Causal Yet Future-Aware: Dual-Path Temporal Modeling for Online Action Segmentation

Zhichao Zheng[1], Peirong Ma[1], Ying Zhou[2⋆], Li Kong[1], and Junsheng Zhou[1]

[1] School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, China
{zhengzhichao, prma, zhoujs}@njnu.edu.cn, kongli@nnu.edu.cn
[2] School of Computer and Software, Nanjing University of Industry Technology, China
Zhouying@niit.edu.cn

Abstract. Temporal Action Segmentation (TAS) assigns action labels to each frame by modeling the temporal structure of actions. While offline TAS leverages full temporal context, online TAS cannot directly access future information during inference. Existing online TAS methods mainly focus on modeling historical dependencies, but temporal ambiguity still arises due to the incomplete temporal dependencies, leading to fragmented predictions. To tackle this challenge, we propose Dual-Path Temporal Modeling (DPTM) for online TAS, a two-stage framework that explicitly models and transfers the future dependencies that are inaccessible during online inference. In the first stage, we adopt an offline-style training scheme where bidirectional dependencies are decomposed into unidirectional historical and future components via a temporal decomposition strategy. In the second stage, we employ a predictive distillation strategy that uses recent historical frames to predict the separated future dependencies without violating causality. These designs enable DPTM to leverage more complete temporal context while remaining suitable for online inference. Extensive experiments demonstrate that DPTM significantly improves temporal coherence in segmentation results and achieves state-of-the-art performance, demonstrating the effectiveness of modeling dependencies beyond historical frames for reducing the temporal ambiguity. The code is available at: https://github.com/zczheng777/DPTM/.

Keywords: Temporal Decomposition · Predictive Distillation · Online Temporal Action Segmentation

1 Introduction

Temporal Action Segmentation (TAS) is a fundamental task for understanding long, untrimmed videos, with broad applications ranging from video surveillance [29] to skill assessment [20]. It aims to produce temporally coherent action segments by modeling both intra-action and inter-action temporal dependencies, formulated as assigning an action label to each frame.

⋆ Corresponding author.

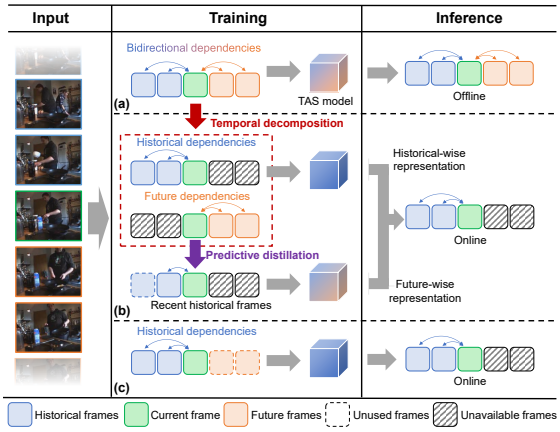


Fig. 1: Illustration of (a) Offline, (b) Our and (c) Online frameworks. The online framework imposes strict causal constraints during both training and inference, restricting the model to historical temporal dependencies. Our framework first captures inaccessible future dependencies during an offline-style training scheme via a temporal decomposition strategy, and then approximates them from recent historical observations using a predictive distillation strategy. This enables online TAS to leverage richer temporal dependencies while preserving causal constraints during inference.

Recently, significant progress [18, 21, 32] has been made in the *offline setting*, where models have access to the full temporal context, as illustrated in Figure 1.(a). However, such models are not applicable to real-time scenarios, as frames arrive sequentially and only historical frames are available, which is known as the *online setting*. To address this issue, prior studies [23, 36] focus on enhancing the modeling of historical dependencies under strict causal constraints, as illustrated in Figure 1.(c). Despite these improvements, they overlook the fundamental limitation of incomplete temporal dependencies, resulting in ambiguous interpretations of ongoing actions, which in turn produce fragmented predictions. One possible solution is to predict future features to assist current prediction [30]. However, temporal dependencies encode higher-level semantic patterns that vary across different actions, which cannot be fully captured by a limited number of future frames. Hence, we revisit these two settings and observe that offline-style training scheme naturally captures both historical and future dependencies, while causal constraints that limit access to future frames are enforced only during online inference. This motivates us to develop a new solution for online TAS, where future dependencies are first captured via offline-style training and then transferred to online model under causal constraints, aiming to mitigate the intrinsic ambiguity from limited temporal context.

Accordingly, we propose a two-stage Dual-Path Temporal Modeling (DPTM) framework tailored for online TAS, as shown in Figure 1.(b). In the first stage, we capture future dependencies in an offline-style training scheme, in which the entire sequence is available. Without causal masking, each frame can establish

interactions with both preceding and subsequent frames, enabling the model to learn complete bidirectional dependencies. However, such bidirectional modeling entangles historical and future information within a unified representation, making future-aware components indistinguishable. Hence, we explicitly disentangle future dependencies from historical ones via a *temporal decomposition strategy*. Specifically, we implement this strategy in a simple yet effective manner by applying complementary temporal masks to the two paths, where one branch is restricted to historical frames and the other to future frames. In the second stage, we transfer the captured future dependencies under causal constraints. Specifically, we design a *predictive distillation strategy* that uses recent historical frames as a proxy to approximate the future dependencies representation via knowledge distillation. This approach is motivated by the inherent temporal continuity of actions [5] and the structured progression of action sequences [23], where recent historical frames provide sufficient cues to approximate high-level future dependencies without explicitly predicting upcoming frames. Finally, during inference, only the model trained in the second stage is used, ensuring the framework suitable for online TAS.

