DART: Deformable Adaptive Reasoning with Temporal Queries for Online Skeleton-Based Action Recognition

Chaoyang Zheng[1], Jinfang Gan[1], Yang Xiao[1†], Tingbing Yan[1], Xintao Zhang[1], Ran Wang[2], Zhiguo Cao[3], and Joey Tianyi Zhou[4,5]

[1] The National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China.
[2] School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan 430074, China
[3] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[4] Singapore Management University, Singapore
[5] Institute of Advanced Intelligence and Computing (IAIC), Agency for Science, Technology and Research (A*STAR), Singapore
{zchaoyang,jinfangan,Yang_Xiao,yantingbing,xintaozhang,rex_wang,zgcao}@hust.edu.cn,zhouty@cfar.a-star.edu.sg

Abstract. Online skeleton-based action recognition requires per-frame classification using only past and current observations. We identify two key challenges in this setting: (1) *Not all frames are equal:* discriminative motion concentrates within specific temporal regions whose location varies across action categories, yet existing temporal modeling relies on fixed sampling patterns and causal attention that treat all positions equally. (2) *Not all futures need to be predicted:* prior methods attempt to explicitly predict future features to compensate for the unobserved portion, but the inherent ambiguity of human motion causes such prediction to degenerate into trivial identity mapping of the input. To address these challenges, we propose DART (Deformable Adaptive Reasoning with Temporal Queries). DART employs a Deformable Multi-scale Temporal Network (DMTN) that introduces input-dependent offsets into causal convolutions, enabling content-adaptive sampling of the most informative historical moments. The learned sampling positions naturally form a compact temporal memory, upon which a Temporal Query Decoder (TQD) operates: learnable queries interact with the memory to distill high-level action evolution patterns, providing complementary temporal reasoning without explicit future forecasting. Extensive experiments on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA demonstrate the effectiveness of DART. On NTU RGB+D 60, DART achieves 73.99% AUC, surpassing the SOTA by +3.04%. Code is available.

Keywords: Online skeleton-based action recognition · Deformable convolution · Temporal query decoding

† Corresponding author: Yang Xiao (Yang_Xiao@hust.edu.cn).

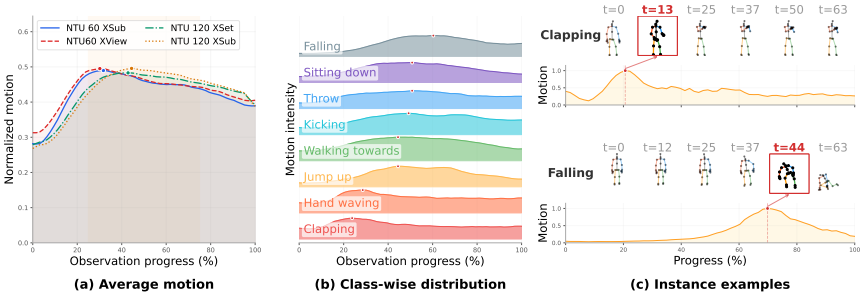

Fig. 1: Temporal distribution of discriminative motion across action sequences. (a) Average normalized motion intensity over observation progress on different datasets. (b) Motion curves for several action classes, where red dots indicate peaks. (c) Two concrete examples with skeleton visualizations and corresponding motion curves.

1 Introduction

Human action recognition is a fundamental task in computer vision, aiming to classify human activities from input sequences [3,23,30,34]. Among various input modalities, skeleton data has gained increasing attention because it represents the human body as compact joint coordinates, making it highly robust to background clutter, viewpoint variations, and appearance changes [6,7,16,23,25,34]. Recent skeleton-based methods [6, 7, 10, 17, 33–35] have achieved remarkable progress by effectively modeling the spatial topology of human joints and the temporal dynamics across frames.

However, most existing skeleton-based models operate in an offline setting, which assumes the entire sequence is available before making a prediction. In many real-world scenarios such as human-robot interaction [21, 26], the system must recognize actions as they unfold. A related task, early action prediction [4, 12, 14, 16, 31], attempts to address this by classifying actions from partial observations at predefined ratios of the full sequence. However, it still requires prior knowledge of the total sequence length and treats predictions at different stages as isolated tasks. In contrast, InfoGCN++ [8] introduced online action recognition for skeleton data, which requires the model to produce a classification at every frame in a streaming fashion, using only past and current observations without access to future frames or the full action duration. This strictly causal constraint poses a significant challenge: the model must extract discriminative features from limited and incrementally growing historical context.

Intuitively, **not all frames are equal:** recognizing an ongoing action does not require equal attention to every observed frame—only a few key moments carry the most discriminative information. As illustrated in Fig. 1(a), discriminative motion in action sequences concentrates within specific temporal regions rather than distributing uniformly across time. Moreover, the location of these informative peaks varies significantly across action categories (Fig. 1(b))—for instance, *Clapping* exhibits its peak as early as 20% of the sequence, while *Falling*

peaks near 70% (Fig. 1(c)). These observations suggest that effective online recognition requires adaptively locating the most discriminative moments from the growing historical context, yet conventional temporal convolutions sample at predetermined positions regardless of action content or observation progress. Meanwhile, causal attention, which restricts each frame to attend only to past observations, tends to dilute its weights uniformly as the sequence grows, further limiting the model's ability to focus on critical moments (see Sec. 3.1).

Furthermore, **not all futures need to be predicted:** beyond knowing where to look, the model must also make sense of what it has observed. Prior online methods attempt to bridge this gap by explicitly predicting future frame features and joint coordinates to compensate for the unobserved portion [8]. However, this is inherently challenging: partial motion sequences are ambiguous, as similar early movements may correspond to very different actions. And such explicit forecasting often degenerates into a trivial identity mapping of the input (see Sec. 3.1). Drawing on prior knowledge of how actions typically unfold, we can often infer the likely motion pattern from just a few key cues—without needing to explicitly predict every future change in the sequence. Online recognition demands a similar capacity: at any given moment, the model sees only a partial and growing sequence, yet must reason about the underlying motion trend that extends beyond the current observation.

