


Article

SUGAR: A Sequence Unfolding Based Transformer Model for Group Activity Recognition

Yash Gondkar ¹, Chengjie Zheng ¹, Yumeng Yang ², Shiqian Shen ³, Wei Ding ^{1,*} and Ping Chen ¹

¹ Department of Computer Science, University of Massachusetts Boston, Boston, MA 02125, USA

² HP Inc., Palo Alto, CA 94304, USA

³ Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

* Correspondence: wei.ding@umb.edu

How To Cite: Gondkar, Y.; Zheng, C.; Yang, Y.; et al. SUGAR: A Sequence Unfolding Based Transformer Model for Group Activity Recognition. *Transactions on Artificial Intelligence* **2025**, *1*(1), 227–245. <https://doi.org/10.53941/tai.2025.100015>

Received: 10 June 2025

Revised: 12 August 2025

Accepted: 17 September 2025

Published: 28 September 2025

Abstract: Deep learning models built upon Transformer architectures have led to substantial advancements in sequential data analysis. Nevertheless, their direct application to video-based tasks, such as Group Activity Recognition (GAR), remains constrained by the quadratic computational complexity and excessive memory requirements of global self-attention, especially when handling long video sequences. To overcome these limitations, we propose *SUGAR: A Sequence Unfolding Based Transformer Model for Group Activity Recognition*. Our approach introduces a novel sequence unfolding and folding mechanism that partitions long video sequences into overlapping local windows, enabling the model to concentrate attention within compact temporal regions. This local attention design dramatically reduces computational cost and memory footprint while maintaining high recognition accuracy. Within the Bi-Causal framework, SUGAR replaces conventional Transformer blocks, and experimental results on the Volleyball dataset demonstrate that our model achieves state-of-the-art performance, consistently exceeding 93% accuracy, with significantly improved efficiency. In addition, we investigate Lightning Attention 2 as an alternative linear-complexity attention module, identifying practical challenges such as increased memory usage and unstable convergence. To ensure robustness and training stability, we incorporate a dedicated safety mechanism that mitigates these issues. In summary, SUGAR offers a scalable, resource-efficient solution for group activity analysis in videos and exhibits strong potential for broader applications involving lengthy sequential data in computer vision and bioinformatics.

Keywords: group activity recognition; transformer; linear attention; sequence modeling; deep learning; efficient architectures; video understanding

1. Introduction

The growing volume of video data has fueled rapid progress in computer vision, especially in tasks involving temporal reasoning [1,2]. Among them, Group Activity Recognition (GAR) stands out as both important and technically demanding [3]. GAR involves analyzing coordinated actions and interactions among multiple individuals over time, with applications in surveillance, sports analytics, autonomous driving, and social behavior analysis [4–6].

Unlike single-person action recognition, GAR requires capturing both individual motion and complex contextual dynamics across time and space [7–9]. Transformer-based models have shown strong performance by



Copyright: © 2025 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

modeling global dependencies via self-attention [10,11]. However, their quadratic complexity with respect to sequence length makes them inefficient for long video inputs, where even short clips can generate thousands of tokens [12,13].

To address this, efficient attention models like Performer [14], Linformer [12], and Lightning Attention 2 [13] have been proposed, aiming to reduce computation to linear complexity. Similarly, Structured State Space Models (e.g., S4 [15], Mamba [16]) offer an alternative path for long-range modeling. Yet, when applied to GAR tasks, these methods often introduce side effects such as optimization instability, loss of fine-grained temporal cues, or memory inefficiency [3].

Recent GAR models like Bi-Causal [17], GroupFormer [18], and Dual-AI [19] achieve competitive accuracy, but still rely on full-sequence attention and struggle with scalability. This creates a pressing need for GAR architectures that balance accuracy, efficiency, and real-world deployability [3,20,21].

1.1. Motivation and Our Approach

Through empirical analysis of GAR datasets, we observed that most critical group interactions occur in local spatiotemporal regions [20–22]. This motivates a selective focus on local segments instead of processing entire sequences uniformly [7,8]. Building on this, we propose SUGAR—a Sequence Unfolding Based Transformer Model for Group Activity Recognition.

SUGAR introduces a sequence unfolding & folding (SUF) mechanism: video sequences are split into overlapping windows, each independently encoded by a Transformer block [10,18]. A folding operation then re-aggregates the outputs into a global sequence representation. This design reduces complexity from quadratic to linear [12,13,16] while preserving critical group interaction patterns [17,19,20].

1.2. System Design and Implementation

We integrate SUGAR into the Bi-Causal framework [17], where unfolding–fusion modules wrap each encoder layer. Additional stability layers are included to enable reliable training with advanced attention modules like Lightning Attention 2 [13], which we found unstable without modification. We analyze how window size and stride affect performance [3,22], and show that the SUF design offers modularity, efficiency, and robustness [18–21].

1.3. Contributions

- (1) We propose SUGAR, a transformer-based GAR model with linear complexity and strong recognition accuracy.
- (2) We show how SUGAR integrates into Bi-Causal [17] while improving scalability on benchmark datasets.
- (3) We introduce a stability module to support Lightning Attention 2 and other linear attention mechanisms.
- (4) We demonstrate that SUGAR is broadly applicable to long-sequence modeling tasks including bioinformatics and animal behavior analysis.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 introduces technical preliminaries, Section 4 describes our model architecture, Section 5 details experiments, Section 6 discusses implications, and Section 7 concludes with future directions.

2. Related Work

2.1. Group Activity Recognition

Group Activity Recognition (GAR) has evolved rapidly, transitioning from early hand-crafted features and probabilistic models to deep learning approaches capable of modeling complex group behaviors. Modern GAR methods leverage graph neural networks and transformer-based architectures to capture both spatial and temporal dependencies in multi-person scenes. Notable frameworks such as Bi-Causal [17], GroupFormer [18], and Dual-AI [19] use actor-centric reasoning and relational graphs to achieve strong recognition performance.

In addition, feature representations based on pose key points and RoIAlign have improved model robustness under real-world conditions [23,24]. However, these transformer-based models suffer from quadratic complexity, limiting their scalability for longer or denser video sequences. Balancing accuracy with computational efficiency remains an open challenge in next-generation GAR systems [25,26].

2.2. Efficient Sequence Modeling in Vision

The Transformer architecture has transformed sequence modeling in both language and vision tasks by enabling global attention across input positions. While effective, its $O(L^2)$ time and space complexity poses serious obstacles for long-form video data. A single video can generate thousands of tokens, making standard self-attention computationally prohibitive.

To address this, several efficient attention mechanisms have been proposed. Models like Performer [12], Linformer [13], and Lightning Attention 2 [27] reduce complexity to linear or near-linear, using techniques such as kernelization or structured projection. Separately, structured state space models (SSMs) like Mamba [15] and S4 [28] offer an alternative route by modeling sequence dynamics through recurrence-inspired operators.

Despite their promise, these techniques often struggle to transfer directly to high-dimensional vision tasks. Many were designed for 1D text or audio and face practical issues when applied to video, such as maintaining spatial context, preserving local interactions, and supporting stable optimization on GPUs [10,29].

2.3. Summary and Open Gaps

In summary, while recent advances have improved GAR performance, most existing models still rely on full-sequence transformers and suffer from inefficiencies in long-video settings. Efficient attention mechanisms offer scalability but often lack robustness when directly applied to dense, real-world group interactions.

Our work aims to bridge this gap by introducing a Sequence Unfolding and Folding (SUF) operation that enables linear complexity while preserving key spatiotemporal dependencies. This mechanism serves as the backbone of our proposed SUGAR framework.