In summary, our contributions are as follows:

1. We propose a new solution for online TAS that alleviates the intrinsic ambiguity caused by limited temporal context, enabling models to leverage richer temporal dependencies while preserving causal constraints during inference.
2. We propose a two-stage Dual-Path Temporal Modeling (DPTM) framework that employs a temporal decomposition strategy to disentangle historical and future dependencies, followed by a predictive distillation strategy to transfer the future dependencies.
3. Extensive experiments on multiple datasets demonstrate that DPTM consistently outperforms prior online TAS methods and produces more temporally coherent segmentations, validating the effectiveness of explicitly modeling and transferring future-aware temporal dependencies in online TAS.

2 Related Work

2.1 Temporal Action Segmentation

To capture temporal dependencies, early offline TAS models adopt sliding windows [11], but are limited by a constrained temporal receptive field. ED-TCN [14] addresses this by employing hierarchical dilated temporal convolutions. Building on this, MS-TCN [7] introduces a multi-round refinement strategy to progressively improve temporal predictions. These architectures have been shown [15,24,31] to effectively capture long-range dependencies and fine-grained temporal dynamics. To overcome the limited global context modeling of temporal convolutions [35], ASFormer [32] incorporates local self-attention within the multi-round refinement. The effectiveness of attention in capturing global temporal dependencies has made it a core component in subsequent methods [1,2,6,19,21],

suggesting that improvements in offline TAS largely stem from advances in temporal modeling. However, these methods cannot be directly applied to online settings, where only partial temporal information is available [23].

Recently, researchers have begun focusing on this challenging scenario. OnlineTAS [36] employs a temporal sliding module that predicts from recent frames and long-term information stored in a memory bank, achieving fast inference but remaining prone to over-segmentation due to the coarse-grained temporal context. ProTAS [23] introduces an action progress estimation module to capture temporal structure within actions, improving the temporal continuity of predictions. While effective in enhancing performance, both methods remain inherently limited by the temporal ambiguity arising from incomplete temporal dependencies. Since action dynamics are driven by higher-level temporal patterns rather than individual frame-level cues, these incomplete dependencies introduce uncertainty, limiting their ability to produce temporally coherent predictions.

2.2 Future-aware Online Action Understanding

Leveraging future information to enhance online prediction has also been explored in other action understanding tasks. For instance, OadTR [30] predicts future features to assist current clip-level prediction, while PK-GCN [26] approximates future skeleton poses to predict human motion. PKD-OAD [34] directly distills knowledge from an offline model to an online action detection model, and OODL [10] leverages an offline model as a prior to supervise the training of an online model, improving temporal coherence in action sequence prediction. However, frame-wise prediction captures only low-level visual features, which are not enough for modeling higher-level temporal dependencies. Additionally, direct distillation from offline representations mixes available historical information with inaccessible future dependencies. Hence, although these approaches incorporate future information for online prediction, they are hindered by the inability to capture future dependencies in online TAS.

3 Methodology

3.1 Preliminaries

Task Definition. Consider a video $V = \{x_t\}_{t=1}^{T}$ consisting of T frames, where $x_t \in \mathrm{R}^D$ is the frame-level feature at time step t with dimension D, pre-extracted using a pretrained I3D [3]. The corresponding frame-wise ground truth labels are denoted as $Y = \{y_t\}_{t=1}^{T}$, where $y_t \in \mathcal{C}$ represents the frame-wise action label, and $\mathcal{C}$ is the set of action classes with total size $N_C = |\mathcal{C}|$. In the online setting, the model sequentially and causally predicts the action label $\hat{y}_t$ as

$$\hat{y}_t = \arg\max_{y \in \mathcal{C}} p_t(y_t | x_{1:t}), \tag{1}$$

where p_t denotes the predicted probability distribution over action classes, without access to future frames $x_{t+1:T}$. In contrast, the offline TAS performs inference with full temporal context, utilizing the entire sequence $x_{1:T}$.

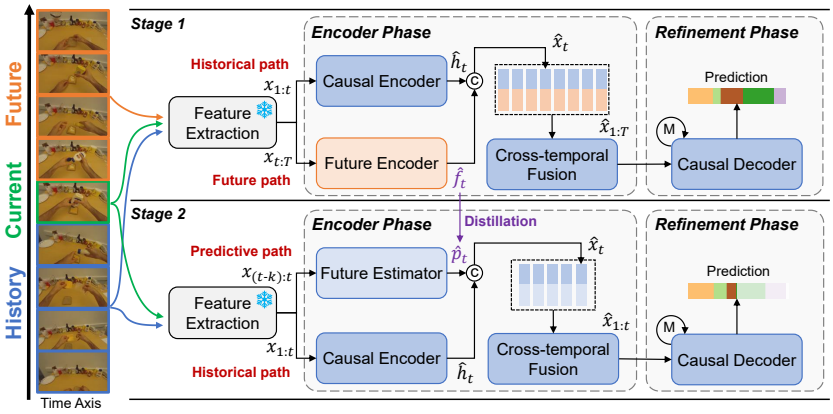


Fig. 2: Illustration of the DPTM within the online multi-round refinement framework. DPTM enhances the encoder phase while keeping the multi-round refinement phase unchanged. Specifically, two paths capture historical and future dependencies, respectively. Their outputs are concatenated and fused to obtain a comprehensive representation. Then, future dependencies are approximated from recent historical frames, enabling causal inference while benefiting from richer temporal context.

