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

Edward Fish
University of Surrey
edward.fish@surrey.ac.uk

Andrew Gilbert
University of Surrey
a.gilbert@surrey.ac.uk

## Abstract

*Few-shot temporal action localization (TAL) methods that adapt large models via single-prompt tuning often fail to produce precise temporal boundaries. This stems from the model learning a non-discriminative mean representation of an action from sparse data, which compromises generalization. We address this by proposing a new paradigm based on multi-prompt ensembles, where a set of diverse, learnable prompts for each action is encouraged to specialize on compositional sub-events. To enforce this specialization, we introduce PLOT-TAL, a framework that leverages Optimal Transport (OT) to find a globally optimal alignment between the prompt ensemble and the video's temporal features. Our method establishes a new state-of-the-art on the challenging few-shot benchmarks of THUMOS'14 and EPIC-Kitchens, without requiring complex meta-learning. The significant performance gains, particularly at high IoU thresholds, validate our hypothesis and demonstrate the superiority of learning distributed, compositional representations for precise temporal localization.*

## 1. Introduction

Temporal Action Localization (TAL) is the task of identifying the start, end, and class labels of actions in continuous videos. While the success of state-of-the-art models has been predicated on access to vast, densely annotated datasets, for TAL to be deployed robustly in real-world applications where data is inherently scarce, these networks need to be able to efficiently learn from only a few samples.

The current strategy for tackling this low-data regime involves adapting large Vision-Language Models (VLMs) [19] via parameter-efficient prompt tuning [35]. However, the limitations of this paradigm become critically apparent in the few-shot setting. By learning a single prompt per class, the model is forced to compress the entire dynamic structure of an action into a single feature vector. This representational bottleneck is severely exacerbated when learning from limited samples. With only a handful of examples, a single prompt is prone to memorizing superficial, non-generalizable cues (e.g., the specific camera angle or background clutter) present in the few shots, leading to poor generalization to novel contexts.

In this work, we propose that robust few-shot generalization stems not from learning a single, monolithic concept, but from discovering the underlying compositional structure of actions. For example, an action like a 'high jump' is a composition of simpler, more reusable sub-events ('running', 'leaping', 'arching the back'). Learning these disentangled concepts from sparse data is a more tractable problem. Consequently, we depart from the single-prompt paradigm and propose modelling each action class with a set of diverse, learnable prompts.

This approach requires a mechanism to guide the specialization of these prompts and prevent their collapse into redundancy. For this, we introduce Optimal Transport (OT) [5], not merely as a matching algorithm, but as a structural regularizer that enforces representational diversity. By seeking the most efficient assignment between the distribution of prompts and the distribution of temporal features, OT implicitly ensures that each prompt finds a unique role in explaining the data. This constraint is a powerful tool against overfitting, preventing the entire set of prompts from aligning with the most prominent feature in the few training examples, thereby fostering a diverse and highly generalizable final representation.

**Our contributions are as follows:**

- We identify the single-prompt architecture as a key source of poor generalization in few-shot TAL.
- We propose a multi-prompt, OT-regulated framework as a direct solution, arguing it learns a more compositional representation inherently better suited to low-data regimes.
- We provide extensive empirical validation, demonstrating state-of-the-art performance on multiple benchmarks.

