Paper

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

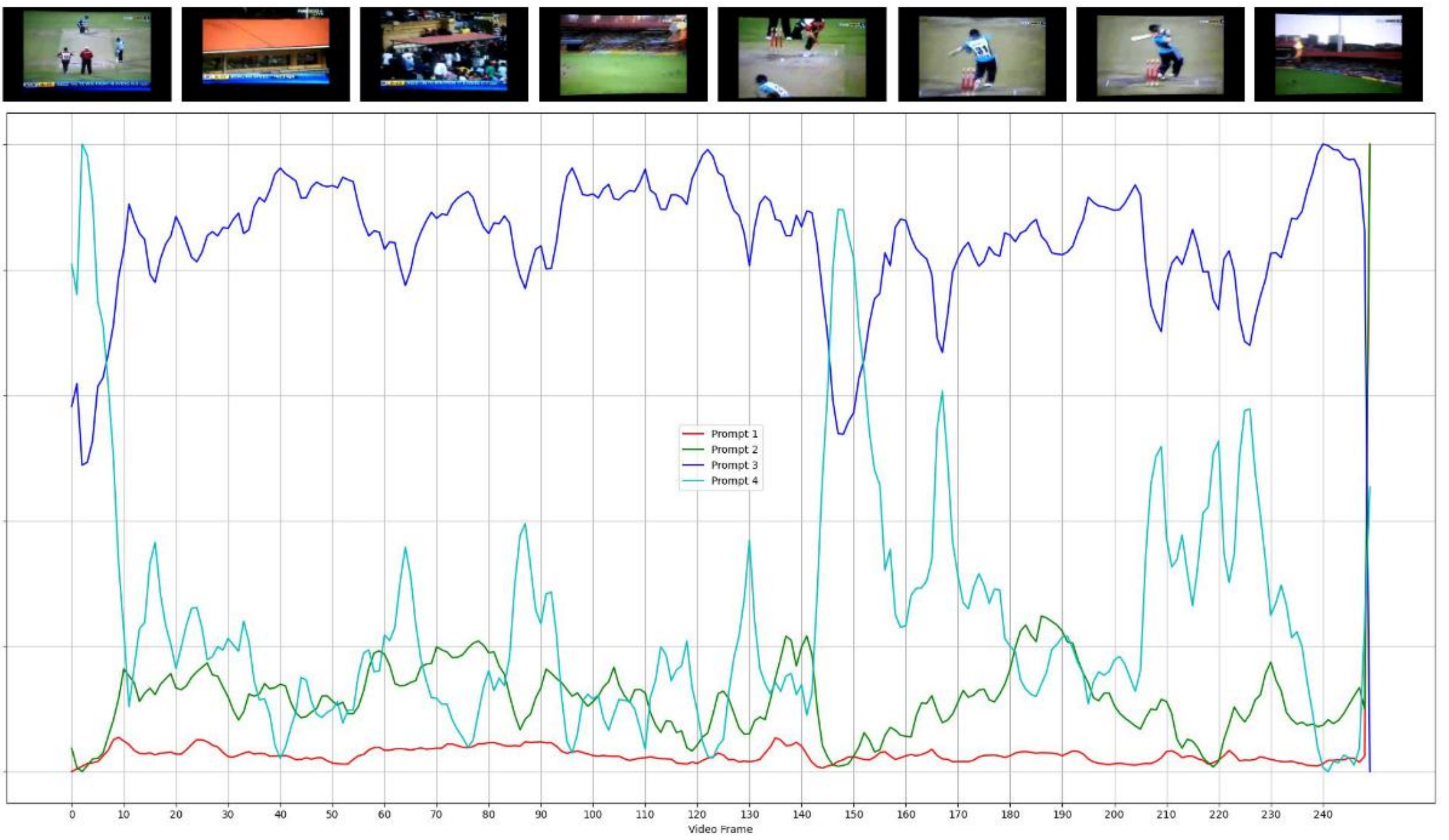
In few-shot temporal action localisation we need to generalise from just 5 instances of an action to the same action in different environments and contexts.

