt, across all pyramid levels.



# 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.



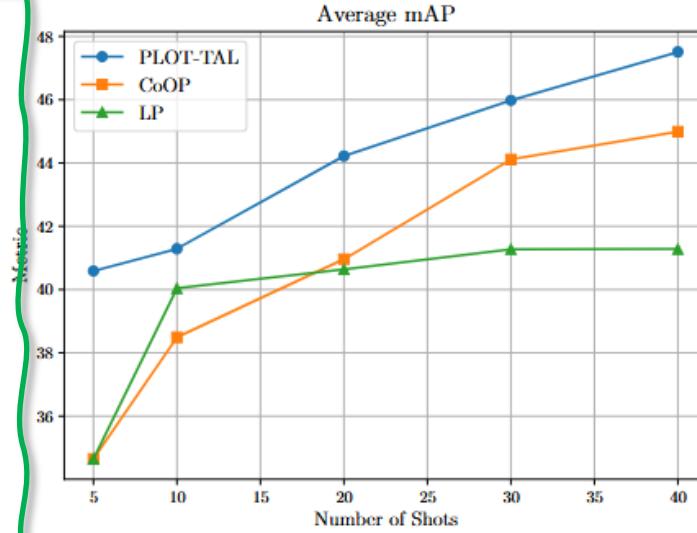
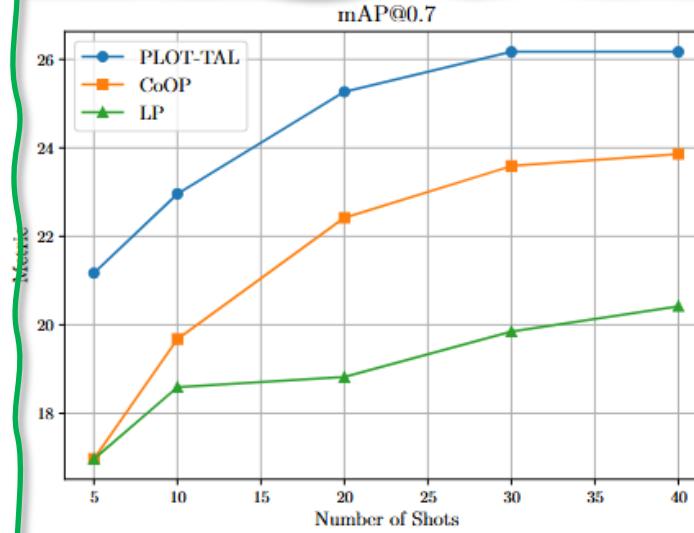
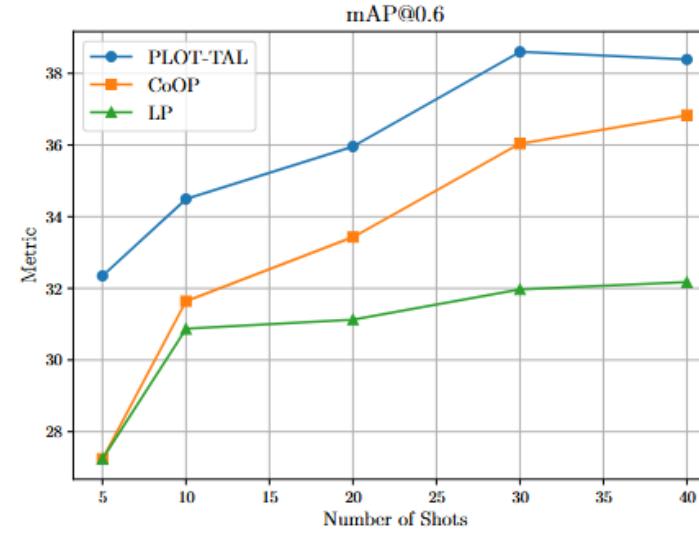
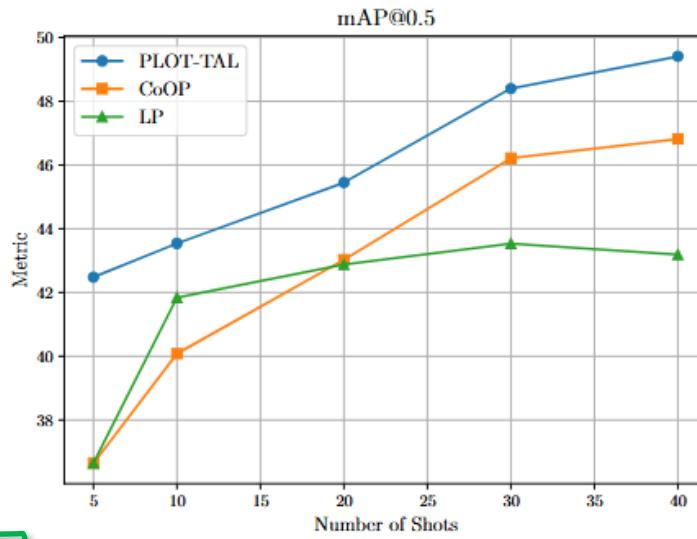
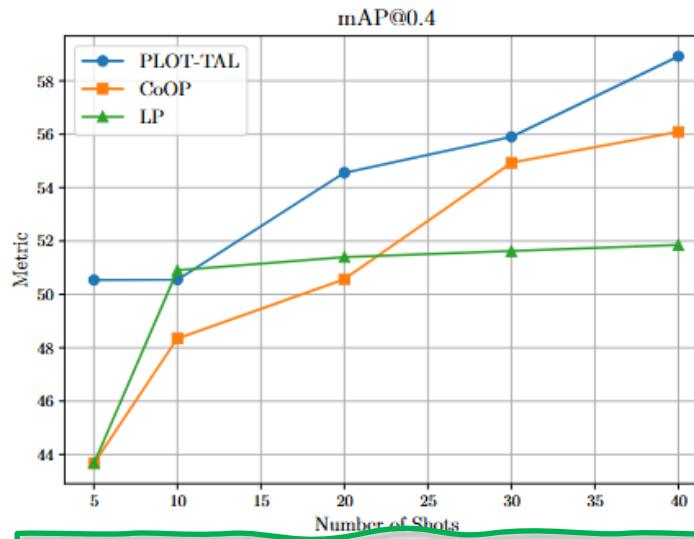
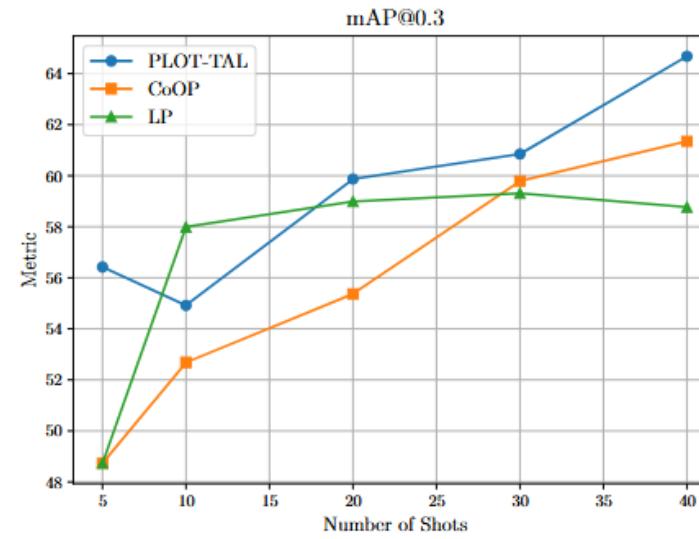