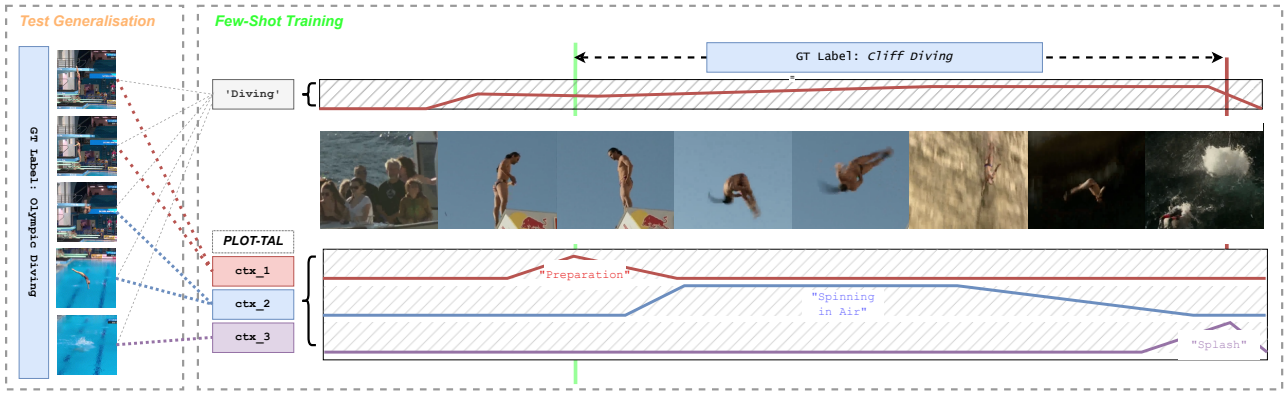


Figure 1. **Conceptual Framework: Compositional Learning for Few-Shot Generalization.** A single prompt trained on a few examples of "diving" in a specific context (top-right) tends to overfit to environmental cues like the cliffs and sea. This holistic representation fails to generalize to a novel environment. In contrast, our method (bottom right) learns an ensemble of prompts that specialize on the compositional, environment-agnostic sub-events of the action: (1) the preparation/stance, (2) the mid-air rotation, and (3) the water entry splash. Optimal Transport is the key mechanism that enforces this specialization, ensuring the prompts remain diverse and discriminative. By identifying these core components, our framework can robustly localize the "diving" action with high precision, even when presented with a completely different environment, such as an indoor swimming pool (left panel), using only a few samples.

## 2. Related Work

The field of TAL has evolved from two-stage methods, which first generate proposals and then classify them [7, 12, 13], towards more efficient single-stage architectures. These unified models, inspired by advances in object detection [15, 20], perform classification and boundary regression in a single pass. Recent state-of-the-art methods frequently leverage Transformer-based backbones [24] combined with feature pyramid networks (FPNs) to handle actions at multiple temporal scales [4, 21, 31]. While powerful, these models' performance relies on large-scale supervision, a limitation our work aims to address.

### 2.1. Few-Shot Learning for Action Localization

Adapting TAL to the few-shot setting has primarily been explored through meta-learning [28, 29]. These methods train a model to quickly adapt to novel classes by learning across a distribution of tasks, or "episodes." While effective, they often involve complex, multi-stage training schedules. Other approaches have tackled zero-shot TAL [18], but often rely on external cues from pre-trained classifiers like UntrimmedNet [26], which may not be available in practical scenarios. Our approach provides a simpler, end-to-end framework for few-shot learning that circumvents complex meta-learning and external dependencies.

### 2.2. Prompt Learning in Vision

Prompt learning has emerged as a parameter-efficient method for adapting large, frozen VLMs to new tasks [34, 35]. By only tuning a small number of context vectors prepended to a text prompt, these methods can steer the model's behaviour without updating all of its parameters. This has been applied to action recognition [9] and video-text alignment [11], but its application to the fine-grained task of localization remains under explored. Works like [17] have combined prompting with meta-learning, but our work focuses on a more direct adaptation for few-shot TAL and extends to multiple prompts.

### 2.3. Optimal Transport in Machine Learning

The Optimal Transport (OT) problem, originating from the work of Monge [25], provides a principled way to measure the distance between probability distributions. Its applicability to machine learning was greatly expanded by the introduction of entropic regularization and the efficient Sinkhorn algorithm [5], which made it computationally tractable for high-dimensional data. OT has since been applied to various vision tasks [23]. Most relevant to our work is Chen et al. [3], who first used OT to align multiple prompts to feature maps for few-shot image classification. Our primary contribution is the novel adaptation and validation of this concept for the temporal domain, demonstrating that OT is a powerful tool for modelling the dynamic, compositional structure of actions over time—a fundamentally different and more complex problem than static image classification.

## 3. Methodology

We propose a novel framework for Temporal Action Localization (TAL), which we term PLOT-TAL. Our approach integrates pre-trained feature extraction, adaptive multi-prompt learning, and an efficient feature-prompt alignment

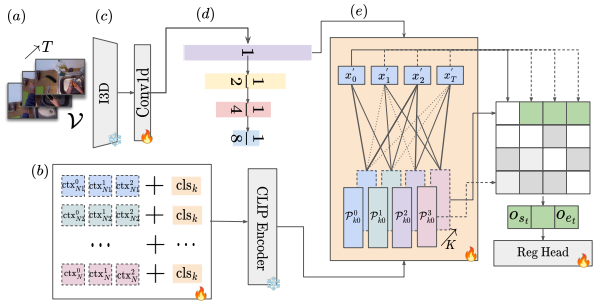


Figure 2. An overview of the PLOT-TAL framework. **(A)** We first extract $T$ frames from a video $V$. **(B)** An ensemble of $N$ learnable prompts is generated for each class. **(C)** Video features are extracted by a frozen visual encoder and text prompts by a frozen VLM text encoder (CLIP). **(D)** A temporal feature pyramid is constructed via max-pooling. **(E)** Optimal Transport aligns the prompt ensemble with video features at each pyramid level. **(F)** The resulting features are passed to lightweight localization heads to predict action instances and a class label. Only modules marked with a flame symbol contain trainable parameters.

mechanism based on the Sinkhorn algorithm for Optimal Transport. The framework is designed to be parameter-efficient and is trained end-to-end. An overview of the network architecture is presented in Figure 2.

