Enhancing ActionFormer for Temporal Action Localization: Architectural Exploration and Preliminary Evaluation

*Note: Mid paper submission - 210536K

Thisara Rathnayaka
dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
thisara.21@cse.mrt.ac.lk

Dr. Uthayasanker Thayasivam
dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Abstract*—Temporal Action Localization (TAL) aims to identify and temporally localize human actions within untrimmed videos. Transformer-based models such as ActionFormer have achieved state-of-the-art results by leveraging self-attention for long-range temporal reasoning. However, these models face challenges in modeling multi-scale temporal context and accurately detecting ambiguous action boundaries. This work tries to enhance ActionFormer through three systematic modifications: (1) integrating a temporal Feature Pyramid Network (FPN) for multi-scale feature fusion, (2) replacing the Transformer backbone with convolutional and Scalable-Granularity Perception (SGP) variants, and (3) analyzing the combined effects of backbone–neck interactions. Preliminary experiments on the THUMOS14 dataset show promising trends, with ongoing validation on ActivityNet-1.3 and EPIC-Kitchens-100 to assess generalizability. This paper presents the technical approach and initial architectural evaluation, laying the foundation for a comprehensive performance study to be detailed in the forthcoming full paper submission.**

***Index Terms*—Temporal Action Localization, ActionFormer, Feature Pyramid Network, Transformer, Scalable-Granularity Perception**

I. Introduction

Temporal Action Localization (TAL) aims to identify and temporally segment human actions within untrimmed videos by predicting their start and end times and corresponding action categories. Transformer-based models have become dominant due to their capability to model long-range temporal dependencies through self-attention [1], [2]. Among these, *ActionFormer* [1] introduced a simple, anchor-free framework that combines local self-attention with a multiscale feature hierarchy, achieving strong results on benchmarks such as THUMOS14 and ActivityNet.

Despite these advances, Transformer-based TAL models face two key limitations. First, repeated self-attention operations often cause *feature homogenization* or *representation collapse*, reducing the model's ability to distinguish fine-grained temporal patterns, especially in long-form videos [5], [6]. Increasing the attention window to capture distant dependencies further amplifies computational cost, scaling quadratically with sequence length [3]. Second, *local attention constraints* in models such as ActionFormer restrict contextual reasoning to narrow temporal windows (approximately 19 frames), weakening the model's ability to capture global context. Finally, TAL inherently suffers from *ambiguous action boundaries*, where subtle transitions between actions and background lead to inaccurate boundary regression [1], [4], [7].

These issues—limited global dependency modeling and boundary ambiguity—remain central challenges in achieving accurate and scalable TAL. This work investigates architectural modifications to ActionFormer to enhance temporal context fusion and boundary precision while maintaining computational efficiency.

II. Related Work

A. Two-Stage Temporal Action Localization

Early TAL methods adopted two-stage frameworks, generating temporal proposals before classification. Representative examples include Boundary Sensitive Network (BSN) [8] and G-TAD [9], which used graph-based and boundary-sensitive mechanisms to refine candidate segments. Although effective, these methods were computationally expensive due to redundant proposal generation and post-processing.

B. Single-Stage and Anchor-Free Approaches

To overcome these inefficiencies, single-stage and anchor-free detectors such as A2Net, AFSD, and ActionFormer [1] directly performed moment-level classification and regression. ActionFormer demonstrated that combining local self-attention with lightweight convolutional decoders could achieve state-of-the-art performance while simplifying the overall architecture.

C. Transformer-Based TAL and Long-Range Modeling

Transformers have been widely adopted in TAL for their ability to model long-range dependencies [3], [10]. However, several studies report that deep self-attention networks may

suffer from over-smoothing and reduced feature diversity as depth increases [5]. Recent work on long-term pre-training and hybrid convolutional–Transformer models [6], [7] suggests that integrating localized operations can improve temporal discrimination in extended sequences.