To address these challenges, we propose DART (Deformable Adaptive Reasoning with Temporal Queries) for online skeleton-based action recognition. For adaptive temporal modeling, DART employs a Deformable Multi-scale Temporal Network (DMTN) that learns content-dependent sampling positions, enabling the model to focus on the most discriminative moments as they shift across different actions and observation stages. These adaptively selected positions naturally compose a compact temporal memory. Rather than forecasting future features from this memory, we introduce a Temporal Query Decoder (TQD), where learnable queries interact with the temporal memory to directly distill high-level motion trends that capture how actions typically progress—providing complementary temporal reasoning without explicit future prediction.

The main contributions of this work are summarized as follows:

- We analyze the existing online recognition method and identify two limitations: attention dilution and degeneracy of explicit prediction (see Sec. 3.1).
- We propose DART, comprising a DMTN for content-adaptive temporal sampling and a TQD that distills motion trends from the resulting temporal memory via learnable queries without explicit future forecasting.
- Extensive experiments on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA demonstrate SOTA performance.

2 Related work

2.1 Skeleton-Based Action Recognition

Skeleton-based action recognition has progressed significantly. GCN-based methods [6, 7, 17, 25, 34, 38] model skeletons as spatial-temporal graphs and focus

on learning adaptive topologies for spatial modeling, such as ProtoGCN [17], which dynamically learns channel-wise and adaptive graph topologies to capture fine-grained relational patterns. More recently, Transformer-based approaches [10, 29, 33, 37] leverage self-attention to capture long-range dependencies across joints and frames, such as SkateFormer [10], which designs partition-specific attention to model skeletal-temporal relationships across different joint groups. However, these methods all operate in the offline setting with fixed temporal sampling patterns. In contrast, our work targets the online setting and introduces content-adaptive temporal modeling that dynamically focuses on the most discriminative moments for more effective per-frame classification.

2.2 Online Action Recognition and Early Prediction

Early action prediction aims to recognize actions from partially observed sequences at predefined observation ratios. Existing methods typically achieve this by learning discriminative representations from incomplete sequences [11,12,31], by generating features for the unobserved portion to complement partial observations [4], or by distilling knowledge from full sequences to guide partial-sequence recognition [14–16]. However, it assumes known sequence length and makes a single prediction at a fixed ratio, unlike the streaming per-frame classification required in real-world applications [1, 22, 26]. To achieve true streaming processing, Online Action Detection (OAD) has been extensively explored in the video domain. Recent OAD methods typically employ Transformer-based architectures [27,32] and memory-refined frameworks [13,20,27], utilizing anticipation queries or generative decoders to forecast future context. However, these methods are designed for dense RGB video features and rely on rich appearance cues that are not available in skeleton data. InfoGCN++ [8] first formalized the online skeleton-based recognition setting, employing causal masked attention for temporal modeling and a Neural ODE [5] module to predict future features. Our analysis reveals two limitations of this paradigm: attention dilution over growing sequences and degeneracy of explicit future prediction into trivial identity mapping, motivating our alternative approach based on adaptive temporal sampling and query-based reasoning.

2.3 Deformable Modeling and Query-based Learning

Deformable convolution [9, 39] enables adaptive spatial sampling by learning input-dependent offsets for the convolution grid. While originally designed for 2D image recognition, this design naturally extends to temporal modeling. Query-based learning, popularized by DETR [2], uses learnable queries with cross-attention to aggregate task-relevant information from encoded features, and has been widely adopted in detection [19, 40, 41], video understanding [24, 36], and action anticipation [20, 27]. In this work, we extend deformable convolution to learn spatio-temporal offsets in causal temporal convolutions, allowing each joint

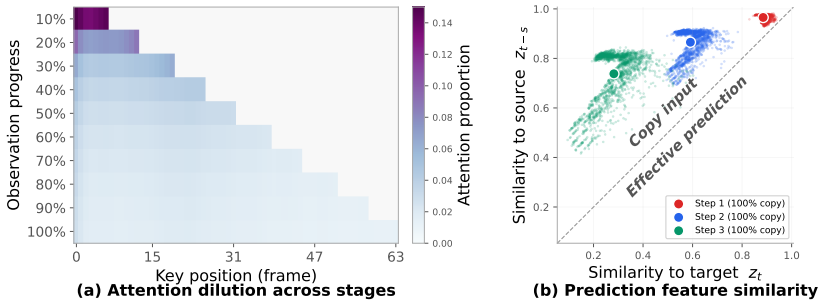

Fig. 2: The limitations of existing online recognition paradigms (e.g., InfoGCN++ [8]). (a) Attention proportion across ten evaluation stages. As the observed sequence grows, causal attention distributes increasingly uniformly across all past frames. (b) Feature similarity analysis of ODE-based future prediction [5]. Each point represents a prediction evaluated at a single frame from one sequence; the x-axis shows similarity to the target $\mathbf{z}_t$ and the y-axis shows similarity to the input source $\mathbf{z}_{t-s}$. Steps 1–3 correspond to predicting 1, 2, and 3 frames ahead.

to attend to distinct temporal positions, and introduce learnable queries that interact with adaptively constructed memory to capture action evolution patterns.

3 Methods

3.1 Preliminary Analysis

As InfoGCN++ [8] represents the current SOTA for online skeleton-based action recognition, we revisit its two core designs—causal temporal attention and Neural ODE-based [5] future prediction—and identify key limitations in both components that motivate the design of our approach.