Causal Backbones. The representative architectures in TAS typically consist of temporal convolution [14] and attention mechanism [28]. To adapt them to the online setting, a simple modification is applied to enforce causal constraints [23, 27]. In the temporal convolution network, the l-th layer of a causal 1D temporal convolution can be formulated as follows:

$$L_c^{(l)}(k,s) = f\left(\sum_{i=0}^{k-1} W_i^{(l)} \cdot x_{t-s^l \cdot i}^{(l-1)} + b^{(l)}\right), \tag{2}$$

where k is convolution size and s denotes the dilation rate, while s^l corresponds to an exponentially growing temporal receptive field with respect to the layer index l [14]. $W_i^{(l)}$ is the i-th convolutional weight, $b^{(l)}$ is the corresponding bias term, and $f(\cdot)$ represents the activation function. The input $x_{t-s\cdot i}^{(l-1)}$ corresponds to the output from the $(l-1)$-th layer and time step $t - s \cdot i$.

For attention mechanism in l-th layer, a binary mask $M_s^{(l)} \in \{0,1\}^{s^l+1}$ can be applied to the attention scores $\alpha_{t,s}^{(l)}$ to prevent information leakage from future frames. Here, $[t-s^l/2, t+s^l/2]$ represents the attention window centered at frame t, with s denotes the dilation rate of window size [32]. In the binary mask, all future positions are set to 0, while all historical and current positions are set to 1. The masked attention scores are then obtained as

$$\tilde{\alpha}_{t,s}^{(l)} = \alpha_{t,s}^{(l)} \odot M_s^{(l)}, \tag{3}$$

where $\odot$ denotes element-wise multiplication operation.

3.2 Dual-Path Temporal Modeling Framework

Overview. Following prior works [23, 36], we adopt the typical online multi-round refinement framework to introduce our DPTM, as illustrated in Figure 2. This framework consists of two phases, where the first uses a causal encoder for temporal aggregation and the second employs M causal decoders for progressively refining predictions.

In the first stage of DPTM, for the t-th frame, the historical features $x_{1:t} \in \mathbb{R}^{t\times D}$ and future features $x_{t:T} \in \mathbb{R}^{(T-t)\times D}$ are fed into historical and future paths, respectively, to explicitly capture complementary dependencies. The outputs $\hat{h}_t \in \mathbb{R}^m$ and $\hat{f}_t \in \mathbb{R}^m$ are concatenated to form the augmented feature $\hat{x}_t \in \mathbb{R}^{2m}$, where m denotes the embedding dimension. A cross-temporal fusion module is then applied over the sequence $\hat{x}_{1:T} \in \mathbb{R}^{T\times 2m}$ to smooth and integrate the enhanced features, producing coherent representations for subsequent perdiction. In the second stage, the future path is substituted with a predictive path, where a future estimator takes recent historical features $x_{(t-k):t} \in \mathbb{R}^{k\times D}$ as input and outputs $\hat{p}_t \in \mathbb{R}^m$, which is distilled to approximate the future dependency representation $\hat{f}_t$. Here, k denotes the number of used historical frames, determined by the temporal receptive field of the corresponding network.

During online inference, only the second stage is used, allowing causal predictions that leverage the richer temporal context acquired during training.

Temporal Decomposition Strategy. In TAS, complete temporal dependencies are essential for modeling action structure but cannot be directly used in online inference. Therefore, we propose a temporal decomposition strategy that splits bidirectional modeling into two unidirectional paths, with a causal path capturing historical dependencies and a future path capturing future dependencies. This design is inspired by prior works [9, 12, 33] on temporal feature aggregation, where different fusion strategies are explored for combining complementary temporal cues. Similarly, we first model unidirectional dependencies and then fuse them in a structured manner. Empirically, this decomposition achieves performance comparable to full bidirectional modeling.

For the causal encoder in the historical path, we adopt a standard causal backbone as described in 3.1, such as stacked temporal convolution layers [7] or stacked attention layers [32]. The future encoder shares the same architecture, differing only in its orientation to capture future temporal dependencies. We maintain architectural symmetry as both types of unidirectional dependencies are equally important. When using stacked temporal convolution layers as the backbone, the l-th layer in the future encoder is formulated as

$$L_f^{(l)}(k,s) = f\left(\sum_{i=0}^{k-1} W_i^{(l)} \cdot x_{t+s^l\cdot i}^{(l-1)} + b^{(l)}\right). \tag{4}$$

The difference from (2) lies in the indices of x, which determine the direction of temporal aggregation.

Given the outputs of the historical and future paths, $\hat{h}_t$ and $\hat{f}_t$, we concatenate them to obtain $\hat{x}_t$, which is subsequently processed by the cross-temporal

fusion module. Since future dependencies are estimated during inference, these per-frame predictions can be noisy or temporally inconsistent. The fusion module mitigates such fluctuations by integrating the concatenated features, producing a more stable and temporally coherent representation. It is implemented as a stacked causal temporal convolutions, as described in (2), which suffices for temporal aggregation while maintaining efficiency for online inference.

Predictive Distillation Strategy. In the first stage, the future encoder captures future dependencies that benefit the causal components but cannot be directly used in online inference, as it relies on future frames. In the second stage, we retain the same model architecture but replace the future path with a predictive path, where a future estimator uses recent historical features to generate a representation of future dependencies. Using only recent history strikes a balance between capturing sufficient temporal cues and maintaining causal inference and efficiency. Specifically, we adopt knowledge distillation to enforce that the estimator produces representations close to $\hat{f}_t$, enabling the online model to approximate future dependencies learned in training, which improves temporal coherence and reduces ambiguity in online frame-level predictions.