**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 start/end times and poor generalization.

**Our Hypothesis:**
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.

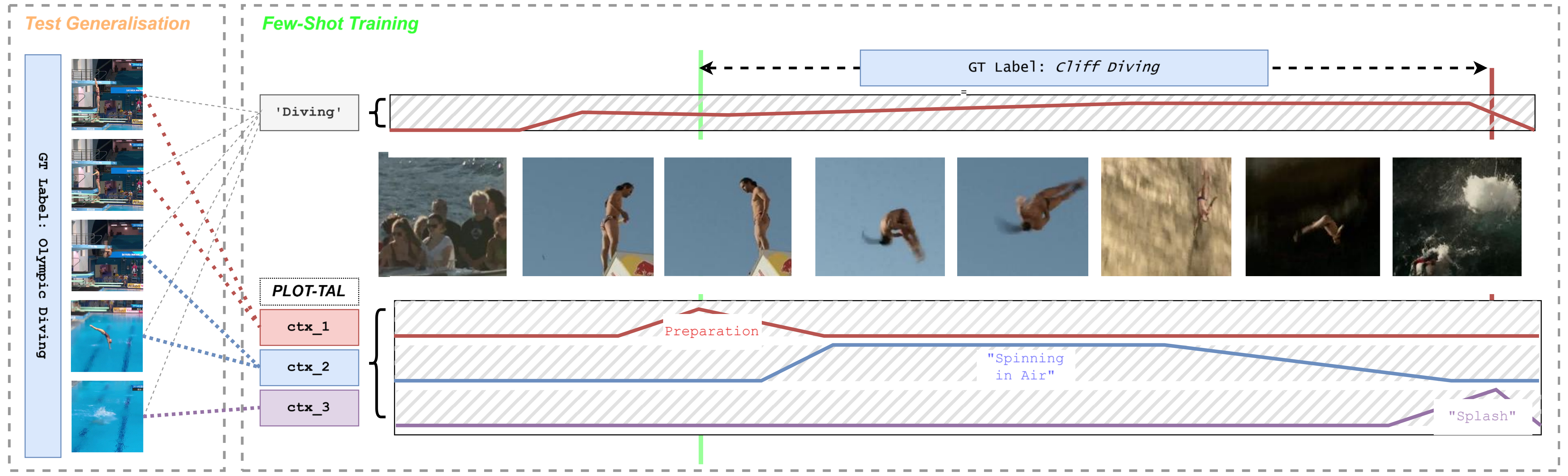
**Our Solution:**
In PLOT-TAL, we represent each action class with a number of learnable prompts. Each prompt is encouraged to become a "specialist" on a distinct sub-event of the action. We use Optimal Transport (OT) as a structural regularizer to find the most efficient alignment between the prompts 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.

## Qualitative Results



In this example, we visualize 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