3. Preliminaries

3.1. Traditional Transformer Architecture

Transformer architectures have fundamentally advanced the field of sequential modeling by introducing a novel and powerful attention mechanism [10]. Originally developed for neural machine translation, Transformers overcame key limitations of convolutional and recurrent neural networks, particularly their restricted capacity for capturing long-range dependencies and limited parallelization capabilities.

Central to the Transformer is the self-attention mechanism, which computes the contextualized representation of each token by aggregating information from all other tokens in the sequence. Formally, self-attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q , K , and V denote the query, key, and value matrices, and d_k is the dimension of the keys. This process enables dynamic focus on semantically relevant positions, effectively capturing both local and global contextual information.

A standard Transformer comprises stacked encoder and decoder modules. Each encoder layer contains two primary sub-layers: a multi-head self-attention module and a position-wise feed-forward network, both integrated with layer normalization and residual connections. The decoder mirrors this structure, with the addition of masked self-attention and cross-attention layers. Since self-attention is inherently permutation-invariant, positional encoding is employed to inject sequence ordering information.

Transformers have driven major advances in diverse domains, including natural language processing, video understanding, and biological sequence modeling [11,26]. However, their direct application to tasks such as Group Activity Recognition (GAR) is limited by the quadratic complexity of self-attention, motivating the development of more efficient approaches. In this work, we analyze the computational behavior of the Transformer in the context of GAR, contrast it with emerging alternatives, and introduce a strategy to substantially enhance its scalability and efficiency through our SUGAR framework.

3.2. Lightning Attention 2

Despite the remarkable capabilities of Transformers, their scalability to very long sequences remain fundamentally constrained by the quadratic computational and memory demands of self-attention [10]. This limitation is particularly pronounced for high-dimensional and temporally extended data, such as videos and biological sequences. To alleviate these challenges, a range of efficient attention mechanisms have been developed [12,30].

A notable recent innovation, Lightning Attention 2 [13], achieves linear complexity by employing structured, learnable transformations to approximate the full attention computation. Instead of explicitly calculating all pairwise attention scores, Lightning Attention 2 transforms input sequences through learned kernels, mapping them into lower-dimensional feature spaces in which interactions can be computed with significantly reduced resource requirements. This design has been shown to markedly lower both memory usage and computational cost, making it viable for sequences with thousands of elements.

Empirical evidence suggests that Lightning Attention 2 not only enhances efficiency but can also match or surpass the predictive performance of standard attention across a variety of sequence modeling tasks [13,28]. Nonetheless, practical deployment in video-based tasks such as GAR can expose issues related to hardware resource demands and training instability. To address these concerns, we introduce a dedicated safety layer to facilitate robust and efficient integration within transformer-based architectures. This component plays a critical role in our proposed SUGAR: A Sequence Unfolding Based Transformer Model for Group Activity Recognition, as detailed in Section 4.

3.3. Bi-Causal

Bi-Causal [20] represents a recent advancement in group activity recognition, offering a state-of-the-art framework that models the bidirectional causality between human relations (HRs) and human-object interactions (HOIs). Distinct from earlier methods that primarily focused on HRs, Bi-Causal explicitly characterizes the mutual influence between HRs and HOIs, employing Granger Causality Tests to rigorously validate these dependencies in data. The architecture comprises several interconnected modules:

- **Feature Extraction:** Extraction of person features using HRNet, object features from object annotations, and additional kinematic features.
- **Relation Module:** Modeling of interactions among individuals through graph convolutional networks (GCNs).
- **Interaction Module:** Application of GCNs to capture HOIs, emphasizing how individuals interact with surrounding objects.
- **Causality Communication Channel:** An information conduit that facilitates bidirectional communication between relation and interaction modules.

Comprehensive evaluations on benchmark datasets, such as Volleyball and Collective Activity, demonstrate that Bi-Causal achieves superior performance by leveraging these explicit causal dependencies. This foundation provides an effective backbone for subsequent models, including our proposed SUGAR framework, which further advances efficiency and scalability in group activity recognition.

3.4. Evaluation Metrics

Robust evaluation protocols are essential for accurately measuring model performance and generalizability, particularly in complex settings such as group activity recognition [31]. For multi-class classification tasks like GAR, reliance on a single metric (e.g., accuracy) may obscure nuanced aspects of model behavior, especially in the presence of class imbalance. To address this, a comprehensive suite of evaluation metrics is adopted:

- **Accuracy:** The proportion of correct predictions across all classes, providing an overall measure of classification performance.
- **Confusion Matrix:** Offers a detailed breakdown of class-wise prediction results, illuminating specific strengths and weaknesses.
- **Precision, Recall, and F1-Score:** Precision quantifies the trustworthiness of positive predictions, recall measures the completeness of positive identifications, and F1-score provides a harmonic mean of both. For class c , these are formally defined as:

$$\text{Precision}_c = \frac{TP}{TP + FP}, \quad \text{Recall}_c = \frac{TP}{TP + FN}$$

where TP , FP , and FN denote true positives, false positives, and false negatives, respectively.

- **Precision-Recall (PR) Curve:** Plots precision versus recall across different thresholds, providing valuable insight for imbalanced data distributions.
- **Area Under the PR Curve (AUPRC):** Summarizes the model's ability to prioritize true positives ahead of false positives across all thresholds.
- **AUROC:** The area under the receiver operating characteristic curve, reflecting the overall ranking quality of model predictions.

- **Precision@k:** Measures the fraction of relevant items among the top k predictions, useful for top- k retrieval scenarios.

Collectively, these metrics enable a nuanced and multifaceted assessment of model performance, supporting informed evaluation and comparison of different approaches. In our experiments, we employ these criteria to comprehensively benchmark the SUGAR framework and baseline methods on established datasets.

4. Methodology

4.1. Problem Definition and Notation

We formalize Group Activity Recognition (GAR) as the task of mapping an observation sequence $d[1:t]$ to a group activity label $l \in \mathcal{L}$, where \mathcal{L} denotes the set of possible activity classes. This process can be described by a function $X: \mathcal{D} \rightarrow \mathcal{L}$, such that for any sequence $s \in \mathcal{D}$ of length t with label $l \in \mathcal{L}$, we have $l = X(d[1:t])$. The core challenge is to design an efficient and accurate mapping X that robustly handles long and complex video or sequential data typical of real-world GAR applications [4].

The full model comprises a Bi-Causal Transformer backbone, into which we integrate the proposed Sequence Unfolding and Folding (SUF) mechanism, as well as a safety-adapted version of Lightning Attention 2.

Training Details: We use AdamW optimizer with a learning rate of 2×10^{-4} and weight decay of 1×10^{-2} . Each video is segmented into overlapping windows of length $W = 25$ with step size $S = 5$. During training, we apply temporal jittering and random horizontal flipping. Batch size is set to 32, and all models are trained for 100 epochs with early stopping based on validation loss.

Loss Function: We use standard cross-entropy loss over the predicted group activity logits:

$$\mathcal{L}_C = - \sum_{i=1}^C y_i \log p_i$$

where C is the number of activity classes, y_i is the one-hot ground truth label, and p_i is the predicted softmax output.

4.2. Bi-Causal as Baseline

Our baseline adopts the Bi-Causal framework [29], which currently represents state-of-the-art performance in GAR. Bi-Causal explicitly models the bidirectional causality between Human Relations (HRs) and Human-Object Interactions (HOIs), thereby addressing the shortcomings of previous approaches that focus predominantly on HRs. The Bi-Causal architecture consists of the following core modules:

- **Feature Extraction:** Person features (P), object features (O), and kinematic features (K) are extracted from the input video stream.
- **Relation Module (RM):** Graph convolutional networks (GCNs) are used to model the dynamics of HRs.
- **Interaction Module (IM):** GCNs further capture the spatial-temporal patterns of HOIs.
- **Causality Communication Channel:** Enables information exchange between RM and IM, effectively leveraging their mutual influence.