The future estimator is implemented as stacked causal temporal convolutions, leveraging the compression effect of knowledge distillation to reduce the computational overhead introduced by the dual-path design. The l-th is formulated as

$$L_p^{(l)}(k,s) = f\left(\sum_{i=0}^{k-1} W_i^{(l)} \cdot x_{t+s\cdot i}^{(l-1)} + b^{(l)}\right). \tag{5}$$

Compared with (2), the dilation in the future estimator grows linearly with the layer index. For instance, a 10-layer future setimator with $s = 2$ covers roughly $k = 100$ historical frames, providing temporal context to approximate future dependencies for ongoing actions while preserving causal inference.

3.3 Loss Functions

Stage 1. We apply the cross-entropy loss to the outputs of all key components, including causal encoder (c), future encoder (f), cross-temporal fusion module (t) and causal decoders (d), ensuring that each component captures essential dependencies to generate accurate frame-wise predictions. The losses are

$$L_1^{cls} = \sum_{i\in\{c,f,t,d\}} CE(P_i, Y), \tag{6}$$

where P_i denotes the frame-wise prediction probability of module i, and Y is the ground-truth label sequence. In addition, we incorporate a smoothing loss [7] to encourage consistency between consecutive frame predictions, mitigating abrupt fluctuations and improving segmentation coherence:

$$L_1^{smo} = \sum_{i\in\{c,f,t\}} \sum_{t=1}^{T-1} \|P_i[t] - P_i[t-1]\|_2^2, \tag{7}$$

where $P[t]$ denotes the predicted probability vector at t frame. The overall loss for stage 1 is

$$L_1 = L_1^{cls} + \lambda_1 L_1^{smo}, \tag{8}$$

with λ_1 controlling the weight of the smoothing term.

Stage 2. similar to the stage 1, cross-entropy and temporal smoothing losses are applied to all key components. For the future estimator, the well-trained future encoder is frozen, and its learned future dependencies are transferred via a predictive distillation loss:

$$L_2^d = \sum_{i=1}^{T-1} \left\| \hat{f}_t - \hat{p}_t \right\|_1, \tag{9}$$

where $\|\cdot\|_1$ denotes the L_1 norm. The overall loss is represented as

$$L_2 = L_2^{cls} + \lambda_1 L_2^{smo} + \lambda_2 L_2^{pd}, \tag{10}$$

where λ_2 controlling the weight of the predictive distillation term. L_2^{cls} and L_2^{smo} are computed in the same manner as L_1^{cls} and L_1^{smo}, respectively.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate our model on three widely used TAS benchmarks: GTEA [8], Breakfast [13], and 50Salads [25]. GTEA contains 28 egocentric videos recorded in a kitchen setting, covering 11 action categories. 50Salads consists of 50 recordings of individuals preparing salads. Breakfast is a largescale dataset comprising 1,712 third-person view videos of breakfast preparation. For all datasets, we adopt the standard training/testing splits commonly used in previous work [7]. We use standard evaluation metrics including framewise accuracy (Acc), segmental edit distance (Edit) and segmental F1 scores at overlapping thresholds of 10%, 25%, 50% (F1@{10, 25, 50}) and the average of these metrics (Avg). Higher values indicate better performance.

Implementation Details. For a fair comparison with existing works, we integrate DPTM into two representative models, MS-TCN [7] and ASFormer [32], both following the multi-round refinement framework with a single encoder and multiple decoders. The original components are kept at default configurations except for converting them to causal. The future encoder adopts the same architecture and settings as the original encoder, while the cross-temporal fusion module and the future estimator are implemented as 10-layer causal TCNs. All modules are configured with a dilation rate of $s = 2$ and the feature embedding dimension is set to 64. Training is performed using the Adam optimizer with a fixed learning rate of 0.0005 and a batch size of 1. The loss weights are set to $\lambda_1 = 0.15$ and $\lambda_2 = 0.5$, and both training stages run for 50 epochs. All experiments are conducted on a single NVIDIA A40 GPU.

Table 1: Comparison of online TAS methods on three benchmark datasets. The best and second-best scores are highlighted in red and blue, respectively. Offline baseline (Line 1, 6) are shown for reference only. Results of OnlineTAS are reported from the original paper, while ProTAS is reproduced using the official code.

L.	Methods	M.	GTEA				50Salads				Breakfast			
			acc	Edit	F1@{10, 25, 50}	Avg	acc	Edit	F1@{10, 25, 50}	Avg	acc	Edit	F1@{10, 25, 50}	Avg
1	MS-TCN [7]	Off.	76.3	79.0	85.8/83.4/69.8	78.9	80.7	67.9	76.3/ 74.0/64.5	72.7	66.3	61.7	52.6/48.1/37.9	53.3
2	proTAS [23]	On.	74.3	69.2	77.1/74.1/59.7	70.9	68.8	29.9	34.5/32.6/25.9	38.3	52.8	26.4	21.2/18.3/11.4	26.0
3	onlineTAS [36]	On.	76.2	63.5	72.6/68.3/58.8	67.9	80.9	28.8	36.1/31.0/23.3	40.0	56.7	19.3	16.8/13.9/9.3	23.2
4	MS-TCN	On.	73.0	70.7	71.1/67.9/54.4	67.4	73.9	24.5	33.2/30.4/24.0	37.2	50.9	20.5	18.1/14.9/10.0	22.9
5	**Ours**	On.	75.5	75.1	77.2/73.6/61.9	72.7	72.1	40.9	48.5/45.3/34.6	48.3	53.6	31.0	28.1/23.8/15.9	30.5
6	ASFormer [32]	Off.	79.7	84.6	90.1/88.8/79.2	84.5	85.6	79.6	85.1/83.4/76.0	81.9	73.5	75.0	76.0/70.6/57.4	70.5
7	proTAS [23]	On.	77.3	74.0	79.9/77.3/65.4	74.8	78.8	43.5	52.7/50.2/43.4	53.7	67.0	60.0	56.0/50.5/37.6	54.2
8	onlineTAS [36]	On.	75.0	69.7	77.7/74.0/62.0	71.7	77.5	29.1	37.9/35.3/28.6	41.7	64.9	32.1	31.2/27.4/20.2	35.2
9	ASFormer	On.	75.9	73.2	79.3/77.7/65.0	74.2	78.4	42.9	52.4/49.1/41.8	52.9	66.9	59.1	55.8/49.9/36.5	53.6
10	**Ours**	On.	76.1	80.0	85.3/81.1/70.9	78.7	79.0	45.8	57.7/53.5/46.8	56.6	66.9	63.0	60.4/54.2/39.9	56.9