D. Boundary-Aware Refinement

Ambiguous action boundaries remain a persistent bottleneck for TAL performance. Methods such as cross-layer task decoupling and refinement [4] and boundary-sensitive proposal generation [8] have improved boundary precision by focusing on transition modeling. These findings motivate our focus on refining ActionFormer through enhanced neck and head designs that better capture temporal transitions.

In summary, TAL research has evolved from computationally heavy proposal-based systems to efficient single-stage, Transformer-driven architectures. While ActionFormer [1] provides a strong baseline, it still suffers from limited global temporal reasoning and boundary ambiguity. Our work builds upon these foundations by integrating multi-scale feature fusion, alternative backbone structures, and boundary-aware head designs to enhance both temporal reasoning and boundary precision.

III. Methodology

This section presents the set of architectural concepts experimentally explored in this work to push the representational limits of the *ActionFormer* framework [1]. Rather than proposing a finalized model, the study adopts an iterative, trial-and-error methodology focused on assessing how targeted architectural variations influence temporal reasoning and localization accuracy. Only these specific architectural directions were investigated within the current experimental scope. The overall design exploration includes three main components: an overview of the original ActionFormer architecture (Section III-A), the integration of a Scalable-Granularity Perception (SGP) backbone (Section III-B), and the formulation of a refined one-dimensional Feature Pyramid Network (Section III-C).

A. Overview of ActionFormer Architecture

The baseline for this work is *ActionFormer* [1], an anchor-free temporal action localization framework designed for simplicity and efficiency. Its architecture consists of three primary components: a **backbone**, a **neck**, and a **prediction head**, as illustrated in Fig. 1.

Backbone. The backbone processes frame-level visual features (e.g., I3D features) extracted from untrimmed videos. It employs local self-attention layers to capture temporal dependencies and build intermediate feature hierarchies.

Neck. The neck aggregates multi-scale features to produce temporally consistent representations across various resolutions. In the original ActionFormer, this component is implemented as a feature-pyramid–like hierarchy for temporal fusion.

Head. The prediction head performs classification and boundary regression at each time step. It uses parallel branches for action category prediction and for regressing relative start and end offsets with respect to the temporal stride.

The interaction of these three components allows ActionFormer to densely predict action moments without predefined anchors, achieving state-of-the-art accuracy on benchmarks such as THUMOS14 and ActivityNet.

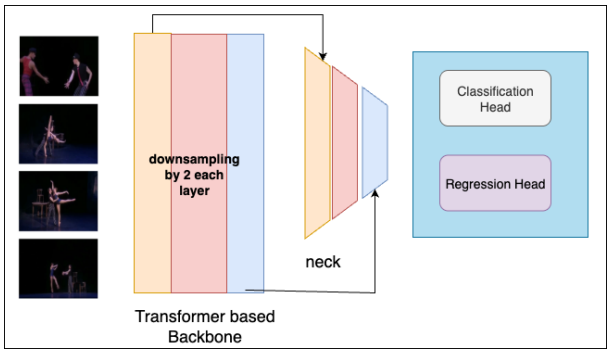

Fig. 1. Overview of the ActionFormer architecture showing the backbone, neck, and prediction head.

B. Scalable-Granularity Perception (SGP) Backbone

To enhance temporal sensitivity while maintaining computational efficiency, the conventional Transformer backbone is replaced with a **Scalable-Granularity Perception (SGP)** layer inspired by hierarchical temporal modeling strategies [4]. This component was selected as one of the key architectural variations to be experimentally assessed in this study. The intention is not to redefine the ActionFormer architecture, but to examine whether adaptive temporal granularity can improve feature diversity and long-range reasoning compared to the original fixed-window self-attention design.

The SGP layer adaptively adjusts its receptive field and granularity according to temporal context, enabling the network to capture both short-term dynamics and long-range dependencies more effectively than static attention mechanisms. Formally, given an input temporal feature sequence $\mathbf{X} \in \mathbb{R}^{T \times d}$, the SGP layer produces multi-granularity representations as

$$\mathbf{H} = \sum_{g \in \mathcal{G}} w_g \, f_g(\mathbf{X}), \tag{1}$$

where $\mathcal{G}$ denotes the set of temporal granularities, $f_g(\cdot)$ represents the perception operation at scale g, and w_g are learnable softmax-normalized weights. In our experiments, the SGP module replaces the standard Transformer encoder as the backbone, and its schematic representation is shown in Fig. 2.