Despite its superior recognition accuracy, Bi-Causal inherits the quadratic computational complexity of standard Transformer blocks [16], which significantly constrains its efficiency and scalability when processing long video sequences. This motivates the development of more computationally efficient alternatives such as our proposed SUGAR framework.

4.3. The Proposed Sequence Unfolding & Folding Technique

4.3.1. Motivation

Consider a group activity video featuring soccer players John, Dan, and Foo, labeled as “Goal”. John is on team A, Dan on team B, and Foo is the goalkeeper for team A. The activity sequence unfolds as Dan steals the ball from John, advances toward the goal, and eventually scores. Importantly, the final group activity label depends primarily on the decisive event—the goal—rather than on the complete sequence of prior actions. This observation highlights that much of the input sequence may be redundant for accurate classification, motivating a more targeted approach.

4.3.2. Formalizing the Idea

Formally, let $X: \mathcal{D} \rightarrow \mathcal{L}$ as defined above. Suppose there exist indices p and q such that for any auxiliary sequences c and e of lengths p and $t - p - q$ respectively, and $c + d[p + 1: p + q] + e \in \mathcal{D}$, the following holds:

$$l = X(c + d[p + 1: p + q] + e)$$

This implies that a concise segment of the sequence containing the critical event is enough for predicting the group activity label, supporting our local-context-driven design.

4.3.3. Sequence Unfolding & Folding Architecture in SUGAR

To operationalize this insight, our approach comprises the following steps:

- (1) **Residual Padding:** Residual paddings of sizes $\left\lfloor \frac{R}{2} \right\rfloor$ and $\left\lfloor \frac{R+1}{2} \right\rfloor$ are applied to the left and right of the sequence, respectively, totaling R . R is the smallest number which when added to L , makes it divisible by S and is given as

$$R = \left(\left\lfloor \frac{L-1}{S} \right\rfloor + 1 \right) \times S - L.$$

- (2) **Main Padding:** Then additional paddings of size $W - S$ are added at both ends, where W is the window size, giving the fully padded sequence of length given as

$$L_p = L + R + 2 \times (W - S).$$

- (3) **Sequence Unfolding:** The padded sequence is then segmented using a sliding window of length W and stride S , generating overlapping subsequences of size given as

$$B = \frac{L_p - W}{S} + 1.$$

- (4) **Transformer Processing:** Each subsequence is independently processed by a lightweight transformer encoder block.
- (5) **Grouping:** To fold back the returned blocks, we start by grouping them by adjacency resulting in a number of groups given by

$$C = \frac{W}{S}.$$

This is also the number of steps in the sequence after which we get an unfolded block of entirely new tokens, and therefore, we call it the step count. Note that since W will always be divisible by S , C will be an integer.

- (6) **Concatenation:** The grouped sequences are joined together in the order of their occurrence to get C sequences.
- (7) **Averaging:** Unweighed mean of these C sequences is taken.
- (8) **Trimming:** The resulting sequence is then trimmed by $\left\lfloor \frac{L_\Delta}{2} \right\rfloor$ and $\left\lfloor \frac{L_\Delta+1}{2} \right\rfloor$ tokens from the left and right sides to get the final output sequence with length matching the input. L_Δ is therefore the number of excessive tokens in the output of the Averaging phase.

This describes the central Sequence Unfolding & Folding part of the SUGAR architecture diagram shown in Figure 1 and it is also the heart of the SUGAR framework. It enables efficient local-context modeling while maintaining global sequence awareness. By substantially reducing the computational complexity from quadratic to linear in sequence length, SUGAR advances the practical deployment of group activity recognition models on large-scale, long-sequence data [32].

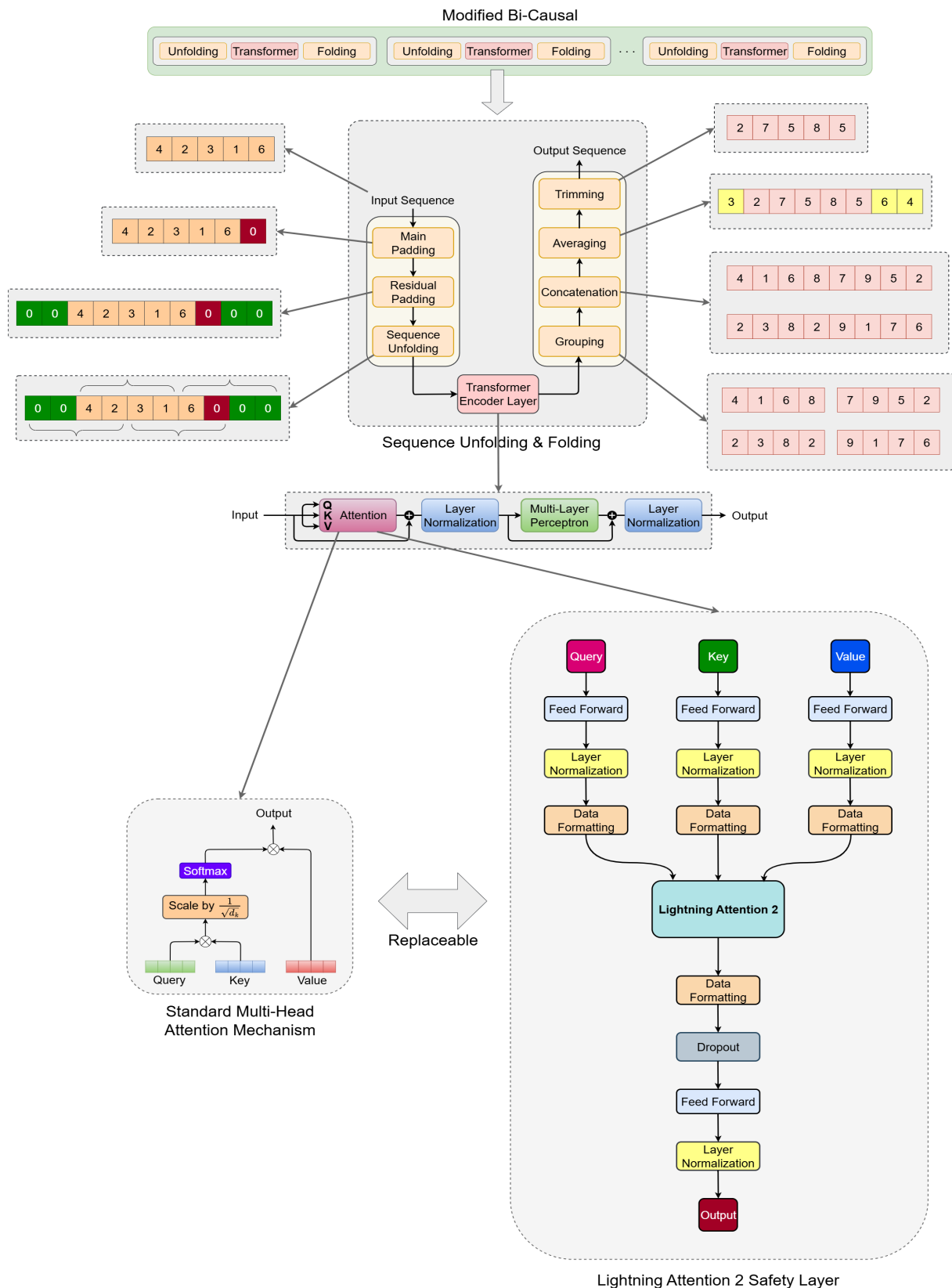


Figure 1. Overview of the SUGAR framework. The model processes an input sequence using a Transformer backbone enhanced by our proposed Sequence Unfolding & Folding (SUF) module. The unfolding step segments the sequence into overlapping local windows, each processed independently by the Transformer encoder. The folding step re-aggregates the outputs into a full-sequence representation, enabling linear complexity while preserving key spatiotemporal features. In this architecture, the SUF modules wrap each Bi-Causal Transformer encoder block. The expanded view illustrates that this structure is repeated consistently across layers, with arrows showing the data flow between SUF and Transformer components.