Table 2: Ablation study on the proposed Temporal Decomposition (T.D.) and Predictive Distillation (P.D.) strategies.

Line	T.D.	P.D.	50salads			
			acc	Edit	F1@{10, 25, 50}	Avg
1	✓	✗	71.1	28.6	37.3/31.8/26.2	39.0
2	✗	✓	70.5	35.7	43.6/38.1/29.5	43.5
3	✓	✓	**72.1**	**40.9**	**48.5/45.3/34.6**	**48.3**

4.2 Comparison with State-of-the-Art

In Table 1, we compare our model with existing online TAS methods. Lines 1-6 use MS-TCN as the backbone, while Lines 7-10 use ASFormer. Compared to offline baselines (Lines 1 and 7), the corresponding online baselines (Lines 4 and 9) exhibit a notable performance drop due to the limited temporal context. ProTAS relies heavily on progress cues from human-object interactions, which are challenging to observe in third-person datasets. OnlineTAS focuses on short-term dynamics, achieving higher frame-level accuracy but substantially lower Edit and F1 scores due to severe over-segmentation. Compared to the online backbones, our model consistently achieves significant improvements, particularly in Edit and F1 metrics, indicating more temporally coherent segmentation. Furthermore, comparing Lines 5 and 10 shows that our framework can further benefit from advances in backbone architecture. These results validate the effectiveness of our framework in enhancing online TAS performance.

4.3 Ablation Study

Effect on Proposed Strategies. Table 2 evaluates the contributions of Temporal Decomposition (T.D.) and Predictive Distillation (P.D.) strategies. Line 1

Table 3: Ablation study on the effect of the Temporal Decomposition Strategy (T.D.). † indicates that the MS-TCN encoder is scaled to match the parameter count.

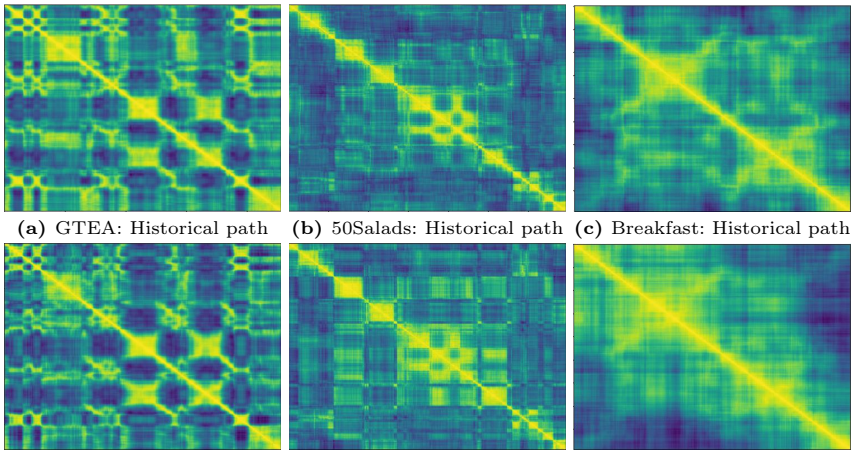
Settings	Para.	GTEA				50Salads				Breakfast			
		Acc	Edit	F1@{10, 25, 50}	Avg	Acc	Edit	F1@{10, 25, 50}	Avg	Acc	Edit	F1@{10, 25, 50}	Avg
MS-TCN	0.8M	76.3	79.0	85.8/83.4/69.8	78.9	80.7	67.9	76.3/74.0/64.5	72.7	66.3	61.7	52.6/48.1/37.9	53.3
MS-TCN †	1.2M	**77.1**	81.6	86.3/**84.7**/72.1	80.4	**81.3**	**68.8**	**77.0**/**75.2**/65.8	**73.6**	**66.4**	61.7	**53.4**/**48.5**/**38.0**	**53.6**
MS-TCN+DS	1.1M	76.5	**83.1**	**86.8**/84.5/**74.0**	**81.0**	81.0	68.7	76.3/74.5/**67.2**	73.5	66.1	**61.9**	53.2/48.0/37.9	53.4

applies only T.D., yielding a purely causal single-path model. Line 2 follows conventional offline-to-online distillation without decomposition, while Line 3 combines both strategies as our full framework. Comparing Lines 1 and 3, consistent gains demonstrate the effectiveness of future-aware supervision, as the predictive path provides temporal cues beyond causal modeling. Comparing Lines 2 and 3 shows that naive distillation offers limited improvement, indicating that directly aligning offline and online representations with mismatched temporal receptive fields is suboptimal. To evaluate the effectiveness of our decomposition strategy in approximating bidirectional temporal modeling, we compare it with offline MS-TCN in Table 3. The results show that replacing full bidirectional modeling with the dual-path structure and temporal fusion module does not noticeably degrade performance. Even when MS-TCN is scaled up to match parameter count, our strategy maintains or improves segmentation quality, demonstrating that it effectively captures complementary dependencies typically obtained through bidirectional modeling.