## 3.1. Problem Formulation

We first formally define the task. Given an untrimmed input video, represented as a sequence of feature vectors $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\}$, the objective of TAL is to predict a set of action instances $\mathcal{Y} = \{(s_i, e_i, c_i)\}_{i=1}^{M}$. Each tuple denotes an action of class $c_i \in \{1, \ldots, C\}$ starting at time $s_i$ and ending at time $e_i$, where the temporal boundaries must satisfy $1 \leq s_i < e_i \leq T$.

In the few-shot setting this paper addresses, the model must learn to perform this task for all $C$ classes using only a small number, $K$, of annotated support examples for each class (e.g., $K = 5$). This constraint requires a model that can generalize effectively from sparse data while minimizing the number of trainable parameters to prevent overfitting.

## 3.2. Multi-Scale Feature and Prompt Representation

Our framework is designed to capture actions at varying temporal scales by using hierarchical representations for both visual features and textual prompts.

### 3.2.1. Temporal Feature Pyramid

The input video is first processed by a frozen, pre-trained 3D Convolutional Network (e.g., I3D [2]) to extract a sequence of clip-level feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. To enhance local temporal context, this sequence is passed through a series of 1D temporal convolutional layers, following modern TAL architectures [22, 31]. This yields a refined feature sequence $\{\mathbf{x'}_1, \ldots, \mathbf{x'}_T\}$. From this base sequence, we construct a temporal feature pyramid of $L$ levels by applying successive max-pooling operations with a stride of 2. This results in a set of multi-scale feature representations $\{\mathbf{F}_1, \ldots, \mathbf{F}_L\}$, where each matrix $\mathbf{F}_l \in \mathbb{R}^{T_l \times D}$ contains the feature sequence at temporal scale $l$.

### 3.2.2. Adaptive Multi-Prompt Ensembles

To address the limitations of a single-prompt representation, we model each action class $c$ with an ensemble of $N$ diverse, learnable prompts. Each of the $N$ prompts is constructed by prepending a unique set of $n_{\text{ctx}}$ learnable context vectors to the class name: $[\text{ctx}_1, \ldots, \text{ctx}_{n_{\text{ctx}}}, \text{class\_name}_c]$. This set of $N$ textual prompts is then passed through the frozen text encoder of a pre-trained VLM like CLIP [19]. The process is formalized as:

$$\mathbf{g}_{ci} = f_{\text{CLIP}}(\text{class\_name}_c, \{\text{ctx}_{cij}\}_{j=1}^{n_{\text{ctx}}}) \tag{1}$$

where $\mathbf{g}_{ci}$ is the $i$-th prompt embedding for class $c$. This yields a set of prompt embeddings $\mathbf{G}_c = \{\mathbf{g}_{c1}, \ldots, \mathbf{g}_{cN}\} \in \mathbb{R}^{N \times D}$. The only trainable parameters in this module are the context vectors $\{\text{ctx}\}$, ensuring our approach remains highly parameter-efficient.

## 3.3. Multi-Resolution Alignment via Optimal Transport

The core technical contribution of our work is the mechanism for aligning the prompt ensemble $\mathbf{G}_c$ with the temporal feature sequence $\mathbf{F}_l$ at each pyramid level $l$.

### 3.3.1. Optimal Transport Formulation

We formulate the alignment as a distribution matching problem. For a given class $c$ and level $l$, we treat the set of $T_l$ video features and $N$ prompt embeddings as empirical samples from two discrete probability distributions, $U_l$ and $V_c$, respectively:

$$U_l = \sum_{i=1}^{T_l} u_i \delta_{\mathbf{f}_i} \quad \text{and} \quad V_c = \sum_{j=1}^{N} v_j \delta_{\mathbf{g}_j} \tag{2}$$