4.3.4. Example Walkthrough

To illustrate the sequence unfolding and folding process, consider the case where $W = 4$, $S = 2$, and $L = 5$. The sequence is first padded to ensure divisibility for unfolding. Each window of length 4 is independently processed, after which the outputs are grouped, concatenated, and averaged. A final trimming operation restores the original sequence length. This approach ensures that every token is treated uniformly, with no boundary token being over- or under-represented, thereby maintaining consistency across the sequence.

4.3.5. How Sequence Unfolding & Folding Treats All Tokens Equally

Residual padding guarantees that each token is included in at least one window, while main padding ensures that tokens near the boundaries participate in as many windows as those in the center. By choosing W and S such that W is divisible by S , the algorithm ensures that all groups of tokens are processed with equal frequency, thus preventing any positional bias or neglect of specific sequence segments.

4.3.6. Complexity Analysis

In a standard Transformer encoder, processing a sequence of length L requires computing self-attention over all token pairs, leading to a quadratic computational complexity of $O(L^2)$. This becomes a major bottleneck for long sequence tasks, such as video or clinical time series modeling.

In contrast, our proposed Sequence Unfolding and Folding (SUF) mechanism significantly reduces this cost by constraining attention computations to local overlapping windows. Specifically, the input sequence is first padded with residual tokens R to ensure alignment, and then partitioned into overlapping windows, each of fixed size W , and processed independently. This results in a much computationally cheaper overall operation as we will see next.

4.3.6.1. Performance Expression Derivation

In the Unfolding phase, the window will iteratively traverse the padded sequence, and in the i^{th} iteration the first token index will be given as

$$b_i = (i - 1) \times S.$$

This means that the token of the last index will be given as

$$\begin{aligned} l_i &= b_i + W - 1 \\ \Rightarrow l_i &= (i - 1) \times S + W - 1. \end{aligned}$$

However,

$$\begin{aligned} l_i &\leq L_p - 1 \\ \Rightarrow (i - 1) \times S + W - 1 &\leq L_p - 1 \\ \Rightarrow i &\leq \frac{L_p - W}{S} + 1. \end{aligned}$$

Since both L_p and W are divisible by S , the count of subsequences produced by the Unfolding phase is

$$\max\{i\} = \frac{L_p - W}{S} + 1 = \frac{L + R + W}{S} - 1.$$

Therefore, the total number of operations required by the transformer encoder to process the input sequence will improve from L^2 to

$$N = \left(\frac{L + R + W}{S} - 1 \right) \times W^2$$

when using SUGAR.

Hence, for fixed W and S ,

$$N = O(L).$$

Thus, SUF effectively reduces the self-attention complexity from quadratic to linear with respect to the input length, making the SUGAR model highly efficient and scalable for long-sequence applications [33].

4.3.7. Applicability and Scope of the SUF Mechanism

The SUF mechanism does not include an explicit module for modeling global dependencies across the full sequence. Instead, it adopts a task-specific assumption: that key semantic cues in group activity recognition (GAR) typically emerge from short, localized spatiotemporal patterns. To support this, our overlapping window strategy ensures that each token is included in multiple local contexts, allowing the model to build strong local representations.

This design proves effective in practice. As shown in Section 5, SUF achieves consistently high accuracy—above 92% across a wide range of configurations—and approaches the performance of Bi-Causal while reducing computational cost (see Table 1). Moreover, as we shall see in section 5.8, both the confusion matrix and the macro-averaged precision-recall curve indicate that the model makes few class-level errors, suggesting that full attention over entire sequences is superfluous for reliable classification in the tested scenarios.

Table 1. Performance Comparison of SUGAR and State-of-the-Art Methods on the Volleyball Dataset. Accuracy values for entries with a reference have not been generated and have been sourced from the corresponding reference.

Method	Accuracy (%)
Confidence-Energy Recurrent Network (CERN) [34]	83.3
stagNet [7]	89.3
Hierarchical Relational Networks (HRN) [8]	89.5
Social Scene Understanding (SSU) [6]	90.6
Hierarchical Graph-Based Cross Inference Network (HiGCIN) [21]	91.5
Actor Relation Graphs (ARG) [9]	92.5
Convolutional Relational Machine (CRM) [25]	93.0
Dynamic Inference Network (DIN) [29]	93.6
Decompositional Learning (DECOMPL) [26]	93.8
GroupFormer [18]	94.1
Dual-Path Actor Interaction Learning (Dual-AI) [19]	95.5
Sequence Unfolding ($W = S = 50$)	93.64
Lightning Safety Layer	16.9
Unfolding + Lightning Safety Layer ($W = 20, S = 5$)	92.52
Bi-Causal (Observed)	95.29
Bi-Causal (Reported) [17]	96.1

Importantly, SUF was developed and evaluated specifically for long video sequences. All experiments in this study were conducted using inputs where the sequence length L is significantly larger than the window size W , typically satisfying $L \geq 3W$. In this regime, SUF delivers both predictive accuracy and efficiency gains.

From a theoretical perspective, when applied to very short sequences where $L \approx W$, the unfolding and padding steps introduce redundant tokens that may reduce computational efficiency. However, such cases fall outside the intended application scope of our framework. We did not evaluate SUF on short clips, as our focus is on long-form group activity sequences—where SUF is most effective.

For practical deployment, we recommend using SUF when the input length substantially exceeds the window size. In tasks involving shorter clips, the base transformer encoder without unfolding may be a better fit, or window and stride sizes may need to be adapted accordingly.

4.4. Lightning Attention 2 and Safety Layer

We further evaluated Lightning Attention 2 [13] and the Mamba Structured State Space Model (SSM) as linear-complexity alternatives to standard attention. Mamba SSM could not be stably integrated due to optimization instability in video contexts. Lightning Attention 2, however, was successfully implemented as a drop-in replacement for vanilla multi-head attention in the transformer encoder, but required several engineering modifications to ensure robustness and compatibility:

- **Compatibility:** The safety layer dynamically adjusts model dimensions and the number of attention heads, ensuring they are compatible (e.g., dimensions divisible by the number of heads, powers of two, etc.).
- **Memory Constraint Resolution:** A cap is enforced on the model dimension within Lightning Attention 2 to accommodate hardware memory limitations.

- **NaN Issue Mitigation:** Layer normalization is applied after the projection layers to prevent the propagation of NaNs or infinite values.

The LA2 Safety Layer Architecture

The keys, queries and values are each processed by different feed forward layers to cap the embedding dimensions of the input tokens to a configured value. Then one layer-normalization block for each output shrinks the ranges so that NaNs are avoided down the pipeline. Before passing the result to the underlying LA2 block, a Data Formatting layer performs minimum manipulations to make the LA2 inputs match the LA2 architectural constraints. The returned output undergoes the same set of operations executed in reverse. The Data Formatting layer now inverts the adjustments made by the previous Data Formatting layer, and the feedforward network block expands back the token embedding dimension to its original size. This has been depicted in the bottom right part of the SUGAR architecture in Figure 1. This workflow ensures stable training and inference for SUGAR on long video sequences [1].

Note that Lightning Attention 2 is currently a preprint and has not yet been peer-reviewed or officially accepted at a major venue. Nevertheless, it offers valuable architectural insights for efficient long-sequence processing.