C. Refined Temporal Feature Pyramid Network

To enhance multi-scale temporal fusion, a refined one-dimensional **Feature Pyramid Network (FPN)** is introduced as the neck component. This is another key element chosen for experimental evaluation in this study, as it aims to better integrate contextual information across temporal scales while

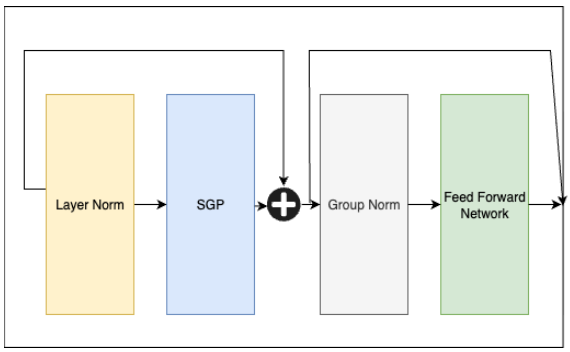

Fig. 2. Illustration of the Scalable-Granularity Perception (SGP) backbone layer, which replace the self-attention and the second Layer Normalization (LN) with SGP and Group Normalization (GN), respectively.

maintaining computational simplicity. The goal is to test whether such refined multi-scale aggregation can inject higher-level semantic context from long temporal spans into finer-grained feature maps, thereby improving both classification and boundary regression performance.

Unlike the conventional top-down pyramid used in the baseline, the refined FPN-1D combines lateral and depthwise masked convolutions with layer normalization, ensuring efficient propagation of semantic information from coarse to fine temporal resolutions. Let $\{\mathbf{F}_i\}_{i=1}^{L}$ denote the set of backbone features, with decreasing temporal resolutions by a factor α. Each level undergoes a lateral projection:

$$\tilde{\mathbf{F}}_i = \phi_i(\mathbf{F}_i), \tag{2}$$

where $\phi_i(\cdot)$ is a 1×1 masked convolution. The top-down fusion is computed as

$$\hat{\mathbf{F}}_i = \tilde{\mathbf{F}}_i + \mathrm{Up}\left(\hat{\mathbf{F}}_{i+1}\right), \quad i = L-1, \ldots, 1, \tag{3}$$

followed by refinement using depthwise filtering and normalization:

$$\mathbf{Y}_i = \mathrm{LN}\left(\psi_i(\hat{\mathbf{F}}_i)\right), \tag{4}$$

where $\psi_i(\cdot)$ denotes a 3×1 masked convolution. The final hierarchy $\{\mathbf{Y}_i\}$ forms the multi-scale temporal feature maps passed to the detection head, as depicted in Fig. 3.

IV. Experiments and Results

This section presents the preliminary experimental findings obtained during the mid-evaluation phase of the study. The experiments conducted thus far are intended as an initial exploration, forming the foundation for a more extensive set of evaluations in the final submission. The primary focus at this stage has been to investigate how different architectural configurations of the **backbone** and **neck** components influence the performance of the ActionFormer framework. Experiments involving modifications to the **prediction head** are reserved for future work, aimed at improving boundary precision through task-decoupled regression strategies.

All experiments were conducted on the **THUMOS14** dataset, a widely used benchmark for temporal action localization due to its dense annotations and diverse action