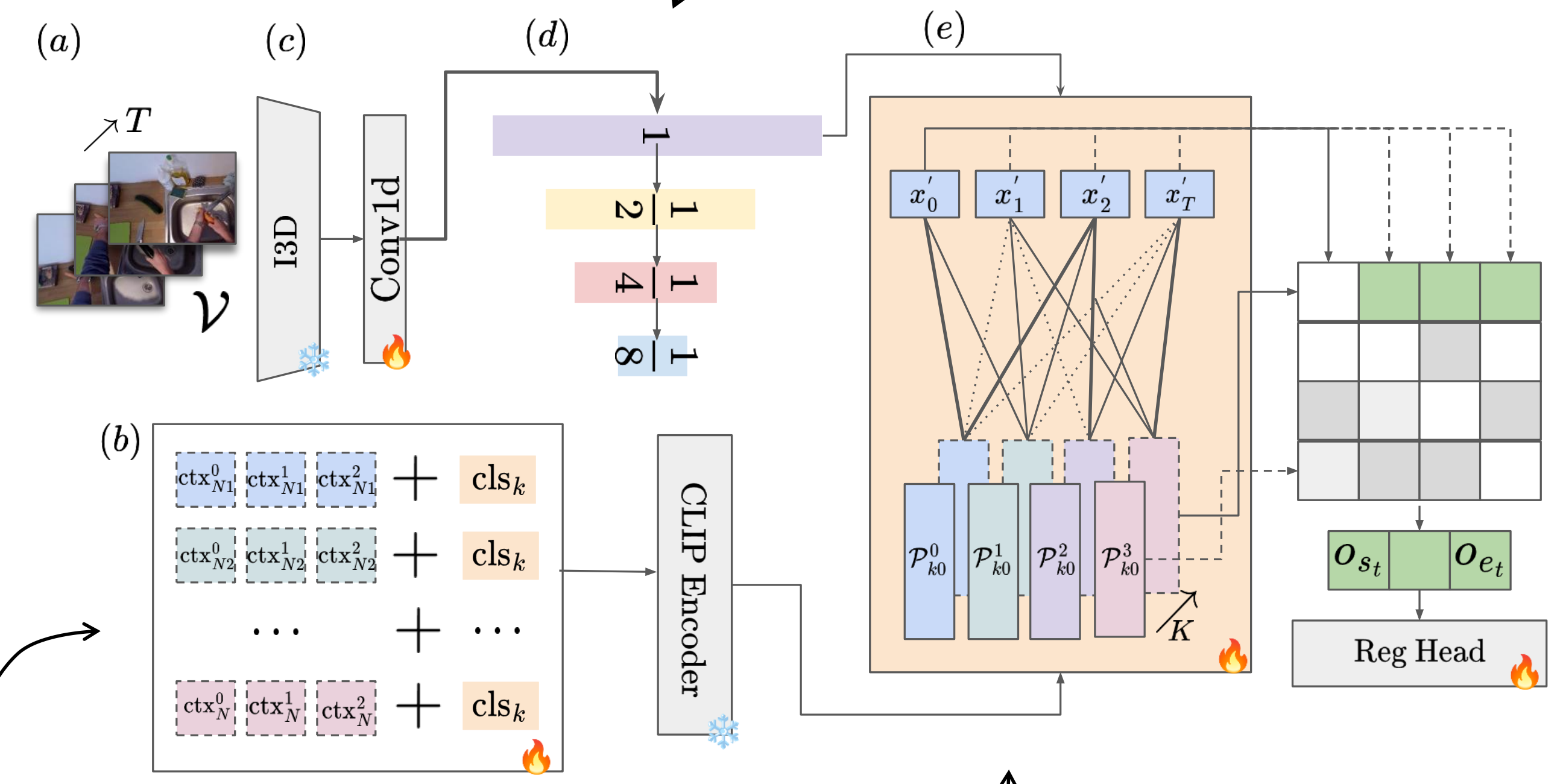


A single prompt trained on a few examples of "diving" in a specific context (top) tends to overfit to environmental cues like the cliffs and sea. This holistic representation fails to generalize to a novel environment. Our method learns an ensemble of prompts that specialize on the compositional, environment-agnostic 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.

## Methodology

(A-C) We first extract **T** frames from a video *V* using a frozen I3D encoder pretrained on Kinetics.

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


(a)
(b)
(c)
(d)
(e)
T
$\mathcal{V}$
I3D
Conv1d
1
2 | 1
4 | 1
8 | 1
$\text{ctx}^0_{N1}$ $\text{ctx}^1_{N1}$ $\text{ctx}^2_{N1}$ + $\text{cls}_k$
$\text{ctx}^0_{N2}$ $\text{ctx}^1_{N2}$ $\text{ctx}^2_{N2}$ + $\text{cls}_k$
… + …
$\text{ctx}^0_{N}$ $\text{ctx}^1_{N}$ $\text{ctx}^2_{N}$ + $\text{cls}_k$
CLIP Encoder
$x'_0$ $x'_1$ $x'_2$ $x'_T$
$\mathcal{P}^0_{k0}$ $\mathcal{P}^1_{k0}$ $\mathcal{P}^2_{k0}$ $\mathcal{P}^3_{k0}$
K
$O_{s_t}$ $O_{e_t}$
Reg Head

(E) The transport plan is fixed and multiplied by the features. We concatenate all temporal layers and perform regression and classification via 1D Convolution and MLP heads.

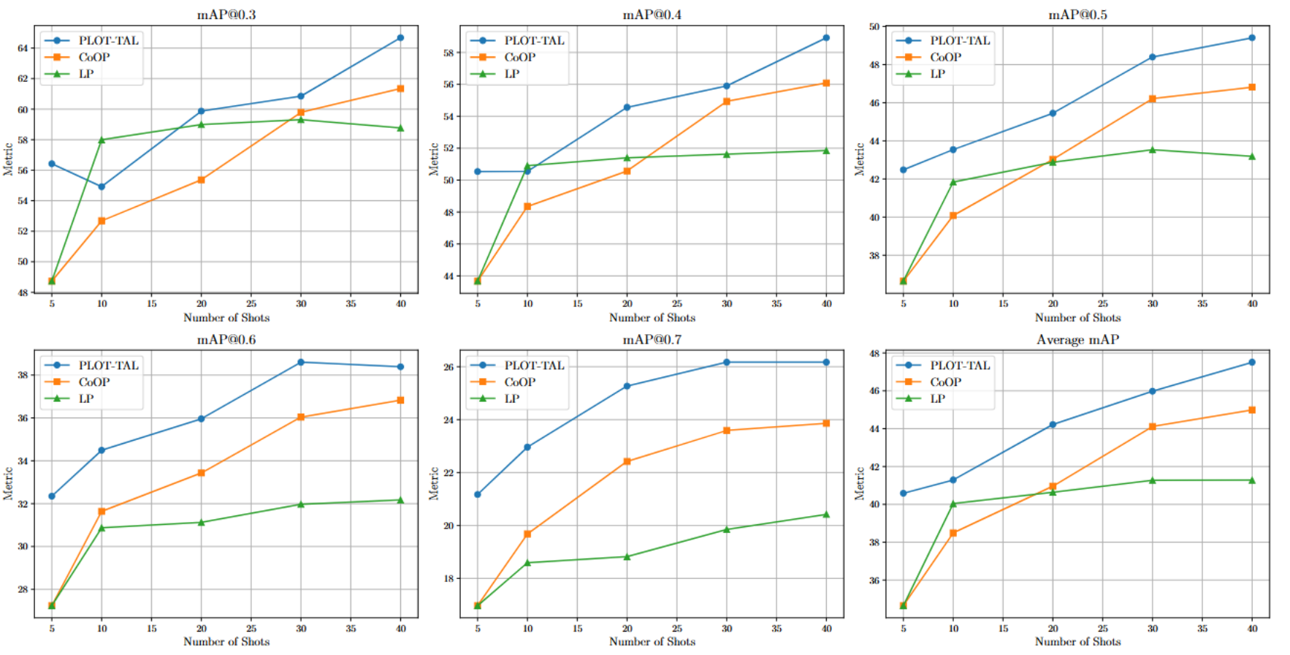
(B) We Initialise **N** learnable prompts for each class K. They are prepended to the prompt and embedded via CLIP.