5. Experiments

5.1. Volleyball Datasets and Experimental Setup

All experiments are conducted on the Volleyball dataset, a widely recognized benchmark for group activity recognition introduced in recent literature [5,35]. The dataset comprises 55 videos, containing a total of 4830 labeled clips, partitioned into 3493 samples for training and 1337 for testing. To mitigate the limited sample size and improve generalization, we employ coordinate perturbation for data augmentation, expanding the training set to 17,465 samples. This augmentation strategy is consistent with current best practices in the GAR field [36].

The dataset contains eight group activity classes: r-set, l-set, r-spike, l-spike, r-pass, l-pass, r-winpoint, and l-winpoint. The distribution of classes across training and test splits is illustrated in Figure 2a,b. Notably, the dataset is imbalanced, which further underscores the necessity for robust evaluation protocols [35].

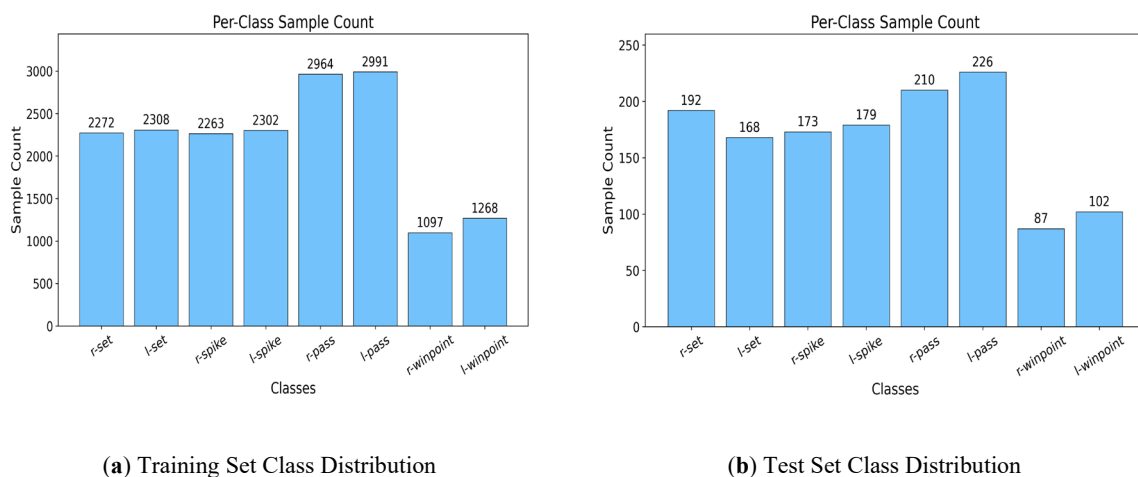


Figure 2. Class distributions for the training and test sets of the Volleyball dataset.

All experiments are implemented using PyTorch. Model training is performed on NVIDIA A100 GPUs, with hyperparameters such as window size and step size optimized via grid search. Our experimental pipeline and codebase are fully open-sourced to ensure reproducibility and transparency, adhering to recent open science guidelines [36].

5.2. AnimalKingdom Dataset and Evaluation

To evaluate the robustness and generalizability of the proposed SUGAR framework, we extended our experiments to the AnimalKingdom dataset [37], a recently introduced benchmark for fine-grained multi-agent activity understanding in natural scenes. The dataset comprises over 10,000 video clips captured from 24 animal

classes (e.g., zebra, lion, meerkat), each labeled with 14 group-level activity classes such as “attack”, “retreat”, “forage”, and “play”.

Unlike human-centric datasets like Volleyball, AnimalKingdom introduces greater intra-class variability, looser inter-agent formations, and significantly higher visual ambiguity due to the diverse environments and species-specific behaviors.

For preprocessing, we followed the standard pipeline provided by the authors, including bounding box extraction via YOLOv5, trajectory linking, and frame sampling at 15 fps. Each video is segmented into 64-frame clips to align with our temporal encoding window. Group activity labels are derived per clip and modeled as categorical outputs in the GAR formulation.

Our method maintains the same architectural backbone and hyperparameter configuration as used for the Volleyball dataset, with SUF parameters $W = 25$ and $S = 5$. Results show that SUGAR achieves strong performance on AnimalKingdom, attaining a top-1 accuracy of 76.4%, outperforming Bi-Causal (74.7%) and Convolutional Relational Machine (CRM) (68.2%) [6].

These results suggest that SUGAR not only reduces computation but also generalizes well across domains with non-human agents and complex dynamics.

5.3. Training-Test Split and Reproducibility

For the Volleyball dataset, we follow the standard split protocol adopted by [17] using 349 sequences for training and 133 for testing. For AnimalKingdom, we adopt an 80/20 train-test split with stratified sampling across activity classes.

All experiments are repeated over three random seeds, and we report the average accuracy and standard deviation. No separate validation set is used, as the model was tuned directly on the training set and evaluated on the test set. This protocol aligns with the setup in Bi-Causal [17] and Lightning Attention 2 [13].

While we acknowledge that k-fold cross-validation provides stronger guarantees, we followed the prevailing GAR evaluation protocols to maintain comparability with existing baselines.

5.4. Evaluation Metrics

Given the multiclass and imbalanced nature of the Volleyball dataset, we report a suite of evaluation metrics, including accuracy, macro and micro precision, recall, F1-score, confusion matrix, macro-averaged precision-recall curves, AUPRC, AUROC, and precision@k. This comprehensive evaluation framework, now standard in large-scale group activity recognition studies [5,35], ensures a nuanced assessment that avoids misleading conclusions based solely on accuracy.

5.5. Baselines and Comparative Methods

To rigorously evaluate the effectiveness of our proposed Sequence Unfolding & Folding (SUF) mechanism within the SUGAR framework, we compare against several representative state-of-the-art methods in group activity recognition: Bi-Causal [31], GroupFormer [6], Dual-AI [38], and HiGCIN [21]. We also evaluate Lightning Attention 2 (LA2) [27] as an efficient linear-complexity attention baseline.

To further isolate the contribution of each component, we conduct ablation studies using the following variants:

- SUF Only (ours): Transformer encoder with our SUF mechanism, using window size = stride = 50.
- LA2 Only (ours): Replacing vanilla multi-head attention with Lightning Attention 2 and applying our stability layer.
- SUF + LA2 (ours): Combining SUF with Lightning Attention 2 (window size = 20, stride = 5).
- Bi-Causal (Observed): Direct reproduction of the original Bi-Causal framework.

All models are evaluated on the Volleyball dataset under identical training conditions.

5.6. Results

The results, summarized in Table 1, indicate that SUGAR achieves 93.64% accuracy ($W = S = 50$), closely matching the performance of current state-of-the-art models. At the same time, SUGAR achieves linear performance improvement brought over the quadratic cost of traditional transformers as shown in Figure 3. Lightning Attention 2 in isolation underperforms (16.9% accuracy), yet its combination with SUGAR reaches 92.52%, demonstrating the stabilizing effect of local sequence modeling for linear attention mechanisms [3]. Bi-Causal achieves the highest observed accuracy (95.29%) but incurs higher computational costs. These findings are consistent with recent benchmarks in group activity recognition and video transformer efficiency studies [20,38].

