Supplementary Material: Is Temporal Prompting All We Need For Limited Labeled Action Recognition?

Shreyank N Gowda[1] Boyan Gao[2] Xiao Gu[2] Xiaobo Jin[3]

[1]University of Nottingham [2]University of Oxford [3]Xi'an Jiaotong-Liverpool University

1. Results on TruZe [4]

We also evaluate on the more challenging TruZe split. Whilst our pre-training dataset does not have any overlapping classes with any of the testing datasets, we run on this split to show even higher performance gains than in the main paper. The proposed UCF101 [10] and HMDB51 [7] splits have 70/31 and 29/22 classes (represented as training/testing). We compare to WGAN [11], OD [8] and E2E [1] on both ZSL and GZSL scenarios. Results are shown in Table 1.

Method	UCF101		HMDB51	
	ZSL	GZSL	ZSL	GZSL
WGAN [11]	22.5	36.3	21.1	31.8
OD [8]	22.9	42.4	21.7	35.5
E2E [1]	45.5	45.9	31.5	38.9
SPOT [2]	25.5	44.1	24.0	37.1
TP-CLIP	**79.6**	**80.1**	**48.6**	**51.8**

Table 1. Results on TruZe. For ZSL, we report the mean class accuracy and for GZSL, we report the harmonic mean of seen and unseen class accuracies. All approaches use sen2vec annotations as the form of semantic embedding and not Stories, for fair comparison.

2. Evaluating Effect of Semantic Embeddings

We conducted an in-depth investigation of different semantic embeddings and their effects on model performance, comparing our TP-CLIP framework against other leading approaches on GZSL tasks. Our analysis covered a range of semantic representations - from manually crafted embeddings to the narrative-based Stories approach described in the literature. For our TP-CLIP model, we tested both manual embeddings and the Stories method, while also benchmarking against competitors using manual annotations, word2vec, sen2vec, and Stories [3] embeddings. This comprehensive comparison across multiple state-of-the-art models helped us understand how different semantic representations influence performance in challenging zero-shot learning scenarios, and demonstrated the particular advantages of our temporal prompting technique when combined with well-chosen semantic embeddings.

3. Theoretical Analysis

3.1. Temporal Representation Capacity of Visual Prompts

Theorem 1 (Temporal Representation Capacity)**.** *Let $F_{1:T} = \{f_1, f_2, \ldots, f_T\}$ be a sequence of video frames, each encoded by CLIP's [9] image encoder E_{image} to produce frame embeddings $\{e_1, e_2, \ldots, e_T\}$. The temporal visual prompting mechanism TE with dimension d has sufficient capacity to capture temporal dynamics between frames with an approximation error bounded by $O\left(\frac{1}{d}\right)$ relative to an ideal temporal encoder with unlimited capacity, when the temporal relationships exhibit Lipschitz continuity.*

Proof. First, we define the frame embeddings produced by CLIP's image encoder:

$$e_t = E_{\text{image}}(f_t) \in \mathbb{R}^D \tag{1}$$

In our TP-CLIP model, the temporal encoder TE processes the sequence of frame embeddings to produce a temporal context vector:

$$T_{\text{context}} = TE(e_1, e_2, \ldots, e_T) \tag{2}$$

The key insight is that any temporal relationship between frames can be modeled as a function $\phi : \mathbb{R}^{D \times T} \to \mathbb{R}^d$ that maps the sequence of frame embeddings to a temporal representation.

Let ϕ^* represent the ideal temporal encoder with unlimited capacity. We need to show that our temporal prompting mechanism TE can approximate ϕ^* within a bounded error.

The temporal encoding in TP-CLIP is performed through a 1D convolutional layer followed by a fully connected