Table 4: Ablation study on various future-aware approaches. 'Direct distillation' refers to directly transferring representations learned by an offline bidirectional model to the online model, while 'Frame-wise prediction' denotes predicting a fixed number of future frame representations to assist prediction of the current frame.

Line	Settings	Breakfast			
		Acc	Edit	F1@{10, 25, 50}	Avg
1	MS-TCN	50.9	20.5	18.1/14.9/10.0	22.9
2	Direct distillation	53.3	26.6	21.4/17.5/13.6	26.5
3	Frame-wise prediction	**54.8**	23.7	23.7/19.5/12.8	26.9
4	Ours	53.6	**31.0**	**28.1**/**23.8**/**15.9**	**30.5**

Effect of Future-aware Approach. To evaluate the effectiveness of our proposed future-aware approach in online TAS, we compare it with two alternative approaches in Table 4, using the baseline (Line 1) for reference. Line 2 performs direct offline-to-online distillation, where online model align representation with a well-trained offline model. Line 3 follows OadTR [30], where a cross-attention module predicts 16 future frames based on 16 historical frames. Line 4 corresponds to our approach, which disentangles future dependencies and transfers



(a) GTEA: Historical path **(b)** 50Salads: Historical path **(c)** Breakfast: Historical path

(d) GTEA: Predictive path **(e)** 50Salads: Predictive path **(f)** Breakfast: Predictive path

Fig. 3: Comparison of frame-level self-similarity in historical and predictive paths on example videos from three datasets. Brighter colors indicate higher similarity.

them through predictive distillation. Direct distillation yields modest improvements due to the transfer of redundant historical information, while OadTR-style prediction has limited impact on Edit, as future dependencies vary across actions and cannot be fully captured by fixed-frame predictions. In contrast, our method explicitly disentangles and selectively transfers future dependencies, providing targeted future-aware signals that lead to more coherent predictions.

To further analyze the role of captured future dependencies, we visualize the frame-wise similarity matrices of the outputs from the historical and predictive paths before fusion, as shown in Figure 3. Both paths exhibit highly consistent temporal structures, with clear block patterns corresponding to action segments, indicating that future dependencies effectively capture the underlying temporal structure rather than introducing unrelated information. Subtle local differences suggest that future-aware cues provide fine-grained complementary information. As supported by our experimental results, these cues help refine ambiguous frames, contributing to improved segmentation coherence. Overall, our future-aware approach captures dependencies that are structurally aligned with historical features while supplying additional detail for predictions.

Effect of Temporal Context on Future Estimator. We evaluate the performance of different dilation schedules for the historical temporal receptive fields of the future estimator in Table 5. Results show that the linear dilation consistently achieves the best overall performance. In contrast, a fixed dilation rate performs the worst, and the exponential schedule underperforms despite rapidly expanding the receptive field, as it tends to over-aggregate distant context and dilute local temporal cues. This finding suggests that using a linearly expanding dilation ensures that the future estimator focuses on recent historical frames when predicting future dependencies. When combined with our temporal de-

Table 5: Ablation study on dilation schedules for future estimator, where dilation rate grows with depth following fixed, linear or exponential schemes. i denotes layer index.

Dilation	GTEA			
	Acc	Edit	F1@{10, 25, 50}	Avg
2 (Fixed)	72.0	63.4	65.6/62.8/51.2	63.0
$2 \times i + 1$ (Linear)	**75.5**	**75.1**	**77.2/73.6/61.9**	**72.7**
2^i (Exp)	73.2	70.4	71.0/67.6/54.5	67.3

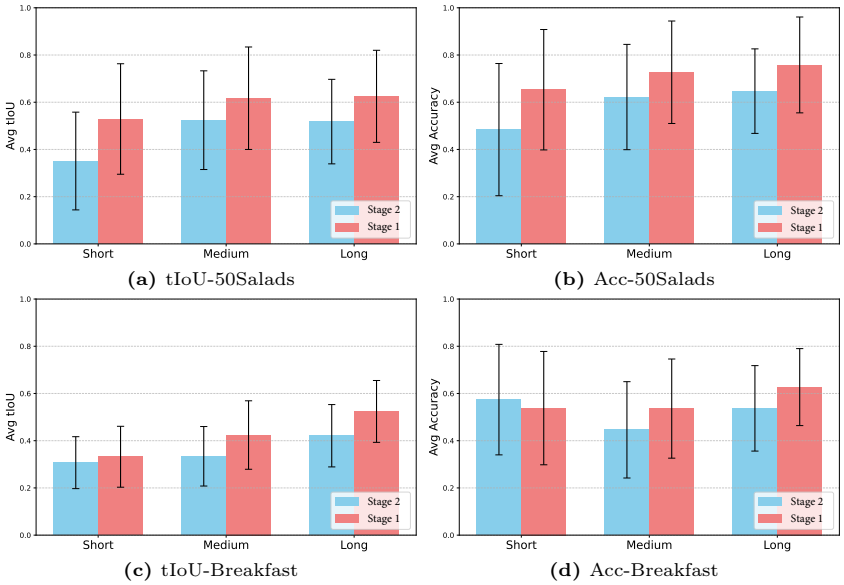
Table 6: Ablation study on the impact of embedding Dimensionality (Dim.) on performance and inference speed (FPS). FPS denotes the number of frames processed per second during inference.