Here we can see the alignment for each learnable prompt using the transport policy. Prompts are aligned with different features across the video including the stadium and players. Cricket shot prediction is shown in green.

# 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
<b>PLOT-TAL (Ours)</b>	E2E + PL	<b>38.24</b>
<b>PLOT-TAL (Verbose) (Ours)</b>	E2E + PL	<b>40.59</b>

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](mailto:Edward.fish@surrey.ac.uk)  
 Andrew Gilbert: [A.gilbert@surrey.ac.uk](mailto:A.gilbert@surrey.ac.uk)

**PAPER** | CVSSP | UNIVERSITY OF SURREY | ICCV HONOLULU, 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**

Most 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 start-end times and poor generalization.

**Our Hypothesis**

Action is 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 metric to regularize to find the most efficient alignment between the prompts and the frames, forcing the prompts to specialize and remember events, thus preventing them from all learning the same redundant information which may not generalise to new contexts.

**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 video such as the cricket shot, while others align with contextual visual features such as the field.

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

A single prompt trained on few examples (e.g., 5) is often forced to learn multiple concepts to generalise well. This results in sub-optimal representation failing 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 specialisation, ensuring the prompts remain diverse and discriminative.

**Methodology**

(A-C) We first extract  $T$  frames from a video  $V$  using a frozen 3D 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, pass them through a linear layer, and classification via 1D Convolution and MLP heads.