Model	SE	Olympics			HMDB51			UCF-101		
		u	s	H	u	s	H	u	s	H
WGAN [11]	M	50.8	71.4	59.4	-	-	-	30.4	**83.6**	44.6
OD [8]	M	61.8	71.1	66.1	-	-	-	36.2	76.1	49.1
CLASTER [5]	M	66.2	71.7	68.8	-	-	-	40.2	69.4	50.9
TP-CLIP	M	**71.6**	**76.9**	**74.2**	-	-	-	**43.1**	77.5	**54.6**
WGAN [11]	W	35.4	65.6	46.0	23.1	55.1	32.5	20.6	73.9	32.2
OD [8]	W	41.3	72.5	52.6	25.9	55.8	35.4	25.3	74.1	37.7
CLASTER [5]	W	49.2	71.1	58.1	35.5	52.8	42.4	30.4	68.9	42.1
WGAN [11]	S	36.1	66.2	46.7	28.6	57.8	38.2	27.5	74.7	40.2
OD [8]	S	42.9	73.5	54.1	33.4	57.8	42.3	32.7	75.9	45.7
CLASTER [5]	S	49.9	71.3	58.7	42.7	53.2	47.4	36.9	69.8	48.3
CLASTER [5]	C	66.8	71.6	69.1	43.7	53.3	48.0	40.8	69.3	51.3
WGAN [11]	Sto	52.5	73.4	61.2	35.2	65.1	45.7	33.8	**84.2**	48.2
OD [8]	Sto	63.3	75.1	68.7	37.2	**67.5**	47.9	40.1	81.7	53.8
CLASTER [5]	Sto	69.1	74.1	71.5	44.3	57.2	49.9	42.1	71.5	53.0
GIL [6]	Sto	-	-	-	52.8	57.8	55.1	**68.2**	**89.8**	**77.5**
TP-CLIP	Sto	**79.4**	**84.4**	**81.8**	**55.5**	60.7	**58.0**	49.1	81.7	61.3

Table 2. Seen and unseen accuracies for TP-CLIP by fine-tuning on different datasets using different embeddings. 'SE' corresponds to the type of embedding used, wherein 'M', 'W', 'S', 'C' and 'Sto' refers to manual annotations, word2vec, sen2vec, combination of the embeddings and Stories respectively. 'u', 's' and 'H' corresponds to average unseen accuracy, average seen accuracy and the harmonic mean of the two. All the reported results are on the same splits.

layer:

$$V_{\text{conv}} = \text{Conv1D}(e_1 : e_2 : \ldots : e_T) \tag{3}$$

$$T_{\text{context}} = \text{ReLU}(\text{FC}(V_{\text{conv}})) \tag{4}$$

By the Universal Approximation Theorem for neural networks with ReLU activations, given sufficient width d, our temporal encoder can approximate any continuous function mapping from the input space to the output space.

For temporal relationships that exhibit Lipschitz continuity (which is a reasonable assumption for most natural videos where adjacent frames are similar), the approximation error is bounded by:

$$\|TE(F_{1:T}) - \phi^*(F_{1:T})\|_2 \leq \frac{C}{d} \tag{5}$$

where C is a constant dependent on the Lipschitz constant of the temporal relationships.

The combined spatial-temporal encoding for frame t is then:

$$v(t) = [E_{\text{image}}(f_t); TE(F_{1:T})] \tag{6}$$

This concatenation ensures that both spatial information (from $E_{\text{image}}(f_t)$) and temporal context (from $TE(F_{1:T})$) are preserved.

Furthermore, by the Johnson-Lindenstrauss lemma, the projection into a d-dimensional space preserves pairwise distances between temporal patterns with high probability when $d = O(\log(T)/\epsilon^2)$, where ϵ is the distortion factor.

Therefore, our temporal visual prompting mechanism has sufficient capacity to capture temporal dynamics with an approximation error bound of $O\left(\frac{1}{d}\right)$ relative to an ideal temporal encoder with unlimited capacity. □

Corollary 2. *The expressive capacity of temporal visual prompting grows linearly with the dimension of the prompt space, making it possible to achieve strong performance with a compact representation.*

This theorem provides the theoretical foundation for why temporal prompting is sufficient for action recognition with limited labeled data, explaining the strong empirical results observed in the experiments.

3.2. Information Transfer through Temporal Prompting

Theorem 3 (Information Preservation in Temporal Prompting)**.** *Let $E_{image} : \mathcal{F} \to \mathbb{R}^D$ be the pre-trained CLIP image encoder with frozen weights θ_{CLIP}. For a video sequence $F_{1:T} = \{f_1, f_2, \ldots, f_T\}$, the TP-CLIP architecture with learnable temporal prompting parameters $\theta_{TP} \ll |\theta_{CLIP}|$ can preserve a $(1-\delta)$ fraction of the mutual information between the temporal dynamics and class labels without modifying the original CLIP architecture.*

Proof. Let $\mathcal{Y}$ be the space of action classes and $\mathcal{T}$ be the space of temporal patterns in videos. The mutual informa-

tion between temporal patterns and class labels is given by:

$$I(\mathcal{T};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{T}) \tag{7}$$