More history, less focus. While the causal mask successfully ensures valid online inference, each frame attends to all available past frames with learned weights. As the observed sequence grows, the attention distribution tends to spread across an increasingly large history. As shown in Fig. 2(a), at later observation stages the attention weights gradually approach a near-uniform distribution, behaving similarly to average pooling over the entire history. This suggests that the model may have difficulty concentrating on the few critical moments that carry the most discriminative information, calling for a more adaptive and content-aware temporal sampling mechanism.

Predicting the future by copying the present. Moreover, InfoGCN++ [8] attempts to predict future motion from partially observed sequences. Specifically, given the encoded feature $\mathbf{z}_{t-s}$ at frame $t-s$, a Neural ODE module is trained with MSE loss to predict the feature $\mathbf{z}_t$ at frame t, where s denotes the prediction step. However, human motion is inherently ambiguous: the same partial observation can lead to vastly different continuations depending on the

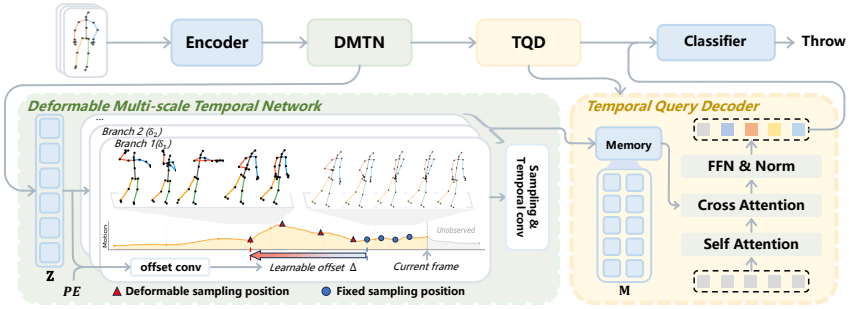

Fig. 3: Overview of model architecture.

person's intent. Under this ambiguity, minimizing the per-sample MSE loss encourages the model to reproduce the current frame features as the predicted future, converging to a trivial solution. As shown in Fig. 2(b), the predicted features exhibit consistently higher cosine similarity to the input source frame than to the intended target frame, suggesting that the prediction has largely collapsed into near-identity copying, motivating an alternative approach that avoids direct instance-level prediction altogether.

3.2 Overview

The above analysis points to two requirements for online recognition: adaptively focusing on discriminative moments from a growing history, and reasoning about motion trends without explicit future forecasting. To this end, we propose DART, which consists of three main components: (1) a spatial-temporal encoder that extracts per-frame representations under the causal constraint, (2) a Deformable Multi-scale Temporal Network that replaces fixed sampling patterns with input-dependent offsets, enabling content-adaptive temporal sampling for each frame and joint, and (3) a Temporal Query Decoder that interacts with the temporal memory formed by DMTN to distill action evolution patterns. An overview of the proposed framework is illustrated in Fig. 3.

3.3 Spatial-Temporal Encoder

Given an input skeleton sequence $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times T \times V \times M}$, where C_{in} denotes the input feature dimension (e.g., 3D joint coordinates), T is the number of frames, V is the number of body joints, and M is the number of persons, the goal of online skeleton-based action recognition is to predict the action label y at each frame t using only the observed frames up to and including t. We encode each frame into a d-dimensional representation using a causal Transformer with SA-GC [7] spatial modeling, following InfoGCN++ [8]:

$$\mathbf{Z}_t = \text{Encoder}(\mathbf{X}_{1:t}) \in \mathbb{R}^{d \times V}, \quad t = 1, 2, \ldots, T. \tag{1}$$

This causal constraint prohibits the model from accessing any future frame information during both training and inference.

3.4 Deformable Multi-scale Temporal Network

As illustrated in Fig. 1, discriminative moments vary in temporal location across actions and observation stages, yet conventional temporal convolutions sample at fixed positions regardless of input content. We address this by proposing a Deformable Multi-scale Temporal Network (DMTN) that learns input-dependent spatio-temporal offsets to dynamically locate informative moments.

Deformable Causal Convolution. For a standard causal convolution with kernel size K and base dilation rate δ, the regular sampling positions relative to the current frame t are strictly historical: $\{t-(K-1)\delta, \ldots, t-\delta, t\}$. We augment each sampling position with a learnable temporal offset $\Delta_{t,v,k}$ predicted directly from the input features. Let $k \in \{0, 1, \ldots, K-1\}$ index the kernel steps. The adaptive sampling position for joint v at time t is formulated as:

$$p_{t,v,k} = t - k\delta + \Delta_{t,v,k}, \tag{2}$$

where the offset $\Delta_{t,v,k}$ is generated by a lightweight network $g(\cdot)$ consisting of two convolutional layers:

$$\Delta_{t,v,k} = g\left([\mathbf{Z}_{t,v} \parallel \mathrm{PE}(t)]\right). \tag{3}$$

Here, $[\cdot \parallel \cdot]$ denotes channel-wise concatenation, $\mathbf{z}_{t,v}$ is the current spatial representation, and $\mathrm{PE}(t)$ is a sinusoidal temporal positional encoding that injects absolute temporal awareness for discriminative frame localization.

Crucially, the offsets are initialized to zero, ensuring the network stably starts from the standard fixed-dilation behavior. To strictly maintain causality during the deformable sampling, the absolute positions $p_{t,v,k}$ are clamped to the valid historical range $[0, t]$, preventing any non-causal information leakage. Since $p_{t,v,k}$ is fractional, the features at these continuous temporal positions are obtained via linear interpolation and subsequently aggregated:

$$\hat{\mathbf{Z}}_{t,v} = \mathrm{Conv}_{1\times K}\left(\left[\phi(\mathbf{Z}_v, p_{t,v,k})\right]_{k=0}^{K-1}\right), \tag{4}$$

where $\phi(\cdot, p)$ denotes the fractional sampling function at temporal position p for joint v, and $\mathbf{Z}_v \in \mathbb{R}^{d\times T}$ represents the temporal feature sequence of joint v.