where $\delta_\cdot$ is the Dirac delta function, and $\mathbf{u}$ and $\mathbf{v}$ are uniform probability vectors (i.e., $u_i = 1/T_l$, $v_j = 1/N$). We define a cost matrix $\mathbf{C} \in \mathbb{R}^{T_l \times N}$ where each entry $C_{ij}$ is the cosine distance between the video feature $\mathbf{f}_i$ and the prompt embedding $\mathbf{g}_j$. The goal of OT is to find a transport plan $\mathbf{T} \in \mathbb{R}^{T_l \times N}$ that minimizes the total transportation cost. We use the entropically regularized formulation, solvable efficiently via the Sinkhorn algorithm [5]:

$$d_{\text{OT}}(\mathbf{F}_l, \mathbf{G}_c) = \min_{\mathbf{T} \in \mathcal{U}(\mathbf{u}, \mathbf{v})} \langle \mathbf{T}, \mathbf{C} \rangle - \lambda H(\mathbf{T}) \tag{3}$$

**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 $\{\text{ctx}\}$

Here, $\langle \cdot, \cdot \rangle$ is the Frobenius dot product, $H(\mathbf{T})$ is the entropy of the transport plan, and $\lambda$ is a regularization parameter. The resulting optimal transport plan $\mathbf{T}^*$ represents a soft, many-to-many assignment map.

#### 3.3.2. Optimization

This alignment process is embedded in a two-stage optimization loop as proposed in [3]. In the inner loop of each training step, we fix the model parameters and iteratively solve Eq. 3 to find the optimal transport plan $\mathbf{T}^*_l$. In the outer loop, with the transport plans fixed, we compute the final task loss and backpropagate the gradients through the OT process to update the learnable prompt context vectors. For the OT-specific parameters, we follow the setup in [3], setting the convergence threshold $\delta = 0.01$, the entropy parameter $\lambda = 0.1$, and we perform a maximum of 100 iterations within the inner Sinkhorn loop. A detailed overview of this process is provided in Algorithm 1.

### 3.4. Decoder and Learning Objective

#### 3.4.1. Decoder Architecture.

Following the multi-scale alignment, the resulting features are passed to two lightweight, parallel heads for the final predictions: a classification head that generates a probability distribution over the $C$ classes using a sigmoid activation, and a regression head that predicts the temporal offsets to the start and end boundaries of a potential action using a ReLU activation.

#### 3.4.2. Learning Objective

The network is trained end-to-end by minimizing a total loss function, $\mathcal{L}_{\text{total}}$. We use the Focal Loss ($\mathcal{L}_{\text{cls}}$) [14] for classification and the Distance-IoU (DIoU) Loss ($\mathcal{L}_{\text{reg}}$) [32] for regression. The total loss, aggregated over all temporal locations $t$ and pyramid levels $l$, is:

$$\mathcal{L}_{\text{total}} = \frac{1}{N_{pos}} \sum_{l,t} \Big( \mathcal{L}_{\text{cls}}(\hat{c}_{lt}, c_{lt}) + \lambda_{reg} \mathbb{1}_{\{c_{lt} > 0\}} \mathcal{L}_{\text{reg}}(\hat{o}_{lt}, o_{lt}) \Big) \tag{4}$$

where $\hat{c}_{lt}$ and $\hat{o}_{lt}$ are the predictions, $N_{pos}$ is the number of positive samples, and the indicator function $\mathbb{1}_{\{\cdot\}}$ applies the regression loss only to foreground frames.

## 4. Experiments

We conduct a comprehensive set of experiments to rigorously validate our proposed framework, PLOT-TAL.

### 4.1. Experimental Setup

#### 4.1.1. Datasets and Metrics

Our evaluation is performed on two standard, yet diverse, benchmarks for temporal action localization:

- **THUMOS'14** [8] is a widely used dataset featuring 20 classes of sports actions in 200 validation and 213 test videos.
- **EPIC-Kitchens-100** [6] is a large-scale egocentric dataset. We evaluate on both the verb (97 classes) and noun (300 classes) localization tasks.

Following standard protocols, we report the mean Average Precision (mAP) at various Intersection over Union (IoU) thresholds: [0.3, 0.4, 0.5, 0.6, 0.7], and report the average of these as our primary metric.

#### 4.1.2. Few-Shot Protocol and Comparison to Prior Work

All our experiments are conducted under a 5-shot, C-way protocol, where $C$ is the total number of classes in the respective dataset. In this challenging setup, the model learns from only 5 examples per class and must then localize actions from among all $C$ classes simultaneously during testing. This differs significantly from the episodic 5-shot, 5-way meta-learning protocol used in some prior works [10, 16]. Due to this fundamental difference in task difficulty, results from those works are presented for context but are not directly comparable.