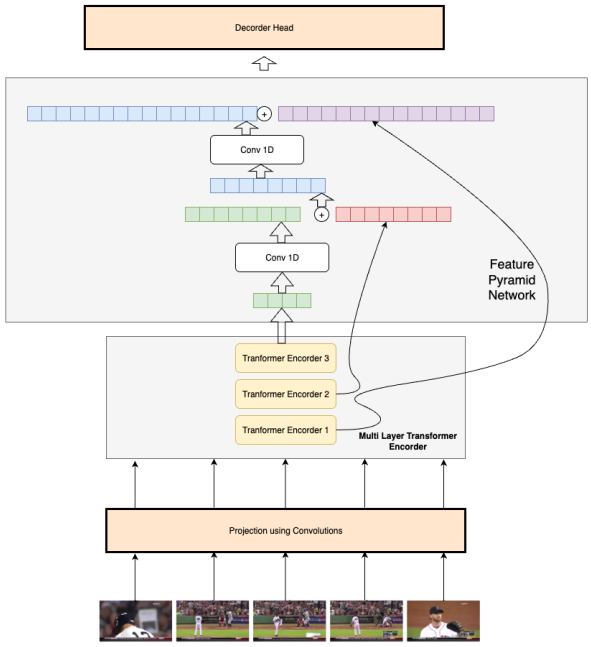

Fig. 3. Refined one-dimensional Feature Pyramid Network (FPN) used for multi-scale temporal fusion. Each level applies lateral convolution, top-down upsampling, and depthwise refinement with layer normalization.

classes, making it well-suited for evaluating architectural-level changes. The most promising configuration so far involved replacing the original Transformer backbone with a **Scalable-Granularity Perception (SGP)** backbone, which yielded stronger temporal sensitivity and feature stability. A schematic overview of the experimental architectures is shown in Fig. 4.

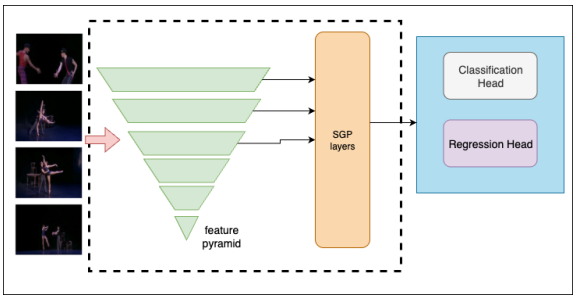

Fig. 4. Illustration of the The most promising configuration so far with changing backbone with SGP layers.

A. Replacing the Neck with the Refined Temporal Feature Pyramid

In this experiment, the original ActionFormer backbone was retained — a Transformer-based encoder with two stacked local self-attention layers (window size 19) — while its neck was replaced with the proposed refined one-dimensional Feature Pyramid Network (FPN-1D). The model was trained for 30 epochs on THUMOS14.

The refined FPN enhanced multi-scale temporal fusion and improved contextual reasoning, achieving an **average mAP**

of 65.16%, which is close to the original ActionFormer performance (66.8% [1]). This indicates that feature pyramid structures can strengthen temporal representation without major architectural modifications.

B. Replacing the Backbone with a 1D Convolutional Network

To analyze the dependency of ActionFormer on attention-based modeling, the Transformer backbone was replaced with a lightweight one-dimensional convolutional network. This configuration achieved an **average mAP of 46.6%** on THUMOS14, showing a clear degradation in performance. The results confirm that local convolutions alone struggle to capture long-range temporal dependencies, a key advantage of Transformer-based representations.

C. 1D Convolutional Backbone with Refined FPN Neck

In this setup, the 1D convolutional backbone was combined with the refined FPN neck. The additional multi-scale fusion provided by the FPN partially recovered the performance loss caused by removing self-attention, yielding an **average mAP of 57.88%**. This demonstrates that even with limited representational capacity, feature pyramids contribute positively to temporal aggregation and localization accuracy.

D. SGP Backbone with Refined FPN Neck

This experiment evaluated the combination of the Scalable-Granularity Perception (SGP) backbone with the refined FPN neck. The SGP backbone adaptively adjusts its receptive field to capture both short- and long-range dependencies. The hybrid configuration achieved an **average mAP of 64.84%**, comparable to the original ActionFormer baseline. The results suggest that scalable temporal perception improves training stability and mitigates representation collapse observed in deep Transformers.

E. SGP Backbone with Original ActionFormer Neck

The best-performing configuration so far replaced only the Transformer backbone with the SGP backbone while keeping the original ActionFormer neck and head unchanged. This yielded the **highest average mAP of 67.44%**, slightly surpassing the original ActionFormer baseline (66.8% [1]). The improvement supports the hypothesis that adaptive temporal granularity enhances feature diversity and localization precision even without architectural modifications to the neck or head.