Multi-scale Architecture. Each temporal layer comprises B deformable causal branches with distinct base dilation rates $\{\delta_1, \ldots, \delta_B\}$, alongside a causal max-pooling branch and an identity branch inspired by [6, 7]. The input channels are evenly split among branches and the outputs from all branches are concatenated along the channel dimension and fused via a residual connection. L such layers are stacked to construct the complete temporal network. Since each deformable branch samples temporal positions independently per joint, two SA-GC [7] layers are applied after the final temporal layer to re-integrate spatial

features within the global skeletal context. Finally, the representations are averaged over all joints V and persons M to produce the compact per-frame feature vectors $\mathbf{z}_t \in \mathbb{R}^d$. The learned sampling positions from each deformable branch are further leveraged to construct the temporal memory for query-based reasoning.

3.5 Temporal Query Decoder

As analyzed in Sec. 3.1, explicit future prediction suffers from degeneracy under motion ambiguity. Instead, we introduce a query-based reasoning module that operates on an adaptively constructed temporal memory to distill complementary temporal cues for classification.

Temporal Memory Construction. The deformable convolutions in Sec. 3.4 learn to dynamically attend to the most informative past frames. We systematically reuse these learned sampling positions to construct a compact temporal memory. Specifically, at a given frame t, each deformable branch has learned K sampling positions. Across all B branches, this yields $N_m = B \times K$ temporal positions. Let n index these aggregated positions, where $n = 1, 2, \ldots, N_m$. We extract d-dimensional features at these fractional positions from the temporal feature maps prior to spatial refinement via linear interpolation. The sampled features are subsequently averaged to form the temporal memory bank:

$$\mathbf{m}_{t,n} = \frac{1}{V} \sum_{v=1}^{V} \phi(\mathbf{Z}_v, p_{t,v,n}), \tag{5}$$

$$\mathbf{M}_t = \left[\mathbf{m}_{t,1}, \mathbf{m}_{t,2}, \ldots, \mathbf{m}_{t,N_m}\right]^\top \in \mathbb{R}^{N_m \times d}, \tag{6}$$

where $\mathbf{M}_t$ is the temporal memory bank at frame t, and $\mathbf{m}_{t,n}$ is the feature extracted at the n-th aggregated sampling position via linear interpolation. This design offers two key advantages. First, unlike standard cross-attention that must attend over the entire observed history, the memory bank provides a compact set of N_m tokens pre-selected at the most informative temporal locations, avoiding the attention dilution problem. Second, since the sampling positions are input-dependent, the memory content naturally adapts to different action classes and observation progress, enabling the model to capture discriminative temporal patterns specific to each instance.

Query Decoder. The temporal memory is first projected into a d_q-dimensional space via a linear projection with Layer Normalization (LN). We then introduce a set of N_q learnable action queries $\mathbf{Q}^{(0)} = \mathbf{Q}_{\text{emb}} \in \mathbb{R}^{N_q \times d_q}$, which interact with the memory through a stack of N_{dec} standard Transformer decoder layers. Each layer consists of three sequential operations with pre-normalization and standard residual connections as follows:

$$\mathbf{Q}' = \mathbf{Q}^{(l)} + \text{SelfAttn}(\text{LN}(\mathbf{Q}^{(l)})), \tag{7}$$

$$\mathbf{Q}'' = \mathbf{Q}' + \text{CrossAttn}(\text{LN}(\mathbf{Q}'), \mathbf{M}_t), \tag{8}$$

$$\mathbf{Q}^{(l+1)} = \mathbf{Q}'' + \mathrm{FFN}(\mathrm{LN}(\mathbf{Q}'')), \tag{9}$$

where the self-attention mechanism enables information exchange among queries, and the cross-attention allows each query to distil relevant evolution patterns from the temporal memory $\mathbf{M}_t$. A final LN is applied to produce the distilled query output $\mathbf{Q}_t^{(N_{\mathrm{dec}})} \in \mathbb{R}^{N_q \times d_q}$ at each frame t.

Feature Fusion and Classification. Instead of regressing future features, we employ the query outputs as condensed representations that encode action evolution patterns. The query outputs are projected back to the original encoder feature dimension, averaged into a single representative vector, and injected into the current frame's representation:

$$\mathbf{z}'_t = \mathbf{z}_t + \frac{1}{N_q} \sum_{i=1}^{N_q} W_{\mathrm{out}} \mathbf{q}_{t,i}^{(N_{\mathrm{dec}})}, \tag{10}$$

where $W_{\mathrm{out}} \in \mathbb{R}^{d \times d_q}$ is the projection weight and $\mathbf{q}_{t,i}^{(N_{\mathrm{dec}})} \in \mathbb{R}^{d_q}$ denotes the i-th individual query vector extracted from the final decoder output matrix $\mathbf{Q}_t^{(N_{\mathrm{dec}})}$. The augmented feature $\mathbf{z}'_t$ is fed into a linear classifier to predict the label $\hat{y}_t$.

3.6 Training Objective

The entire framework is trained end-to-end. The total training objective consists of a main classification loss and an auxiliary query loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{main}} + \lambda \mathcal{L}_{\mathrm{aux}}, \tag{11}$$

where λ is a balancing hyperparameter.