#### 4.1.3. Baselines

We compare PLOT-TAL against three strong baselines trained under the exact same few-shot protocol for a fair comparison:

1. **Linear Probe (LP)**: A linear classifier trained on top of the frozen video features.

2. **CoOp** [35]: The canonical single-prompt learning method.
3. **Ours (Avg.)**: An ablation of our model where the $N$ prompts are combined by simple averaging, removing the Optimal Transport module.

#### 4.1.4. Implementation Details

We use standard I3D (RGB+Flow) features for THUMOS'14 and SlowFast for EPIC-Kitchens. Models are trained for 100 epochs using the Adam optimizer with a batch size of 2 on a single NVIDIA RTX 3090 GPU. Based on our ablation studies, we set the number of prompts $N = 6$ and context tokens $n_{ctx} = 16$. The OT regularization $\lambda$ is set to 0.1 following [3]. All results are averaged over 4 random seeds.

### 4.2. Evaluation

This section evaluates our approach against existing methods for both few-shot temporal action localisation and prompt learning. To compare with previous works, we report the mean average precision (mAP) at various intersections over union for all results.

#### 4.2.1. THUMOS-14

In Tab 1, we show results for 5-shot 20-way TAL on the THUMOS'14 dataset for our approach *PLOT-TAL CLS*. Adding additional class prompts can improve performance over a single prompt by a large margin (↑ 5.9). We also show how it's possible to achieve higher accuracy by handcrafting prompts (Verbose). In this setting, we use GPT-3.5 [1] to produce additional descriptions of the actions that will replace the class label.

The Baseline $I$ method represents performance when we add additional prompts but exclude optimal transport, demonstrating how optimal transport is highly effective at aligning the features (↑ 15.77). While Baseline $II$ (linear probe) based on the work of [19] and [3] has an average performance of 5% less than our method.

In Fig 3, we demonstrate how the optimal transport improves performance at higher IoU thresholds than single prompt or linear probe methods.

At low IoU thresholds, the predicted segment only needs to overlap with a small section of the ground truth, meaning that single prompt methods and linear probes achieve relatively good performance as they distribute the attention between prompts and features across the temporal domain. However, as we increase the IoU threshold, we can see that our PLOT-TAL method becomes more effective, demonstrating the network's higher discriminative ability.

#### 4.2.2. EPIC-KITCHENS-100

In Tab 2, we show results on the EPIC Kitchens verb and noun partitions, showing a slight improvement over single prompt methods for the noun classes (↑ 1.19) but achieve a more significant performance boost for the verb classes (↑ 2.96).

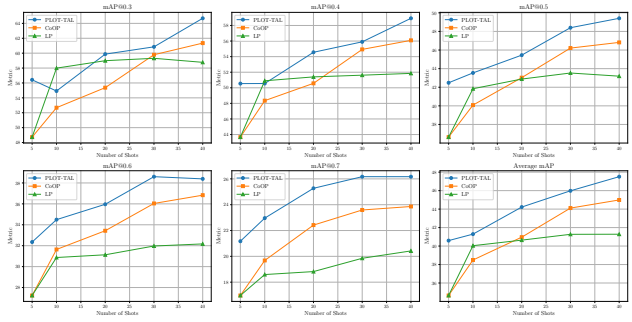


Figure 3. mAP over various IOU thresholds and number of training samples on the THUMOS-14 dataset. Additional prompts demonstrate improved performance, especially at high IoU thresholds indicating improved discriminative ability.

This demonstrates the effectiveness of the additional prompts in distinguishing between complex temporal features. However, the performance improvement is less pronounced for the noun partition. This suggests that nouns, which are generally static and visually distinct, are inherently easier to classify with a single prompt. As a result, they do not derive as much benefit from the added context provided by multiple prompts. Nouns typically represent objects with consistent visual appearances, reducing the need for additional context to disambiguate them. Therefore, the application of optimal transport, which excels in aligning distributions of more dynamic and context-dependent features (such as verbs), does not yield a substantial advantage in this case.

In Tab 3, we compare with other SOTA methods for few-shot temporal action localisation, which utilise meta-learning and perform few-shot localisation at a 5-shot, 5-way setting, whereas our results are from the 5-shot, 20-way configuration. Not only is the 5-shot, 20-way few-shot setting more challenging, but PLOT-TAL also benefits from being trained end-to-end without the requirement for pre-training and episodic adaptive contrastive learning as in current meta-learning approaches.