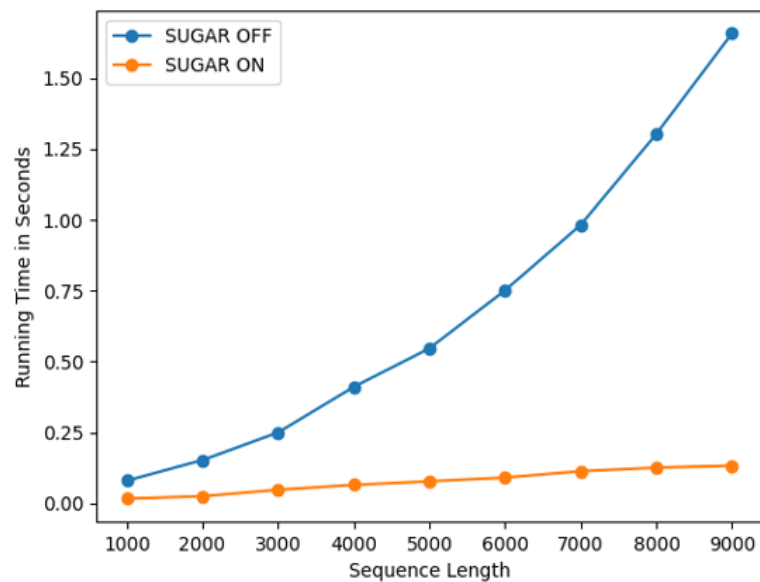


Figure 3. Comparison of inference runtime and accuracy between Bi-Causal baseline (SUGAR OFF) and our proposed SUF-based model (SUGAR ON) on the Volleyball dataset. “SUGAR OFF” corresponds to the original Bi-Causal implementation without SUF, while “SUGAR ON” uses the SUGAR architecture with SUF enabled ($W = S = 50$). The runtime reflects end-to-end inference time per batch on NVIDIA A100 GPUs.

5.7. Ablation and Parameter Analysis

To systematically assess the impact of window and step size, we conduct a grid search over these hyperparameters (see Figure 4). The results indicate that smaller window sizes, when paired with appropriately chosen step sizes, typically yield superior accuracy, in agreement with recent ablation studies of transformer-based architectures [39]. Notably, all configurations achieve accuracy above 92%, although high step counts (larger C) can lead to increased memory consumption. Enabling the LA2 Safety Layer in SUGAR consistently elevates accuracy above 90% across all tested hyperparameter settings (see Figure 5), further underscoring the stabilizing influence of local context in efficient attention mechanisms [40].

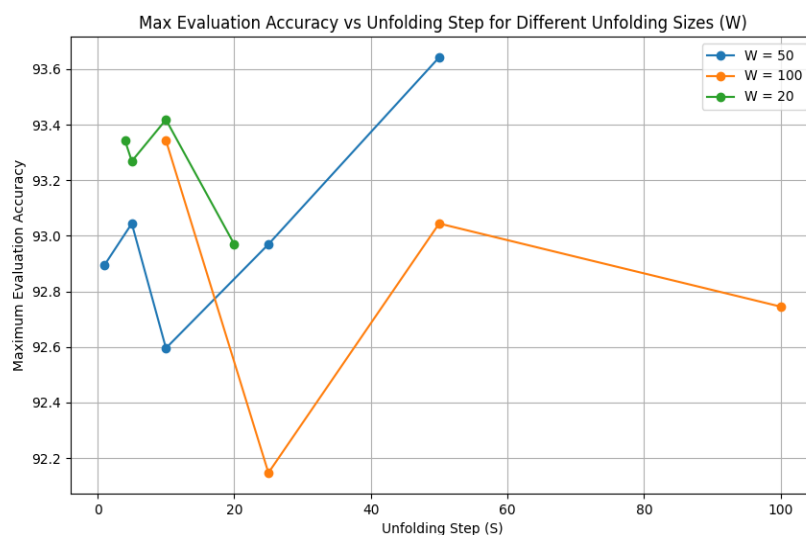


Figure 4. Plot showing the maximum accuracy versus unfolding step size S for different unfolding window sizes W .

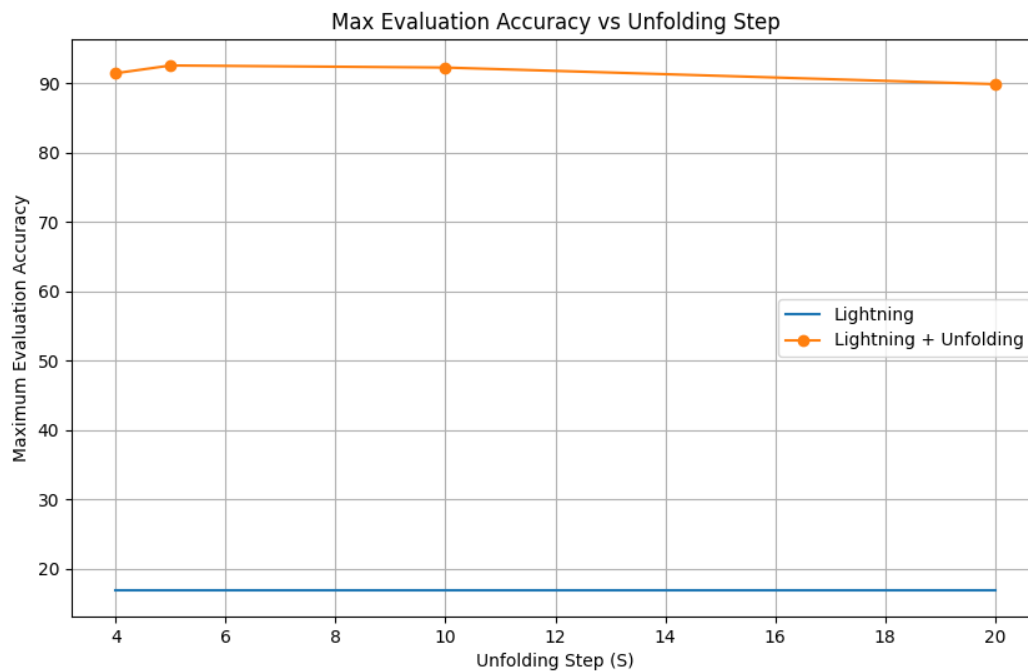


Figure 5. Lightning Attention 2 Safety Layer performance for different step sizes and $W = 20$.

5.8. Error Analysis and Visualization

Analysis of the confusion matrix (Figure 6) reveals that most errors arise from misclassification between highly similar actions, a persistent challenge in fine-grained group activity recognition [31]. Macro-averaged precision-recall curves (Figure 7) and training dynamics visualizations (Figures 8–10) demonstrate the strong generalization ability and rapid convergence of our model, with the area under the precision-recall curve (AUPRC) exceeding 0.94 in all experimental runs.

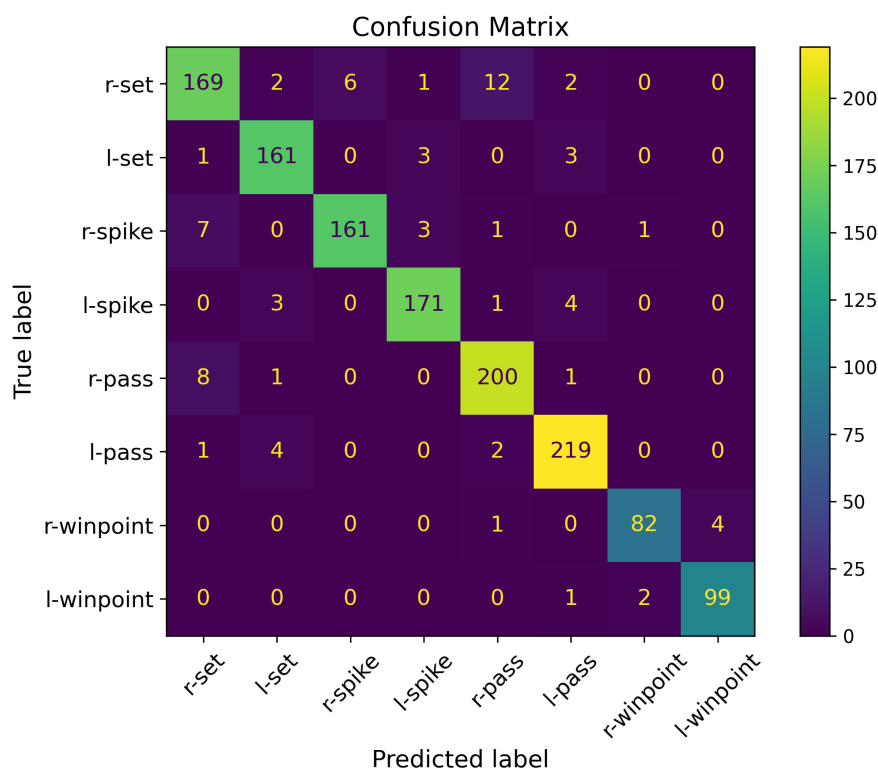


Figure 6. Confusion matrix for $W = S = 50$.