The main objective $\mathcal{L}_{\mathrm{main}}$ is a label-smoothed Cross-Entropy (CE) loss applied to the per-frame predictions across the entire sequence of length T:

$$\mathcal{L}_{\mathrm{main}} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_{\mathrm{CE}}(\hat{y}_t, y), \tag{12}$$

where y is the ground-truth action label, and $\hat{y}_t$ is the prediction at the t-th frame. Furthermore, to explicitly encourage each learnable query to independently distill discriminative temporal patterns from the temporal memory, we apply an auxiliary classifier to each individual query output at every time step. These outputs are supervised with the same ground-truth label y:

$$\mathcal{L}_{\mathrm{aux}} = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N_q} \sum_{i=1}^{N_q} \mathcal{L}_{\mathrm{CE}}\big(W_{\mathrm{aux}} \mathbf{q}_{t,i}^{(N_{\mathrm{dec}})}, y\big) \right), \tag{13}$$

where $W_{\mathrm{aux}} \in \mathbb{R}^{N_c \times d_q}$ is a shared linear classifier across all N_q queries, N_c is the number of action classes. This auxiliary supervision provides a direct learning signal to each query, complementing the indirect gradient from the main classification through feature fusion.

Table 1: Performance comparisons at different observation ratios for a single model trained using the full observation on NTU RGB+D 60 [23] and NTU RGB+D 120 [18]. The best results are highlighted in **bold**.

Datasets	Methods	Observation Ratio										AUC
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
NTU 60 X-Sub	ST-GCN [34]	6.86	10.78	18.09	33.44	53.60	67.68	74.71	78.34	79.81	81.50	50.48
	2S-AGCN [25]	8.69	12.34	20.79	33.54	47.70	61.88	73.31	81.45	85.58	88.50	51.38
	CTR-GCN [6]	5.19	9.02	19.11	33.93	51.93	66.32	77.89	86.07	88.95	89.93	52.83
	InfoGCN [7]	8.26	12.14	21.51	35.65	54.88	70.65	81.55	87.33	**89.42**	89.80	55.12
	MixFormer [33]	5.32	11.48	22.19	36.66	55.23	69.45	79.53	85.45	88.90	90.73	54.94
	STST [37]	12.91	23.71	32.64	42.34	62.04	69.42	75.47	79.10	85.70	**91.90**	57.92
	InfoGCN++ [8]	28.56	44.57	64.66	73.59	78.47	81.68	83.15	84.42	85.05	85.38	70.95
	DART(Ours)	**29.79**	**48.75**	**67.84**	**76.63**	**81.74**	**85.12**	**86.78**	**87.69**	87.96	87.63	**73.99**
NTU 60 X-View	ST-GCN [34]	10.24	14.02	21.40	35.81	53.78	70.85	82.07	86.55	88.72	88.00	55.14
	2S-AGCN [25]	9.12	14.25	23.83	38.34	54.56	69.57	81.56	89.57	93.21	95.10	56.91
	CTR-GCN [6]	6.47	11.07	21.59	35.10	55.22	71.37	83.71	91.60	94.00	94.67	56.48
	InfoGCN [7]	8.10	12.59	24.32	39.09	59.47	75.41	86.47	92.50	**94.51**	95.20	58.77
	MixFormer [33]	6.12	12.04	23.98	38.91	57.75	73.49	84.89	91.25	93.17	95.76	57.73
	STST [37]	14.03	28.80	42.70	54.39	64.78	70.65	77.10	83.20	90.00	**96.80**	62.25
	InfoGCN++ [8]	35.88	55.57	76.39	84.62	88.68	90.79	91.80	92.35	92.67	92.55	80.13
	DART(Ours)	**37.49**	**58.09**	**78.59**	**86.43**	**90.53**	**92.34**	**93.29**	**93.87**	94.07	93.93	**81.86**
NTU 120 X-Sub	ST-GCN [34]	5.30	7.09	11.41	19.02	29.57	43.04	56.30	67.74	75.23	77.44	39.21
	2S-AGCN [25]	7.04	10.59	17.41	28.67	42.19	55.50	66.70	74.34	79.09	80.36	46.19
	CTR-GCN [6]	5.19	5.08	10.52	20.64	38.50	54.09	67.38	77.91	83.07	84.93	44.73
	InfoGCN [7]	3.80	5.77	10.25	18.86	33.98	50.17	64.76	77.13	82.82	85.10	43.26
	MixFormer [33]	4.64	6.49	12.86	22.66	39.41	53.77	64.55	75.19	80.98	85.82	44.67
	STST [37]	1.72	7.14	22.01	37.51	57.91	67.08	73.36	79.82	**84.32**	**87.60**	51.85
	InfoGCN++ [8]	22.13	33.63	48.16	58.86	66.28	71.72	74.45	76.35	77.26	77.42	60.63
	DART(Ours)	**24.92**	**36.73**	**52.60**	**63.43**	**70.80**	**76.19**	**78.90**	**80.62**	81.28	81.12	**64.66**
NTU 120 X-Set	ST-GCN [34]	6.20	8.35	13.04	21.37	32.64	46.20	60.34	71.74	78.10	80.36	41.83
	2S-AGCN [25]	7.47	10.95	18.62	30.18	44.04	57.02	68.44	76.98	81.24	82.53	47.75
	CTR-GCN [6]	5.19	9.02	11.68	21.02	37.06	51.99	66.32	78.63	84.35	85.93	45.12
	InfoGCN [7]	4.91	7.59	14.76	25.67	43.05	58.47	71.04	80.58	84.96	86.30	47.73
	MixFormer [33]	5.14	8.66	13.47	24.88	40.30	53.95	65.81	76.34	81.50	87.64	45.77
	STST [37]	2.51	8.21	24.10	40.11	59.79	71.33	77.32	82.62	**86.62**	**89.44**	54.20
	InfoGCN++ [8]	**27.09**	38.70	54.93	64.90	71.32	76.01	78.54	80.31	81.04	81.20	65.40
	DART(Ours)	26.89	**39.92**	**56.89**	**67.11**	**73.79**	**78.69**	**81.16**	**82.81**	83.47	83.32	**67.40**

4 Experiment

4.1 Datasets and Settings

NTU RGB+D 60 [23] contains 56,880 skeleton sequences across 60 action categories performed by 40 subjects. We evaluate under two standard benchmarks: Cross-Subject (X-Sub), where training and testing are split by subject identity, and Cross-View (X-View), where the split is based on camera viewpoints.