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

## 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** |

Results on THUMOS compared to existing few-shot approaches.



Performance over increasing number of training samples per class.

| | EPIC-Kitchens Noun | | | | | | EPIC-Kitchens Verb | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. |
| Ours (Avg.) | 14.3 | 13.5 | 13.1 | 10.3 | 9.3 | 12.1 | 21.2 | 19.9 | 18.0 | 15.2 | 11.9 | 17.3 |
| Linear Probe (LP) | **18.0** | 15.4 | 14.1 | 12.2 | 9.5 | 13.9 | **22.5** | **21.3** | 19.2 | 17.1 | 13.3 | 18.7 |
| CoOp [35] | 16.1 | 15.0 | 13.8 | 11.8 | 9.5 | 13.3 | 18.5 | 17.6 | 16.3 | 14.6 | 12.5 | 15.9 |
| **PLOT-TAL (Ours)** | 17.9 | **16.7** | **15.1** | **12.7** | **10.0** | **14.5** | 21.8 | 20.9 | **19.4** | **17.6** | **14.6** | **18.9** |

Results on Epic Kitchens against simpler single prompt learning (CoOp) and averaging prompt features.

## Ablations

| Prompts (*N*) | mAP @ IoU | | | | | Avg |
|---|---|---|---|---|---|---|
| | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | |
| 4 | 55.88 | 50.21 | 43.06 | 31.97 | 21.16 | 40.46 |
| 6 | **56.42** | **50.54** | 42.48 | 32.35 | 21.17 | **40.59** |
| 8 | 53.60 | 48.72 | 41.74 | 31.68 | 20.70 | 39.29 |
| 10 | 54.96 | 50.27 | **43.45** | **32.53** | **21.44** | 40.53 |
| 12 | 53.74 | 48.25 | 41.02 | 30.57 | 20.06 | 38.73 |
| 14 | 54.25 | 48.94 | 40.90 | 30.78 | 18.86 | 38.75 |
| 16 | 53.66 | 48.28 | 41.04 | 30.84 | 20.15 | 38.79 |

Effect of varying number of learnable context prompts per class on THUMOS dataset.

| FPN Levels | mAP@0.5 | Avg. mAP (%) |
|---|---|---|
| 1 | 25.82 | 26.16 |
| 2 | 37.80 | 35.81 |
| 3 | 39.10 | 36.58 |
| 4 | 40.02 | 38.03 |
| 5 | **43.06** | **40.46** |
| 6 | 42.21 | 39.57 |
| 7 | 41.56 | 38.92 |

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

| Embedding Type | mAP @ IoU | | | | | Avg. mAP (%) |
|---|---|---|---|---|---|---|
| | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | |
| CLIP Vision (ViT-B-16) | 46.99 | 42.09 | 34.26 | 25.34 | 15.82 | 32.90 |
| RGB (I3D) | 43.13 | 38.76 | 31.71 | 23.15 | 14.46 | 30.24 |
| Optical Flow (I3D) | 26.03 | 23.10 | 19.54 | 14.07 | 8.93 | 18.33 |
| **RGB + Flow (I3D)** | **55.88** | **50.21** | **43.06** | **31.97** | **21.16** | **40.46** |

Results with alternative vision encoders. We find that I3D including both Optical Flow and RGB is more effective than the CLIP visual embeddings.