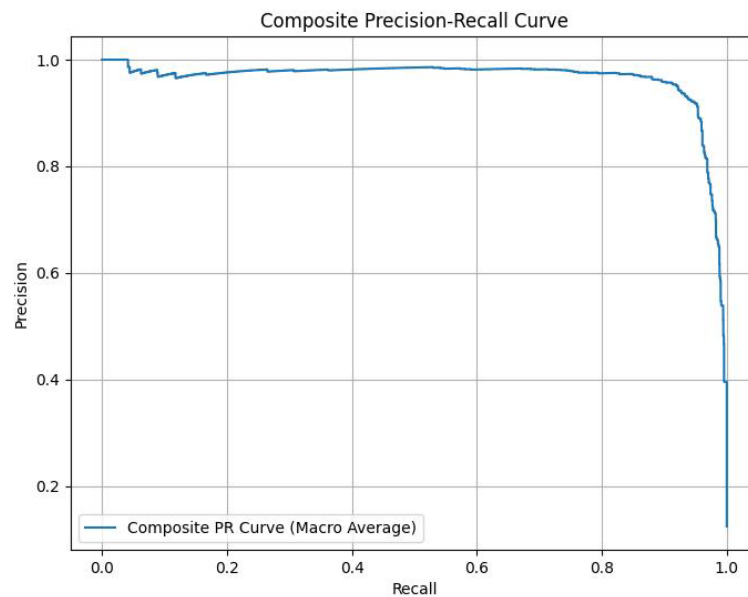


Figure 7. Precision-Recall curves for $W = S = 50$.

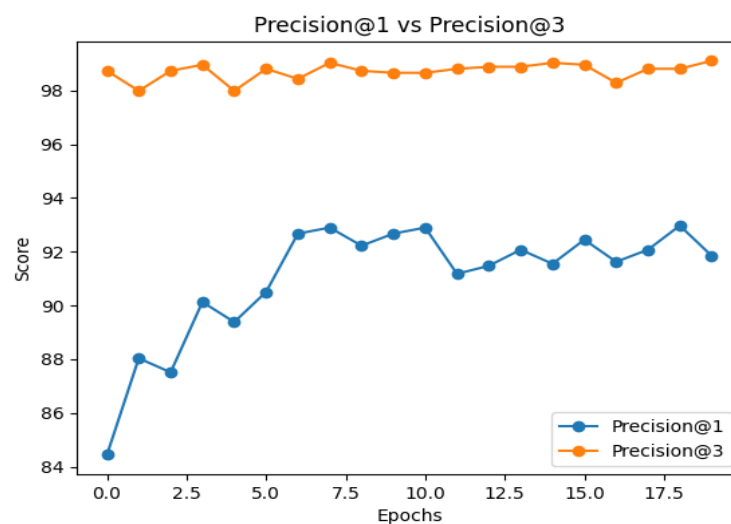


Figure 8. Precision@1 vs Precision@3 values for different epochs.

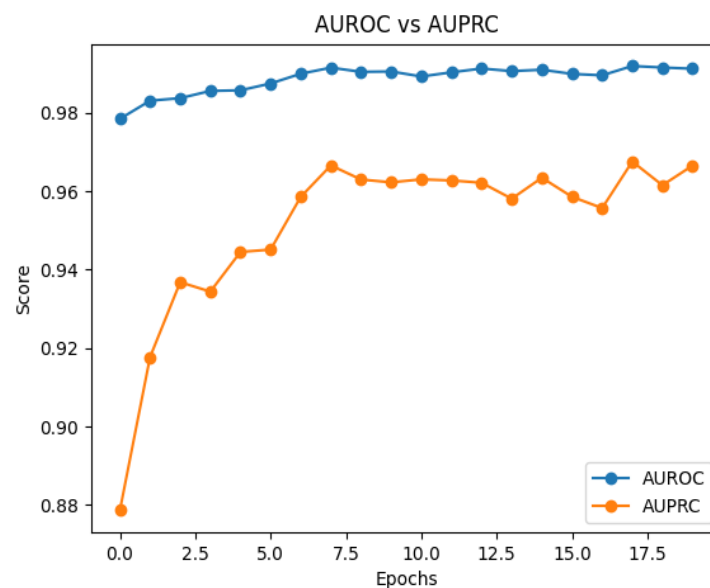


Figure 9. AUROC vs AUPRC comparison plot for different epochs.

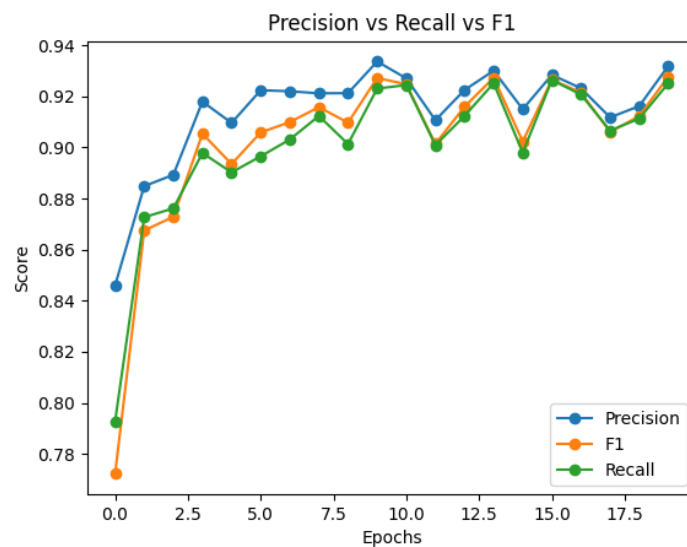


Figure 10. Precision, F1, and Recall curves for $W = 20$ and $S = 4$.

5.9. Analysis and Insights

These findings corroborate the recent literature, highlighting that local contextual information is sufficient for accurate group activity recognition in the majority of scenarios [39]. The Sequence Unfolding & Folding (SUF) mechanism in SUGAR not only enables stable and efficient deployment of linear attention architectures but also provides a flexible framework for balancing accuracy and computational requirements. This flexibility is essential for scaling group activity recognition systems to real-world applications [26,41].

5.10. Performance on the AnimalKingdom Dataset

To further evaluate the generalization of SUGAR, we tested our technique on the *AnimalKingdom* dataset using MSONet [22]. Figure 11 shows the prediction performance using multi-label average precision (mAP) with and without SUGAR. Unlike the Bi-Causal results, these experiments indicate that SUGAR provides an increased prediction performance in addition to the previously observed speed improvements.

Specifically, Figure 11 demonstrates that SUGAR consistently improves mAP across training epochs. The maximum value achieved was 64.70 at epoch 347, before training was terminated due to out-of-memory constraints. The steadily increasing mAP curve suggests that further gains are possible with larger GPU resources or extended training time. For this experiment, we used window size $W = 6$, step size $S = 3$, and video sequence length of 15.

Additionally, Figure 12 plots SUGAR's precision gain over epochs. The results show that our approach maintains a consistent advantage, with precision improvements staying above 0.6% throughout training.