NTU RGB+D 120 [18] extends NTU 60 to 114,480 sequences covering 120 action categories, performed by 106 subjects across 32 camera setups. We eval-

Table 2: Performance comparisons at different observation ratios for a single model trained using the full observation on the NW-UCLA [28] dataset. The best results are highlighted in **bold**.

Dataset	Methods	Observation Ratio										AUC
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
NW-UCLA	ST-GCN [34]	12.72	12.93	21.98	33.84	47.41	57.54	66.59	78.45	82.76	87.50	50.17
	2S-AGCN [25]	17.89	24.57	38.36	51.94	61.21	68.53	76.08	83.19	88.36	90.87	60.10
	CTR-GCN [6]	10.56	13.79	29.09	43.32	56.90	71.55	81.68	91.16	**93.53**	**95.04**	58.66
	InfoGCN [7]	18.32	23.71	37.50	57.54	69.40	79.74	85.78	90.52	91.38	94.00	64.79
	MixFormer [33]	24.38	33.72	44.45	59.00	69.81	81.40	85.38	88.19	92.15	95.50	67.40
	STST [37]	23.38	39.25	47.44	66.28	73.48	81.34	85.20	88.54	91.49	94.42	69.08
	InfoGCN++ [8]	**70.69**	**78.02**	83.84	85.99	86.20	88.15	89.80	90.73	90.95	90.09	85.47
	DART(Ours)	70.26	77.80	**84.91**	**86.85**	**89.01**	**89.44**	**90.30**	**91.38**	91.81	91.59	**86.34**

uate under Cross-Subject (X-Sub) and Cross-Setup (X-Set) benchmarks, where the latter splits training and testing by camera configuration.

NW-UCLA [28] consists of 1,494 skeleton sequences of 10 action categories captured simultaneously by three Kinect cameras from different viewpoints. Following the standard protocol, we use samples from the first two cameras for training and the third for evaluation.

Implementation Details. All skeleton sequences are uniformly sampled to T=64 frames. The spatial-temporal encoder follows the configuration of InfoGCN++ [8] with feature dimension d=128 and we follow the same evaluation protocol. The deformable temporal network consists of L=4 layers with kernel size K=5, base dilations δ_1, δ_2=1, 2. The query decoder uses N_q=10 queries, N_{dec}=4 decoder layers. The auxiliary loss weight is λ=0.5. All experiments are conducted on a single NVIDIA RTX 3090 GPU. We train for 90 epochs using SGD with momentum 0.9, and weight decay 3×10^{-4}. The initial learning rate is 0.1, decayed by a factor of 0.1 at epochs 50 and 70. The offset prediction network uses a 10× higher learning rate to accelerate convergence.

4.2 Comparison with SOTA Methods

NTU RGB+D 60. Tab. 1 presents the results on NTU RGB+D 60 [23]. Our method achieves AUC of 73.99% and 81.86% on the Cross-Subject and Cross-View splits, outperforming InfoGCN++ [8] by 3.04% and 1.73%, respectively. The improvement is most pronounced at early-to-mid observation stages (e.g., 20%–50%), where the available history is short and fixed sampling patterns are particularly wasteful. DMTN's adaptive sampling can concentrate on the few informative frames rather than being locked to predetermined positions. At 90%–100%, some offline methods (e.g., STST [37]) achieve higher per-stage accuracy as they optimize solely for full-sequence recognition without the need to maintain performance across all observation stages. As an online method, DART must maintain strong performance across all observation stages, yet still achieves the highest overall AUC on both splits.

Table 3: Ablation studies on the NTU RGB+D 60 [23] X-Sub dataset. We systematically evaluate the effectiveness of our proposed components and key hyperparameters. Default settings for our full model are highlighted in **bold**.

DMTN	TQD	AUC (%)
		69.25
	✓	70.27
✓		73.37
✓	✓	**73.99**

(a) Component-wise ablation.

Sampling Strategy	AUC (%)
Fixed dilation	72.37
Learnable offset	**73.37**

(b) Deformable vs. Fixed.

L	AUC (%)
2	71.68
3	72.42
4	**73.37**
5	73.03

(c) DMTN layers L.

N_q	AUC (%)
2	73.42
6	73.62
10	**73.99**
14	73.29

(d) Number of queries N_q.

Model	Params	FLOPs
InfoGCN++	2.50M	10.00G
DART (Ours)	**2.35M**	**4.59G**

(e) FLOPs and params.

N_{dec}	AUC (%)
2	73.36
3	73.58
4	**73.99**
5	73.57

(f) TQD layers N_{dec}.

NTU RGB+D 120. On the larger-scale NTU RGB+D 120 [18] (Tab. 1), our method achieves 64.66% and 67.40% AUC on the Cross-Subject and Cross-Setup splits, surpassing InfoGCN++ [8] by 4.03% and 2.00%, respectively. The larger margin on this more challenging benchmark suggests that as the action vocabulary grows and temporal patterns become more diverse, content-adaptive sampling and query-based temporal reasoning become more beneficial.

NW-UCLA. On NW-UCLA [28], our method achieves 86.34% AUC, improving over InfoGCN++ [8] by 0.87%. While the margin is smaller on NW-UCLA, as its limited training set (∼1.5K samples) constrains the learning of input-dependent adaptive sampling patterns and the query priors, our method consistently outperforms at the 30%–80% stages and attains the highest overall average.

4.3 Ablation Studies

We conduct comprehensive ablation studies on the NTU RGB+D 60 [23] Cross-Subject split to evaluate the contribution of each proposed component and the sensitivity to key hyperparameters. All results are summarized in Tab. 3.