For a standard image-based model like CLIP, each frame is processed independently:

$$e_t = E_{\text{image}}(f_t;\theta_{\text{CLIP}}) \tag{8}$$

The information captured by independently processing frames is:

$$I_{\text{indep}} = I(\{e_1, e_2, \ldots, e_T\};\mathcal{Y}) \tag{9}$$

However, this approach fails to model the temporal dependencies:

$$I_{\text{indep}} < I(\mathcal{T};\mathcal{Y}) \tag{10}$$

In TP-CLIP, we introduce temporal prompting:

$$T_{\text{context}} = \text{TE}(e_1, e_2, \ldots, e_T;\theta_{\text{TP}}) \tag{11}$$

$$v(t) = [e_t; T_{\text{context}}] \tag{12}$$

The key insight is that by concatenating the temporal context T_{context} with the frame embeddings, we create a representation that preserves temporal information without modifying the original CLIP architecture.

Let $\mathcal{V} = \{v(1), v(2), \ldots, v(T)\}$ be the set of enhanced frame representations. We can establish the following inequality:

$$I(\mathcal{V};\mathcal{Y}) \geq (1-\delta)\cdot I(\mathcal{T};\mathcal{Y}) \tag{13}$$

where δ is a small constant that depends on the complexity of the temporal patterns and the dimension of the temporal context.

This is because:

$$I(\mathcal{V};\mathcal{Y}) = I(\{e_t; T_{\text{context}}\}_{t=1}^{T};\mathcal{Y}) \tag{14}$$

$$\geq I(\{e_t\}_{t=1}^{T};\mathcal{Y}) + I(T_{\text{context}};\mathcal{Y}|\{e_t\}_{t=1}^{T}) \tag{15}$$

The temporal encoder TE is designed to capture temporal dependencies, ensuring that:

$$I(T_{\text{context}};\mathcal{Y}|\{e_t\}_{t=1}^{T}) \approx I(\mathcal{T};\mathcal{Y}|\{e_t\}_{t=1}^{T}) \tag{16}$$

Furthermore, the Data Processing Inequality ensures that the information content does not increase through processing, which means the upper bound of information is preserved:

$$I(\mathcal{V};\mathcal{Y}) \leq I(\mathcal{T};\mathcal{Y}) \tag{17}$$

The parameter efficiency comes from the fact that $|\theta_{\text{TP}}| \ll |\theta_{\text{CLIP}}|$. Specifically, if we denote the number of parameters in the temporal encoder as $|\theta_{\text{TP}}|$ and in CLIP as $|\theta_{\text{CLIP}}|$, then:

$$\frac{|\theta_{\text{TP}}|}{|\theta_{\text{CLIP}}|} = O\left(\frac{1}{|\theta_{\text{CLIP}}|}\right) \tag{18}$$

Therefore, with a minimal number of additional parameters θ_{TP}, TP-CLIP can preserve a $(1-\delta)$ fraction of the mutual information between temporal patterns and class labels, without modifying the original CLIP architecture. □

Corollary 4. *The TP-CLIP framework achieves efficient temporal modeling with parameter count scaling as $O(d^2)$ where d is the embedding dimension, compared to $O(Td^2)$ required for full cross-attention mechanisms in alternative approaches.*

This theorem explains why TP-CLIP can effectively capture temporal information without modifying CLIP's core architecture, leading to parameter efficiency while maintaining or improving performance on video understanding tasks.

References

[1] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4613–4623, 2020. 1

[2] Shreyank N Gowda. Synthetic sample selection for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 58–67, 2023. 1

[3] Shreyank N Gowda and Laura Sevilla-Lara. Telling stories for common sense zero-shot action recognition. In *Proceedings of the Asian Conference on Computer Vision*, pages 4577–4594, 2024. 1

[4] Shreyank N Gowda, Laura Sevilla-Lara, Kiyoon Kim, Frank Keller, and Marcus Rohrbach. A new split for evaluating true zero-shot action recognition. In *DAGM German Conference on Pattern Recognition*, pages 191–205. Springer, 2021. 1

[5] Shreyank N Gowda, Laura Sevilla-Lara, Frank Keller, and Marcus Rohrbach. Claster: clustering with reinforcement learning for zero-shot action recognition. In *European Conference on Computer Vision*, pages 187–203. Springer, 2022. 2

[6] Shreyank N Gowda, Davide Moltisanti, and Laura Sevilla-Lara. Continual learning improves zero-shot action recognition. In *Proceedings of the Asian Conference on Computer Vision*, pages 3239–3256, 2024. 2

[7] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 1

[8] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9985–9993, 2019. 1, 2

[9] 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

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[11] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 1, 2