Methods	Dim.	FPS	50Salads			
			Acc	Edit	F1@{10, 25, 50}	Avg
MS-TCN	64	18.96	73.9	24.5	33.2/30.4/24.0	37.2
Ours	32	16.67	68.6	32.4	42.8/39.9/28.0	42.3
	64	11.11	72.1	40.9	48.5/45.3/34.6	48.3
	128	6.25	72.5	**42.7**	**50.7/48.6/36.5**	**50.2**
	256	3.70	**73.0**	41.9	48.8/46.1/32.5	48.5

composition and predictive distillation framework, this design allows the model to selectively capture and transfer temporally relevant cues, effectively reducing ambiguity and improving the coherence of online predictions.

Efficiency. We investigate the effect of embedding dimensionality on both performance and inference speed, as shown in Table 6. When our framework uses the same feature dimension as the baseline, the inference speed decreases, but substantial improvements in segmentation quality are observed. As the feature dimensionality increases, performance consistently improves, with the best results achieved at a dimension of 128. Notably, for a fair comparison with existing methods, we still adopt a feature dimension of 64 in all experiments. These findings suggest that our framework benefits from increased representational capacity. Although the reduced inference speed may limit applicability in real-time scenarios, this issue can potentially be mitigated by further architectural optimizations. We will further explore the balance between performance and efficiency in future work.

Transfer Performance Across Actions. To evaluate transfer performance, we compare the stage 1 (teacher) and stage 2 (student) models across actions of different durations in Figure 4. For 50Salads, actions are categorized as short ($<$3%), medium (3-10%), and long ($>$10%) duration, where the percentage refers to the proportion of the video's total duration that each action occupies. For Breakfast, the thresholds are 5% and 20%. Temporal Intersection over Union (tIoU) and frame-wise accuracy are computed for each action category, where



(a) tIoU-50Salads **(b)** Acc-50Salads

(c) tIoU-Breakfast **(d)** Acc-Breakfast

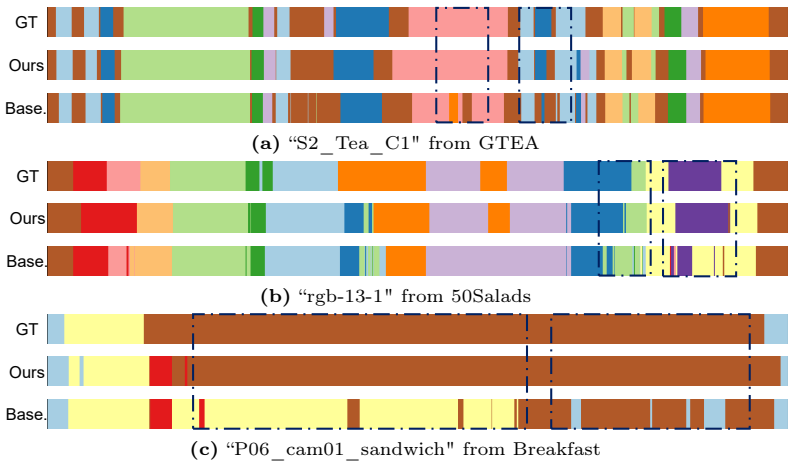
Fig. 4: Bar charts comparing the average tIoU and accuracy (Acc) of stage 1 and stage 2 models for short-, medium-, and long-duration actions on the 50Salads and Breakfast datasets, with error bars representing standard deviation.

tIoU measures the overlap between the predicted segments and the ground-truth annotations. The student model improves medium- and long-duration actions in 50Salads while maintaining competitive short-duration performance. In Breakfast, the future estimator emphasizes near-future frames, yielding strong results on short-duration tasks. These results indicate that implicitly capturing future dependencies better suits online TAS, and that the estimator adaptively learns future context to enhance segmentation across varying action durations.

Generalization on skeleton data. To further assess the generalization of our framework, we evaluate its performance on skeleton-based online TAS. Experiments are conducted on two widely used benchmarks, LARA [22] and PKU-MMD [17], following the standard evaluation protocols [4, 16]. Specifically, we insert a 1×1 convolutional layer into MS-TCN to adapt the dimensionality. As reported in Table 7, our framework consistently improves performance across all metrics on both datasets. Unlike RGB inputs, skeleton-based data lack appearance cues such as object interactions or environmental context, providing only the motion dynamics of the human skeleton. Despite this limitation, our approach effectively captures temporal dependencies and transfers future-aware information, demonstrating its robustness and applicability.

Table 7: Performance comparison on two skeleton-based TAS datasets.

Datasets	Methods	Acc	Edit	F1@{10, 25, 50}	Avg
LARA	MS-TCN	63.4	37.6	40.6/35.3/23.4	40.1
	ProTAS [23]	66.3	38.4	43.3/38.5/27.0	42.7
	Ours	**66.8**	**47.1**	**52.2/45.3/31.4**	**48.6**
PKU-sub	MS-TCN	54.7	42.4	45.1/38.7/26.8	41.5
	ProTAS [23]	55.0	43.6	45.8/39.0/27.6	42.2
	Ours	**55.8**	**47.2**	**51.1/44.4/30.8**	**45.9**



(a) "S2_Tea_C1" from GTEA

(b) "rgb-13-1" from 50Salads

(c) "P06_cam01_sandwich" from Breakfast

Fig. 5: Qualitative results on three datasets. "Base." denotes predictions from the baseline. Dashed boxes highlight regions with improved accuracy and temporal coherence.

4.4 Qualitative Results

We visualize predictions on selected video samples in Figure 5. Compared to the baseline, our method significantly alleviates the over-segmentation problem. Fragmented or misaligned segments are corrected by incorporating future dependencies, which provide richer temporal context for ongoing actions. Additionally, our method consistently improves performance across different action durations by implicitly capturing future temporal dependencies, enhancing current predictions and enabling adaptive use of future information for diverse action understanding in online TAS. These results highlight the effectiveness of our framework, offering a solution that goes beyond relying solely on historical dependencies by explicitly capturing and leveraging future dependencies.