**Results**

**ICCV 2025 Performance**

Method	Approach	Arg. mAP (%)
Majority Voting (5shots, 5ways)		22.8
Convex Action Loc. [1]	ML + PL	24.9
MILPE [2]	ML + PL	24.9
MLPE [3]	ML	32.7
Q.A. Transformer [4]		32.7
<b>PLOT-TAL (Ours)</b>	OTPL + PL	<b>34.68</b>
<b>PLOT-TAL (Ours)</b>	OTPL + PL	<b>36.24</b>
<b>PLOT-TAL (Ours)</b>	OTPL + PL	<b>36.24</b>

Results on THUMOS compared to existing few-shot approaches.

**ICCV ActionNet**

Method	OTPL	OTPL + PL							
	0.81	0.82	0.83	0.84	0.82	0.83	0.84	0.85	0.85
Arg.	14.3	13.1	13.1	10.3	9.3	12.1	21.2	18.9	18.9
MLPE (LP)	16.1	15.0	13.8	11.3	9.5	13.2	17.6	16.3	14.6
ConvT [1]	16.1	15.0	13.8	11.3	9.5	13.2	17.6	16.3	14.6
<b>PLOT-TAL (Ours)</b>	<b>17.9</b>	<b>16.7</b>	<b>15.1</b>	<b>12.3</b>	<b>10.0</b>	<b>14.1</b>	<b>21.8</b>	<b>20.9</b>	<b>19.4</b>
<b>PLOT-TAL (Ours)</b>	<b>17.9</b>	<b>16.7</b>	<b>15.1</b>	<b>12.3</b>	<b>10.0</b>	<b>14.1</b>	<b>21.8</b>	<b>20.9</b>	<b>19.4</b>

Results on Epic-Kitchens dataset sample prompt learning (OTPL) and averaging prompt features.