### 4.3. Qualitative Results

In Fig 4, we show the normalised transport cost for each frame and $N$ embedding for the class label 'Cricket Shot'. This figure shows how each $N$ prompts diverge and focuses on different elements and views within the videos. For example, we can see that $N_1$ or Prompt 1 learns global information across all frames. This shows how we may distribute alignment across all frames in a single prompt framework and lose discriminative ability since it learns global information over the whole video. In the figure, we can

Table 1. Performance comparison of our proposed method PLOT-TAL on the THUMOS-14 dataset against baselines.

| Method | mAP@0.3 | mAP@0.4 | mAP@0.5 | mAP@0.6 | mAP@0.7 | Avg (mAP) |
|---|---|---|---|---|---|---|
| Baseline I (avg) | 37.3 | 32.93 | 26.88 | 18.17 | 8.83 | 24.82 |
| Baseline II (lp) | 51.98 | 46.5 | 36.79 | 25.62 | 14.66 | 35.11 |
| CoOP | 48.73 | 43.67 | 36.64 | 27.24 | 16.97 | 34.65 |
| PLOT-TAL CLS | 53.46 | 48.93 | 38.2 | 30.2 | 18.8 | 38.24 |
| PLOT-TAL Verbose | **56.42** | **50.54** | **42.48** | **32.35** | **21.17** | **40.59** |

Table 2. Few-shot TAL performance (mAP@IoU) on the EPIC-Kitchens-100 Noun and Verb partitions. Our method shows the most significant gains on the more dynamic verb classes, supporting our thesis that it excels at modeling complex temporal structure.

| | **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** |

Table 3. Few-shot TAL performance on THUMOS'14. Our end-to-end (E2E) method is compared against prior meta-learning (ML) work. Note that the evaluation settings are different, making a direct comparison of scores challenging; our 20-way task is significantly harder than the 5-way task.

| **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** |

Figure 4. The normalised transport cost of each $N$ prompt for the class 'Cricket Shot' after training. Prompt one aligns with global information, while the other prompts learn additional, complementary views. In the transport cost algorithm, a lower value indicates closer alignment.

note that Prompt 4 appears to learn background information and is more closely aligned to frames where we can see the stadium stands. Prompts 2 and 3, however, indicate a closer alignment with objects related to the class of 'cricket shot,' including when the cricket strip is in the shot and there are people on the field. The plot clearly reveals that each prompt, when considered in isolation, demonstrates varying degrees of alignment with different sections of the video. This suggests that each prompt is capturing unique and complementary aspects of the video content, allowing for a more nuanced understanding of the temporal dynamics.

## 4.4. Ablation Experiments

We perform several ablation experiments to evaluate each component of the architecture. We experiment with the number of learnable context tokens and prompts per class, alternative feature alignment metrics, and the number of feature pyramid network levels. We also experimented with the types of RGB embeddings and several prompt-engineering strategies.

### 4.4.1. Number of Learnable Prompts

In Tab 4, we perform an ablation experiment on the number of learnable prompts $N$. The results show that the optimum number of prompts is $N = 6$, while with an increased number of prompts, e.g., $N = 10$, we can achieve better results in the more difficult IoU thresholds. This is due to the

Table 4. Ablation on the number of prompts ($N$) per class, evaluated on THUMOS'14. We report mAP (%) at various IoU thresholds. Performance is optimal at $N = 6$. Best result in each column is in bold.

| **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 |

increased temporal discriminative ability of the additional prompts. As the N increases, performance degrades as the model overfits due to the increased number of learnable parameters.

#### 4.4.2. Number of Learnable Context Tokens

Each prompt also has several learnable context tokens as described in [33] and [27]. These context tokens are randomly initialised so that for the class 'Basketball Dunk' with 4 $ctx$ tokens, the full prompt will be

$$P = \{X, X, X, X, \text{Basketball Dunk}\} \tag{5}$$

In Tab 6, we show the effect of varying the number of learnable $ctx$ tokens appended to each prompt. For each $N$ prompt, $n_{ctx}$ tokens are randomly initialised. The figure shows that the optimum number of tokens is between 10 and 20. As per the existing literature [33, 34], we select 16 tokens for all methods unless otherwise stated and train and test using the 5-shot, 20-way setup.

#### 4.4.3. FPN Levels

In Tab 7, we show the effect of increasing or decreasing the number of feature pyramid levels in the network. The results show that six is the optimum number. Additional FPN layers beyond six will tend to increase the number of parameters for optimisation while not providing any additional benefit.