5 Conclusion

In this paper, we propose a two-stage Dual-Path Temporal Modeling (DPTM) framework to enhance online temporal action segmentation. By leveraging an

offline-style training scheme, our approach explicitly disentangles historical and future dependencies through the temporal decomposition strategy. The future dependencies are then transferred to the online model using a predictive distillation strategy, preserving causal constraints. Unlike prior online TAS methods that primarily focus on modeling historical dependencies, DPTM improves temporal coherence by reducing the ambiguity arising from incomplete temporal context. Extensive experiments validate the effectiveness of our framework, showing substantial gains in segmentation coherence.

Acknowledgements

This work was supported by Natural Science Foundation of Jiangsu Province (Grant BK20240583), National Natural Science Foundation of China (Grant Nos. 62407021, 62306145, 62277031, 62377029), Frontier Technologies R&D Program of Jiangsu (Grant BF2024076), Natural Science Foundation of Jiangsu Higher Education Institutions (Grant Nos. 25KJD520006, 25KJB520024), and Open Fund of the Industrial Software Engineering Technology Research and Development Center of Jiangsu Province (Grant ZK25-04-04).

References

1. Aziere, N., Todorovic, S.: Multistage temporal convolution transformer for action segmentation. Image and Vision Computing **128**, 104567 (2022)
2. Bahrami, E., Francesca, G., Gall, J.: How much temporal long-term context is needed for action segmentation? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10351–10361 (2023)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Chen, B., Nie, W., Ji, H., Ren, W., Tong, Q., Wang, Z., Liu, H.: Multiscale skeleton-based temporal action segmentation using hierarchical temporal modeling and prediction ensemble. IEEE Transactions on Cybernetics (2025)
5. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 1011–1030 (2023)
6. Du, D., Su, B., Li, Y., Qi, Z., Si, L., Shan, Y.: Do we really need temporal convolutions in action segmentation? In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 1014–1019. IEEE (2023)
7. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3575–3584 (2019)
8. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: CVPR 2011. pp. 3281–3288. IEEE (2011)
9. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)

10. Ghoddoosian, R., Dwivedi, I., Agarwal, N., Choi, C., Dariush, B.: Weakly-supervised online action segmentation in multi-view instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13780–13790 (2022)
11. Karaman, S., Seidenari, L., Del Bimbo, A.: Fast saliency based pooling of fisher encoded dense trajectories. In: ECCV Thumos Workshop. vol. 1, p. 5 (2014)
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
13. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 780–787 (2014)
14. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
15. Li, S., Farha, Y., Liu, Y., Cheng, M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 6647–6658 (2023)
16. Li, Y., Li, Z., Gao, S., Wang, Q., Hou, Q., Cheng, M.M.: A decoupled spatio-temporal framework for skeleton-based action segmentation. arXiv preprint arXiv:2312.05830 (2023)
17. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: A large scale benchmark for skeleton-based human action understanding. In: Proceedings of the workshop on visual analysis in smart and connected communities. pp. 1–8 (2017)
18. Liu, D., Li, Q., Dinh, A.D., Jiang, T., Shah, M., Xu, C.: Diffusion action segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10139–10149 (2023)
19. Liu, D., Li, Q., Dinh, A.D., Jiang, T., Shah, M., Xu, C.: Diffact++: Diffusion action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
20. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9522–9531 (2021)
21. Lu, Z., Elhamifar, E.: Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18175–18185 (2024)
22. Niemann, F., Reining, C., Moya Rueda, F., Nair, N.R., Steffens, J.A., Fink, G.A., Ten Hompel, M.: Lara: Creating a dataset for human activity recognition in logistics using semantic attributes. Sensors **20**(15), 4083 (2020)
23. Shen, Y., Elhamifar, E.: Progress-aware online action segmentation for egocentric procedural task videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18186–18197 (2024)
24. Singhania, D., Rahaman, R., Yao, A.: C2f-tcn: A framework for semi-and fully-supervised temporal action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 11484–11501 (2023)
25. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. pp. 729–738 (2013)

26. Sun, X., Cui, Q., Sun, H., Li, B., Li, W., Lu, J.: Overlooked poses actually make sense: Distilling privileged knowledge for human motion prediction. In: European Conference on Computer Vision. pp. 678–694. Springer (2022)
27. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 **12**, 1 (2016)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer **29**(10), 983–1009 (2013)
30. Wang, X., Zhang, S., Qing, Z., Shao, Y., Zuo, Z., Gao, C., Sang, N.: Oadtr: Online action detection with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7565–7575 (2021)
31. Wang, Z., Gao, Z., Wang, L., Li, Z., Wu, G.: Boundary-aware cascade networks for temporal action segmentation. In: European Conference on Computer Vision. pp. 34–51. Springer (2020)
32. Yi, F., Wen, H., Jiang, T.: Asformer: Transformer for action segmentation. BMVC (2021)
33. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015)
34. Zhao, P., Xie, L., Wang, J., Zhang, Y., Tian, Q.: Progressive privileged knowledge distillation for online action detection. Pattern Recognition **129**, 108741 (2022)
35. Zheng, Z., Zhou, Y., Chen, Y., Zhou, J., Gu, Y.: What, when and where: Spatial-aware temporal action segmentation. Pattern Recognition p. 111833 (2025)
36. Zhong, Q., Ding, G., Yao, A.: Onlinetas: An online baseline for temporal action segmentation. Advances in Neural Information Processing Systems **37**, 58984–59005 (2024)