**Evaluations**

Prompts (N)	mAP@Inf	mAP@100	Arg. mAP (%)				
0.3	0.8	0.8	0.8				
4	55.88	50.27	45.06	39.97	21.16	40.46	26.16
8	53.60	48.77	41.74	31.06	20.70	39.29	26.16
16	53.74	48.52	41.02	30.57	20.83	39.73	26.16
32	53.74	48.52	41.02	30.57	20.83	39.73	26.16
64	53.66	48.48	41.04	30.64	20.15	39.79	26.16
128	53.66	48.48	41.04	30.64	20.15	39.79	26.16
256	53.66	48.48	41.04	30.64	20.15	39.79	26.16
512	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1024	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2048	53.66	48.48	41.04	30.64	20.15	39.79	26.16
4096	53.66	48.48	41.04	30.64	20.15	39.79	26.16
8192	53.66	48.48	41.04	30.64	20.15	39.79	26.16
16384	53.66	48.48	41.04	30.64	20.15	39.79	26.16
32768	53.66	48.48	41.04	30.64	20.15	39.79	26.16
65536	53.66	48.48	41.04	30.64	20.15	39.79	26.16
131072	53.66	48.48	41.04	30.64	20.15	39.79	26.16
262144	53.66	48.48	41.04	30.64	20.15	39.79	26.16
524288	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1048576	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2097152	53.66	48.48	41.04	30.64	20.15	39.79	26.16
4194304	53.66	48.48	41.04	30.64	20.15	39.79	26.16
8388608	53.66	48.48	41.04	30.64	20.15	39.79	26.16
16777216	53.66	48.48	41.04	30.64	20.15	39.79	26.16
33554432	53.66	48.48	41.04	30.64	20.15	39.79	26.16
67108864	53.66	48.48	41.04	30.64	20.15	39.79	26.16
134217728	53.66	48.48	41.04	30.64	20.15	39.79	26.16
268435456	53.66	48.48	41.04	30.64	20.15	39.79	26.16
536870912	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1073741824	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2147483648	53.66	48.48	41.04	30.64	20.15	39.79	26.16
4294967296	53.66	48.48	41.04	30.64	20.15	39.79	26.16
8589934592	53.66	48.48	41.04	30.64	20.15	39.79	26.16
17179869184	53.66	48.48	41.04	30.64	20.15	39.79	26.16
34359738368	53.66	48.48	41.04	30.64	20.15	39.79	26.16
68719476736	53.66	48.48	41.04	30.64	20.15	39.79	26.16
137438953472	53.66	48.48	41.04	30.64	20.15	39.79	26.16
274877906944	53.66	48.48	41.04	30.64	20.15	39.79	26.16
549755813888	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1099511627776	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2199023255552	53.66	48.48	41.04	30.64	20.15	39.79	26.16
4398046511104	53.66	48.48	41.04	30.64	20.15	39.79	26.16
8796093022208	53.66	48.48	41.04	30.64	20.15	39.79	26.16
17592186044416	53.66	48.48	41.04	30.64	20.15	39.79	26.16
35184372088832	53.66	48.48	41.04	30.64	20.15	39.79	26.16
70368744177664	53.66	48.48	41.04	30.64	20.15	39.79	26.16
140737488355328	53.66	48.48	41.04	30.64	20.15	39.79	26.16
281474976710656	53.66	48.48	41.04	30.64	20.15	39.79	26.16
562949953421312	53.66	48.48	41.04	30.64	20.15	39.79	26.16
112589990684264	53.66	48.48	41.04	30.64	20.15	39.79	26.16
225179981368528	53.66	48.48	41.04	30.64	20.15	39.79	26.16
450359962737056	53.66	48.48	41.04	30.64	20.15	39.79	26.16
900719925474112	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1801439850948224	53.66	48.48	41.04	30.64	20.15	39.79	26.16
3602879701896448	53.66	48.48	41.04	30.64	20.15	39.79	26.16
7205759403792896	53.66	48.48	41.04	30.64	20.15	39.79	26.16
14411518807585792	53.66	48.48	41.04	30.64	20.15	39.79	26.16
28823037615171584	53.66	48.48	41.04	30.64	20.15	39.79	26.16
57646075230343168	53.66	48.48	41.04	30.64	20.15	39.79	26.16
115292150460686336	53.66	48.48	41.04	30.64	20.15	39.79	26.16
230584300921372672	53.66	48.48	41.04	30.64	20.15	39.79	26.16
461168601842745344	53.66	48.48	41.04	30.64	20.15	39.79	26.16
922337203685490688	53.66	48.48	41.04	30.64	20.15	39.79	26.16
184467440737098136	53.66	48.48	41.04	30.64	20.15	39.79	26.16
368934881474196272	53.66	48.48	41.04	30.64	20.15	39.79	26.16
737869762948392544	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1475739525896785088	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2951479051793570176	53.66	48.48	41.04	30.64	20.15	39.79	26.16
5902958103587140352	53.66	48.48	41.04	30.64	20.15	39.79	26.16
11805916207174280704	53.66	48.48	41.04	30.64	20.15	39.79	26.16
23611832414348561408	53.66	48.48	41.04	30.64	20.15	39.79	26.16
47223664828697122816	53.66	48.48	41.04	30.64	20.15	39.79	26.16
94447329657394245632	53.66	48.48	41.04	30.64	20.15	39.79	26.16
188894659314788491264	53.66	48.48	41.04	30.64	20.15	39.79	26.16
377789318629576982528	53.66	48.48	41.04	30.64	20.15	39.79	26.16
755578637259153965056	53.66	48.48	41.04	30.64	20.15	39.79	26.16
151115727458230793012	53.66	48.48	41.04	30.64	20.15	39.79	26.16
302231454916461586024	53.66	48.48	41.04	30.64	20.15	39.79	26.16
604462909832923172048	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1208925819665846344096	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2417851639331692688192	53.66	48.48	41.04	30.64	20.15	39.79	26.16
4835703278663385376384	53.66	48.48	41.04	30.64	20.15	39.79	26.16
9671406557326770752768	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1934281311465354150536	53.66	48.48	41.04	30.64	20.15	39.79	26.16
3868562622930708301072	53.66	48.48	41.04	30.64	20.15	39.79	26.16
7737125245861416602144	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1547425049172283320428	53.66	48.48	41.04	30.64	20.15	39.79	26.16
3094850098344566640856	53.66	48.48	41.04	30.64	20.15	39.79	26.16
6189700196689133281712	53.66	48.48	41.04	30.64	20.15	39.79	26.16
1237940039337826656344	53.66	48.48	41.04	30.64	20.15	39.79	26.16
2475880078675653312688	53.66	48.48	41.04	30.64	20.15	39.79	26.16
4951760157351306625376	53.66	48.48	41.04	30.64	20.15	39.79	26.16
9903520314702613250752	53.66	48.48	41.04	30.64	20.15	39.79	26.16
19807040629405226501504	53.66	48.48	41.04	30.64	20.15	39.79	26.16
39614081258810453003008	53.66	48.48	41.04	30.64	20.15	39.79	26.16
79228162517620906006016	53.66	48.48	41.04	30.64	20.15	39.79	26.16
158456325352401812012032	53.66	48.48	41.04	30.64</td			

# 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  $d_{OT}(l, c) = \langle \mathbf{T}_{l,c}^*, \mathbf{C}_{l,c} \rangle$ 
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  $\{\text{ctx}\}$ 
```

---