Effectiveness of Each Component. As shown in Tab. 3(a), adding DMTN to the baseline yields an AUC of 73.37%, and further incorporating TQD brings the performance to 73.99% (+0.62%), demonstrating that both components provide complementary benefits. Notably, applying TQD alone with uniform memory already improves the baseline by 1.02%, confirming that the query decoder is independently effective. Tab. 3(b) further isolates the effect of deformable sampling: replacing the learnable offsets with fixed dilation patterns causes a 1.00% drop (72.37% vs. 73.37%). This confirms that content-adaptive temporal sam-

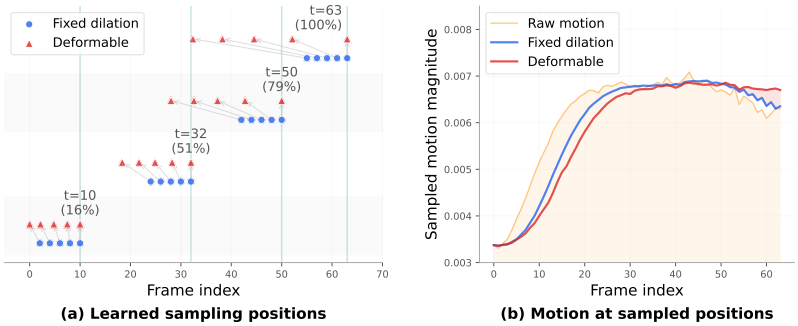

Fig. 4: Visualization of the learned deformable temporal sampling on the NTU RGB+D 60 [23] Cross-Subject split. (a) Learned sampling positions: fixed dilation (blue circles) vs. deformable (red triangles) at four observation stages. (b) Aggregated motion magnitude at the sampled positions, with the red shaded area indicating where deformable sampling exceeds fixed dilation.

pling is essential, as fixed patterns cannot accommodate the diverse temporal dynamics across different action categories and instances.

DMTN Depth. Tab. 3(c) varies the number of deformable temporal layers L. Performance steadily improves from L=2 (71.68%) to L=4 (73.37%), as deeper networks progressively expand the effective temporal receptive field through stacked deformable convolutions. However, increasing to L=5 causes a 0.34% drop, indicating that excessive depth introduces optimization difficulty without further representational gain. We therefore adopt L=4 as the default setting.

Query Decoder Configuration. Tab. 3(d) and Tab. 3(f) examine the number of queries N_q and decoder layers N_{dec}, respectively. For N_q, performance improves as the number of queries increases from 2 to 10, reaching the peak of 73.99%, as more queries provide greater capacity for extracting relevant cues from the temporal memory. However, performance drops at N_q=14 (73.29%), suggesting that excessive queries introduce redundancy and dilute the learning signal from the auxiliary supervision. A similar trend holds for N_{dec}, where 4 layers achieve the best result and adding a fifth layer yields diminished returns (73.57%), likely due to overfitting on the limited temporal memory tokens.

Computational Efficiency. As shown in Tab. 3(e), DART requires only 2.35M parameters and 4.59G FLOPs, reducing computational cost by 54% compared to InfoGCN++ (10.00G FLOPs) while using fewer parameters (2.35M vs. 2.50M). This is because DART replaces the computationally expensive Neural ODE module which requires iterative numerical integration with lightweight deformable offsets and a compact query decoder, achieving both stronger performance (AUC +3.04% on NTU RGB+D 60) and higher efficiency.

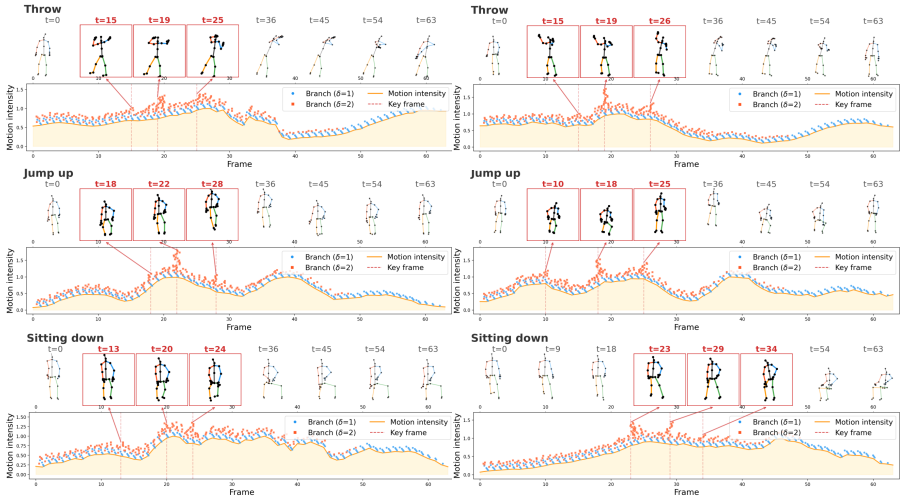

Fig. 5: Qualitative visualization of the learned deformable sampling on individual action instances from three categories. For each sample, the top row shows skeleton frames at selected time steps, with red boxes highlighting the key frames that receive the most sampling attention. The bottom row plots the motion intensity curve alongside the sampling positions from two branches (δ=1 and δ=2), where vertical stacking indicates higher sampling frequency.

5 Visual Analysis

Visualization of Deformable Temporal Sampling. We visualize the learned sampling behavior of the last DMTN layer on the NTU RGB+D 60 [23] Cross-Subject split in Fig. 4. As shown in Fig. 4(a), at early stages (e.g., t=10), the deformable positions spread more uniformly across the available history than the clustered fixed pattern, achieving a more comprehensive temporal coverage. At later stages (e.g., t=63), the offsets shift substantially toward earlier frames, expanding the receptive field beyond what fixed dilations can reach. Fig. 4(b) further shows that this adaptive sampling enables the model to maintain high motion capture even when overall motion declines in the later portion, confirming that the learned offsets effectively steer toward informative temporal regions.