#### 4.4.4. Feature Matching Strategy

To assess the efficacy of using Optimal Transport (OT) with the Sinkhorn Algorithm to align video features with adaptive prompts, we conducted ablation experiments in which OT was replaced with more straightforward distance metrics, precisely Euclidean distance and Hungarian distance. Our goal was to determine the impact of these substitutions on alignment performance and overall method effectiveness.

#### 4.4.5. Euclidean Distance

We replaced the OT metric with the Euclidean distance in the first variant. Here, the alignment between the refined video features $\{\mathbf{x'_1}, \mathbf{x'_2}, \ldots, \mathbf{x'_T}\}$ and the adaptive prompts $P_k$ for each action category $k$ was performed directly using the Euclidean distance:

$$d_{\text{Euc}}(\mathbf{U}, \mathbf{V_k}) = \sum_{t=1}^{T} \sum_{i=1}^{N} \|\mathbf{x'_t} - \mathbf{P_{ki}}\|^2$$

In this formulation, the cost matrix $C_{ti}$ is defined as the squared Euclidean distance between video feature $\mathbf{x'_t}$ and prompt embedding $\mathbf{P_{ki}}$:

$$C_{ti} = \|\mathbf{x'_t} - \mathbf{P_{ki}}\|^2$$

The alignment process involves directly computing the sum of these distances without optimising a transport plan.

#### 4.4.6. Hungarian Distance

In the second variant, we utilised the Hungarian algorithm to find an optimal one-to-one matching between video features and prompts, minimising the overall distance. The cost matrix $C_{ti}$ is defined similarly to the Euclidean distance case, but the Hungarian algorithm ensures a unique assignment of each video feature to a prompt:

$$d_{\text{Hung}}(\mathbf{U}, \mathbf{V_k}) = \min_{\mathbf{T} \in \Pi} \sum_{t=1}^{T} \sum_{i=1}^{N} C_{ti} T_{ti} \tag{6}$$

Here, $\Pi$ represents the set of all possible permutations that allow a one-to-one matching between the sets of video features and prompts. In Tab 8, we show that OT outperforms both methods.

The superior performance of OT can be attributed to several key factors:

- **Global Distribution Matching**: OT aligns the entire distribution of video features with the prompts distribution, considering the global structure and interdependencies within the data. In contrast, Euclidean distance considers each pair independently, which can lead to suboptimal alignments in the presence of complex feature distributions.
- **Flexible Many-to-Many Matching**: OT allows for a many-to-many correspondence between video features and prompts, providing more flexibility in the alignment process. On the other hand, the Hungarian algorithm enforces a strict one-to-one matching, which may not capture the underlying relationships effectively, especially when the number of video features and prompts differ significantly.
- **Entropic Regularization**: The Sinkhorn algorithm introduces entropic regularisation, promoting smoother and

Table 5. Ablation on visual feature embeddings, evaluated on THUMOS'14. Combining RGB and Optical Flow (Flow) features from I3D yields the best performance, highlighting the importance of explicit motion cues for the TAL task.

| 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** |

Table 6. Ablation on the number of context tokens ($n_{\text{ctx}}$) per prompt, evaluated on THUMOS'14. We report mAP (%) at various IoU thresholds. Performance is robust for values between 10-20, with the optimum at $n_{\text{ctx}} = 16$.

| $n_{\text{ctx}}$ | mAP @ IoU | | | | | Avg. |
|---|---|---|---|---|---|---|
| | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | |
| 1 | 52.25 | 46.94 | 40.73 | 31.26 | 20.17 | 38.27 |
| 10 | 54.94 | 49.55 | **42.49** | 31.14 | 20.08 | 39.64 |
| 16 | **56.42** | **50.54** | 42.48 | 32.35 | **21.17** | **40.59** |
| 20 | 53.39 | 48.38 | 42.19 | **33.00** | 20.78 | 39.55 |
| 30 | 50.27 | 45.54 | 38.30 | 29.64 | 18.83 | 36.52 |
| 40 | 53.55 | 47.30 | 40.35 | 31.06 | 19.46 | 38.34 |

Table 7. Ablation on the number of FPN levels, evaluated on THUMOS'14. Performance peaks with a 5-level pyramid.

| 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 |

more stable solutions by avoiding challenging assignments. This regularisation helps mitigate the impact of noisy or outlier features, leading to more robust alignments.