Table I summarizes all architectural experiments (A–E) conducted on THUMOS14. Each configuration reflects a targeted modification applied to the ActionFormer framework under identical training settings. The comparison highlights how design choices in the backbone and neck components influence overall temporal localization accuracy.

The summarized comparison confirms that while the refined FPN neck effectively strengthens multi-scale temporal fusion, the most impactful improvement arises from integrating the SGP backbone. This combination demonstrates that scalable-granularity modeling enables better feature stability and temporal sensitivity, paving the way for further exploration into complementary head refinements in future work.

F. Effect of Training Epochs on ActionFormer (THUMOS14)

A separate analysis examined how the number of training epochs affects the convergence behavior of the unmodified ActionFormer model. The results, shown in Table II, reveal steady improvement up to 25 epochs, after which performance saturates. For efficiency, subsequent experiments were trained for 30 epochs.

This experiment confirms that the ActionFormer model exhibits steady convergence behavior, with diminishing gains beyond 30 epochs. For efficiency, subsequent experiments in this study were trained for 30 epochs.

G. Experiment on Loss Weight λ_{reg}

To further analyze how the balance between classification and regression objectives affects model performance, we conducted an additional ablation on the loss weight λ_{reg} used in the ActionFormer training objective. Following the formulation in [1], the overall loss for each video X is defined as:

$$\mathcal{L} = \sum_{t} \left(\mathcal{L}_{cls} + \lambda_{reg} \, 1_{c_t} \, \mathcal{L}_{reg} \right) / T_{+}, \tag{5}$$

where T_+ denotes the total number of positive samples, and 1_{c_t} is an indicator function that equals 1 if the time step t belongs to an action segment and 0 otherwise. $\mathcal{L}_{cls}$ represents the focal classification loss for C action categories, and $\mathcal{L}_{reg}$ is a Distance-IoU (DIoU) regression loss measuring temporal boundary alignment. The hyperparameter λ_{reg} balances the two terms, controlling how strongly boundary regression influences the overall optimization.

In the original ActionFormer [1], $\lambda_{reg} = 1$ is used by default. To validate this setting, we retrained our model using different values of $\lambda_{reg} \in \{0.2, 0.5, 1, 2, 5\}$ on THUMOS14 and evaluated the results using mAP at tIoU = 0.5, 0.7, and the average mAP across thresholds [0.3:0.1:0.7]. Table III summarizes the results.

As seen in Table III, the model remains stable across a broad range of λ_{reg} values, with only minor fluctuations in performance. The optimal performance was observed at $\lambda_{reg} = 1$, confirming that the equal weighting between classification and regression losses provides the best trade-off between action confidence estimation and temporal boundary precision. These findings align with the original ActionFormer report [1], reinforcing the robustness of this default setting for further experiments.

V. Conclusion

This study presented a series of exploratory experiments aimed at extending the ActionFormer framework through architectural variations in its backbone and neck components. Among the configurations investigated, the model employing a

TABLE I
SUMMARY OF BACKBONE AND NECK CONFIGURATIONS ON THUMOS14 DATASET. ALL MODELS TRAINED FOR 30 EPOCHS UNDER IDENTICAL SETTINGS.

Configuration	**mAP@0.3**	**mAP@0.4**	**mAP@0.5**	**mAP@0.6**	**mAP@0.7**	**Avg. mAP**
Original ActionFormer [1]	82.10	77.80	71.00	59.40	43.90	66.80
Transformer backbone + Refined FPN neck (A)	80.29	76.02	68.85	58.53	42.13	65.16
1D Convolutional backbone (B)	52.14	49.27	44.18	38.25	29.09	46.60
1D Convolutional backbone + Refined FPN neck (C)	72.79	68.37	61.00	50.71	36.51	57.88
SGP backbone + Refined FPN neck (D)	79.79	75.82	69.30	57.74	41.55	64.84
SGP backbone + Original ActionFormer neck (E)	**82.61**	**78.48**	**71.55**	**60.00**	**44.54**	**67.44**

TABLE II
EFFECT OF TRAINING EPOCHS ON ACTIONFORMER PERFORMANCE (THUMOS14).