Visualization of samples. To qualitatively examine whether the learned offsets adapt to individual action dynamics, we visualize the deformable sampling behavior on representative instances from three action categories in Fig. 5. For each sample, we identify the key frames—those most frequently attended by the deformable branches—and overlay the sampling positions onto the motion intensity curve. It is also notable that branch 1 (δ=1) consistently focuses on nearby frames, while branch 2 (δ=2) freely shifts to distant positions, naturally forming a short-term and long-term temporal memory. Across all examples, the sampling concentrates around the motion peaks: the throwing phase in *Throw*, the takeoff in *Jump up*, and the postural transition in *Sitting down*. Importantly, even for

the same action class, the temporal location of these key frames shifts across different instances (e.g., *Sitting down* peaks at t=6–19 in the left sample vs. t=23–34 in the right), demonstrating that the offsets are driven by per-instance content rather than fixed class-level priors.

6 Conclusion

We propose DART, a deformable adaptive reasoning framework for online skeleton-based action recognition. We first identify two limitations of existing online paradigms and thus propose two complementary components. The DMTN replaces fixed dilation patterns with input-dependent offsets, enabling each frame and joint to adaptively sample from the most discriminative temporal positions. The TQD avoids explicit future forecasting and instead leverages learnable queries to distill action evolution patterns. Extensive experiments on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA demonstrate that DART achieves state-of-the-art performance across all benchmarks.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62271221 and the National Social Science Foundation of China under Grant No. 25BXW041. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation.

References

1. Brehar, R.D., Muresan, M.P., Mariţa, T., Vancea, C.C., Negru, M., Nedevschi, S.: Pedestrian street-cross action recognition in monocular far infrared sequences. IEEE Access **9**, 74302–74324 (2021)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
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, L., Lu, J., Song, Z., Zhou, J.: Ambiguousness-aware state evolution for action prediction. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6058–6072 (2022)
5. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
6. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13359–13368 (2021)

7. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20186–20196 (2022)
8. Chi, S., Chi, H.G., Huang, Q., Ramani, K.: Infogcn++: Learning representation by predicting the future for online skeleton-based action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(1), 514–528 (2025)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
10. Do, J., Kim, M.: Skateformer: skeletal-temporal transformer for human action recognition. In: European Conference on Computer Vision. pp. 401–420. Springer (2024)
11. Foo, L.G., Li, T., Rahmani, H., Ke, Q., Liu, J.: Era: Expert retrieval and assembly for early action prediction. In: European Conference on Computer Vision. pp. 670–688. Springer (2022)
12. Foo, L.G., Li, T., Rahmani, H., Ke, Q., Liu, J.: Unified pose sequence modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13019–13030 (2023)
13. Guermal, M., Ali, A., Dai, R., Bremond, F.: Joadaa: joint online action detection and action anticipation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 6889–6898 (2024)
14. Ke, Q., Bennamoun, M., Rahmani, H., An, S., Sohel, F., Boussaid, F.: Learning latent global network for skeleton-based action prediction. IEEE Transactions on Image Processing **29**, 959–970 (2019)
15. Li, T., Liu, J., Zhang, W., Duan, L.: Hard-net: Hardness-aware discrimination network for 3d early activity prediction. In: European conference on computer vision. pp. 420–436. Springer (2020)
16. Li, T., Luo, Y., Zhang, W., Duan, L., Liu, J.: Harder-net: Hardness-guided discrimination network for 3d early activity prediction. IEEE transactions on circuits and systems for video technology **34**(12), 12112–12126 (2024)
17. Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., Sun, Z.: Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29248–29257 (2025)
18. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
19. Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., Bai, X.: End-to-end temporal action detection with transformer. IEEE Transactions on Image Processing **31**, 5427–5441 (2022)
20. Pang, Z., Sener, F., Yao, A.: Context-enhanced memory-refined transformer for online action detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8700–8710 (2025)
21. Pei, M., Jia, Y., Zhu, S.C.: Parsing video events with goal inference and intent prediction. In: 2011 international conference on computer vision. pp. 487–494. IEEE (2011)
22. Pop, D.O., Rogozan, A., Chatelain, C., Nashashibi, F., Bensrhair, A.: Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction. IEEE Access **7**, 149318–149327 (2019)

23. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
24. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18857–18866 (2023)
25. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
26. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 98–106 (2016)
27. Wang, J., Chen, G., Huang, Y., Wang, L., Lu, T.: Memory-and-anticipation transformer for online action understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13824–13835 (2023)
28. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2649–2656 (2014)
29. Wang, L., Koniusz, P.: 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5620–5631 (2023)
30. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
31. Wang, W., Chang, F., Liu, C., Wang, B., Liu, Z.: Todo-net: Temporally observed domain contrastive network for 3-d early action prediction. IEEE Transactions on Neural Networks and Learning Systems **36**(4), 6122–6133 (2024)
32. 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)
33. Xin, W., Miao, Q., Liu, Y., Liu, R., Pun, C.M., Shi, C.: Skeleton mixformer: Multivariate topology representation for skeleton-based action recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2211–2220 (2023)
34. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
35. Yan, T., Zeng, W., Xiao, Y., Tong, X., Tan, B., Fang, Z., Cao, Z., Zhou, J.T.: Crossglg: Llm guides one-shot skeleton-based 3d action recognition in a cross-level manner. In: European Conference on Computer Vision. pp. 113–131. Springer (2024)
36. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022)
37. Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: Stst: Spatial-temporal specialized transformer for skeleton-based action recognition. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3229–3237 (2021)
38. Zhou, Y., Yan, X., Cheng, Z.Q., Yan, Y., Dai, Q., Hua, X.S.: Blockgcn: Redefine topology awareness for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2049–2058 (2024)

39. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)
40. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
41. Zhu, Y., Zhang, G., Tan, J., Wu, G., Wang, L.: Dual detrs for multi-label temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18559–18569 (2024)