#### 4.4.7. Visual Feature Embeddings

To evaluate the effectiveness of adding motion information via optical flow, we also performed additional experiments using only the RGB embeddings, the optical flow embeddings, and RGB CLIP embeddings from a ViT-B-16 encoder, with results shown in Tab 5. The results show that the CLIP embeddings perform better than the RGB from the I3D network $\uparrow$ 2.67. This is because of the implicit alignment between the image and text encoder embeddings before temporal convolution. However, when combined with optical flow, the performance is improved by a large margin of $\uparrow$ 7.56, demonstrating the enhanced classification ability of the network when we add additional temporal information.

Table 8. Ablation on the prompt alignment strategy, evaluated on THUMOS'14. Our Optimal Transport (OT) approach significantly outperforms both hard-assignment (Kuhn-Munkres) and simple distance-based methods.

| Alignment Method | mAP@0.5 | Avg. mAP (%) |
|---|---|---|
| Euclidean Distance | 21.97 | 22.27 |
| Kuhn-Munkres (Hungarian) | 29.48 | 29.09 |
| **Optimal Transport (OT)** | **43.06** | **40.46** |

## 5. Conclusion

This work addressed a fundamental limitation in prevailing few-shot TAL methods: the inability of a single prompt vector to effectively model the compositional and dynamic nature of human actions from sparse data. We introduced PLOT-TAL, a framework that departs from this paradigm by modeling actions as a distribution of concepts, learned via an ensemble of diverse prompts. Crucially, we demonstrated that Optimal Transport serves not merely as a matching algorithm, but as a powerful structural regularizer that enforces prompt specialization, a key requirement for robust generalization in low-data regimes.

Through extensive experiments on THUMOS'14, EPIC-Kitchens, and ActivityNet 1.3, we established a new state-of-the-art in few-shot TAL without resorting to complex meta-learning schedules. Our analyses, particularly the significant performance gains at high IoU thresholds and the qualitative visualizations of prompt specialization, confirm that our method's success stems from learning a more precise and compositional representation of actions. By moving beyond mean-based representations towards structured, distributional alignments, our work opens a promising new direction for developing more generalizable and data-efficient models for video understanding.

## References


[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[3] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 2, 4, 5

[4] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. *ECCV*, 2022. 2

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 1, 2, 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 4

[7] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 768–784. Springer, 2016. 2

[8] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2014. 4

[9] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 2

[10] Kanchan Keisham, Amin Jalali, Jonghong Kim, and Minho Lee. Multi-level alignment for few-shot temporal action localization. *Information Sciences*, 650:119618, 2023. 4, 6

[11] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 2

[12] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[13] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 2

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 2

[16] Sauradip Nag, Xiatian Zhu, and Tao Xiang. Few-shot temporal action localization with query adaptive transformer. *arXiv preprint arXiv:2110.10552*, 2021. 4, 6

[17] Sauradip Nag, Mengmeng Xu, Xiatian Zhu, Juan-Manuel Pérez-Rúa, Bernard Ghanem, Yi-Zhe Song, and Tao Xiang. Multi-modal few-shot temporal action detection via vision-language meta-adaptation. *arXiv preprint arXiv:2211.14905*, 2022. 2, 6

[18] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *European Conference on Computer Vision*, pages 681–697. Springer, 2022. 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[21] Dingfeng Shi, Qiong Cao, Yujie Zhong, Shan An, Jian Cheng, Haogang Zhu, and Dacheng Tao. Temporal action localization with enhanced instant discriminability. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023. 2

[22] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. *arXiv preprint arXiv:2303.07347*, 2023. 3

[23] Luis Caicedo Torres, Luiz Manella Pereira, and M Hadi Amini. A survey on optimal transport for machine learning: Theory and applications. *arXiv preprint arXiv:2106.01963*, 2021. 2

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[25] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009. 2

[26] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[27] Yuetian Weng, Zizheng Pan, Mingfei Han, Xiaojun Chang, and Bohan Zhuang. An efficient spatio-temporal pyramid transformer for action detection. In *ECCV*, 2022. 7

[28] Huijuan Xu, Ximeng Sun, Eric Tzeng, Abir Das, Kate Saenko, and Trevor Darrell. Revisiting few-shot activity detection with class similarity control. *arXiv preprint arXiv:2004.00137*, 2020. 2

[29] Hongtao Yang, Xuming He, and Fatih Porikli. One-shot action localization by learning sequence matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2018. 2

[30] Pengwan Yang, Vincent Tao Hu, Pascal Mettes, and Cees GM Snoek. Localizing the common action among a few videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 505–521. Springer, 2020. 6

[31] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 492–510. Springer, 2022. 2, 3

[32] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12993–13000, 2020. 4

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 7

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 5, 6