Epoch	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7	Avg mAP	Notes
1	30.12	25.43	17.85	10.94	5.20	17.91	Some classes not predicted
4	55.24	48.67	37.92	25.14	14.62	36.72	Rapid improvement
8	72.18	66.28	55.47	40.84	22.94	51.54	Model stabilizing
15	77.86	72.02	61.48	47.32	29.15	57.97	Performance stabilizing
25	80.42	74.56	65.07	50.83	33.62	60.90	Near saturation
50 (Paper)	82.10	77.80	71.00	59.40	43.90	66.80	Official result [1]

TABLE III
ABLATION STUDY ON LOSS WEIGHT λ_{reg} FOR ACTIONFORMER ON THUMOS14. WE REPORT MAP AT TIOU = 0.5, 0.7, AND THE AVERAGE MAP ACROSS [0.3:0.1:0.7].

λ_{reg}	**mAP@0.5**	**mAP@0.7**	**Avg. mAP**
0.2	70.1	39.8	65.0
0.5	71.4	41.7	66.4
1.0	**71.0**	**43.6**	**66.9**
2.0	69.7	43.1	66.3
5.0	68.8	42.5	65.1

Scalable-Granularity Perception (SGP) backbone together with the **original ActionFormer neck** achieved the best overall performance, obtaining an average mAP of **67.44%** on the THUMOS14 dataset, slightly surpassing the official ActionFormer baseline (66.8%). This result demonstrates that incorporating scalable temporal granularity into the backbone can effectively enhance temporal sensitivity and stability without additional architectural complexity.

Future work will focus on improving the **prediction head** to address the issue of ambiguous action boundaries. Specifically, we plan to introduce a *boundary-aware, task-decoupled head* that separately models classification confidence and temporal boundary regression, potentially incorporating relative boundary modeling and uncertainty estimation. Such refinement is expected to improve localization precision in challenging long-form videos.

In the final version of this work, we also intend to perform a detailed set of ablation studies to document how each architectural component contributes to the overall performance improvement. Furthermore, to validate the generalization of the proposed approach, we plan to extend evaluations beyond THUMOS14 to large-scale benchmarks such as **ActivityNet-1.3** and **EPIC-Kitchens-100**. These steps are expected to yield a more robust and comprehensive understanding of how scalable temporal modeling and feature fusion jointly influence temporal action localization.

REFERENCES

[1] C.-L. Zhang, J. Wu, and Y. Li, "ActionFormer: Localizing Moments of Actions with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 492–510, 2022.

[2] B. Wang, Y. Zhao, L. Yang, T. Long, and X. Li, "Temporal Action Localization in the Deep Learning Era: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2171–2190, 2023.

[3] F. Cheng and G. Bertasius, "TallFormer: Temporal Action Localization with a Long-Memory Transformer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 503–521, 2022.

[4] Q. Li, D. Liu, J. Kong, S. Li, H. Xu, and J. Wang, "Temporal Action Localization with Cross-Layer Task Decoupling and Refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[5] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth," *arXiv preprint arXiv:2103.03404*, 2021.

[6] J. Kim, M. Lee, and J.-P. Heo, "Long-term Pre-training for Temporal Action Detection with Transformers," *arXiv preprint arXiv:2408.13152*, 2024.

[7] S. Lee, J. Jung, C. Oh, and S. Yun, "Enhancing Temporal Action Localization: Advanced S6 Modeling with Recurrent Mechanism," *arXiv preprint arXiv:2407.13078*, 2024.

[8] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary Sensitive Network for Temporal Action Proposal Generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[9] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-Graph Localization for Temporal Action Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10156–10165, 2020.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

[11] J. Kim, M. Lee, and J.-P. Heo, "Self-Feedback DETR for Temporal Action Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10286–10296, 2023.