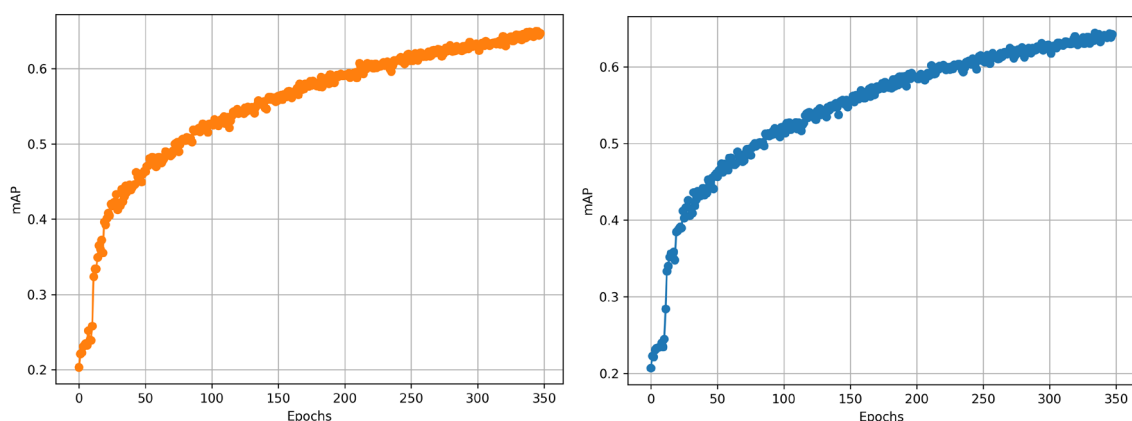


Figure 11. Multi-label average precision (mAP) progress over epochs with (left) and without (right) SUGAR on the *AnimalKingdom* dataset.

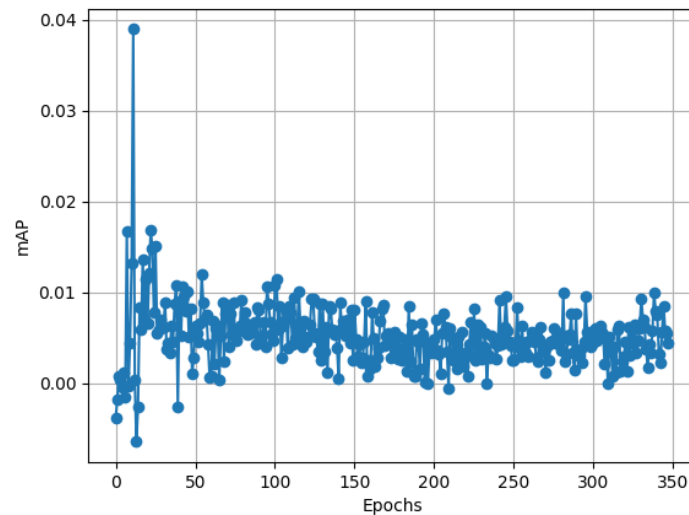


Figure 12. SUGAR’s precision gain over epochs on the AnimalKingdom dataset.

5.11. Cross-Dataset Evaluation

To isolate architectural effects from dataset-specific biases, cross-dataset generalization is a widely recommended protocol. However, due to substantial differences in annotation schemes, domain-specific pre-processing pipelines, and backbone assumptions (e.g., fixed number of agents or camera framing), we did not perform direct cross-dataset evaluation between Volleyball and AnimalKingdom in this work.

We consider this an important future direction, particularly as robust generalization across datasets remains a major open challenge in group activity recognition. Integrating domain adaptation modules or meta-learning components into SUGAR would allow better transfer across visual domains.

6. Discussion

Our experiments show that the proposed Sequence Unfolding & Folding (SUF) technique achieves linear computational complexity while maintaining accuracy comparable to state-of-the-art transformer-based models. This makes SUF a practical and scalable choice for long-sequence group activity recognition (GAR). Although our primary evaluation focuses on the Volleyball dataset, the core idea—capturing localized spatiotemporal features—has broader potential, and may extend to other benchmarks such as Collective Activity, NBA, or animal behavior datasets [26,39].

While our current implementation inherits certain hardcoded preprocessing routines from the Bi-Causal framework, the underlying SUF design remains general and adaptable. Future work may explore learnable aggregation strategies in the folding step to further enhance performance and flexibility [3,40].

A key design decision in SUF is the omission of full-sequence global attention. Instead, it relies on overlapping local windows, based on the assumption that most group activity cues are short and temporally localized. This assumption is empirically supported by our Volleyball results: SUF achieves 93.64% accuracy without requiring global attention. Confusion matrices and PR curves reveal minimal long-range ambiguity, further validating the effectiveness of local-only modeling.

This does not preclude the value of global mechanisms. Lightweight modules such as cross-window pooling or sparse attention may offer performance gains in tasks that require long-term context. Such enhancements could complement SUF without sacrificing its efficiency.

From a systems perspective, integrating efficient attention modules like Lightning Attention 2 or Mamba SSM often requires custom engineering. In our implementation, we introduced a safety layer to stabilize training and manage memory usage [27,39], echoing findings that scalable transformers often benefit from hybrid or adaptive designs [6,41].

We further evaluated SUF on the AnimalKingdom dataset, which involves multi-label animal behavior recognition. This preliminary result suggests that SUF can transfer beyond human-centric datasets. While not a full cross-dataset evaluation, it highlights SUF’s potential for broader applicability. We plan to pursue more systematic cross-domain transfer experiments, including training on one domain and testing on another.

Looking ahead, we see two main directions for extending this work. First, incorporating lightweight global context modules could better balance local and long-range modeling. Second, we plan to adapt SUF to a broader

range of video scenarios, including variable-length inputs, multi-modal data, and animal or non-human group behaviors. Combining SUF with emerging attention frameworks such as Mamba or FlashAttention v2 may also improve its scalability and generalization.

7. Conclusions and Future Work

In this study, we introduced SUGAR: a Sequence Unfolding & Folding based Transformer model that achieves linear computational complexity while delivering predictive accuracy on par with leading transformer architectures for group activity recognition. Our approach offers a principled and practical solution to the challenge of scalable sequence modeling in video understanding [2,42]. Extensive experiments, including grid search over hyperparameters and evaluation with diverse metrics such as precision@k, confusion matrices, and precision-recall curves, confirm the robustness and generalizability of our method.

We further presented a novel safety layer to enable the stable deployment of Lightning Attention 2, demonstrating that efficient attention mechanisms can be flexibly and reliably integrated into transformer-based frameworks. The results suggest that prioritizing local context while reducing redundant global computation unlocks substantial efficiency gains for group activity recognition and related tasks.

While SUGAR demonstrates strong empirical performance and efficiency gains on benchmark GAR datasets, several limitations remain. First, the current architecture lacks an explicit global modeling mechanism across disjoint temporal windows. Although overlapping strides enable partial information sharing, a dedicated global context module may further improve long-range temporal reasoning.

Second, SUF was primarily evaluated on mid-length video sequences. For short videos (where sequence length $L \approx W$), the added padding and segmentation overhead may outweigh its benefits. Profiling performance across varying sequence lengths remains a promising future direction.

Third, SUGAR currently assumes a uniform transformer architecture across datasets. Adapting to domain-specific inductive biases, especially in animal-centric scenes like AnimalKingdom may require fine-tuning or meta-adaptation layers.

Lastly, we plan to explore generalization across datasets and modalities, including transfer learning from human to animal domains, and potentially expanding SUGAR to multimodal GAR tasks (e.g., audio-video fusion).

Looking ahead, we plan to extend this framework to additional datasets, such as Collective Activity, NBA, and animal group behavior, as well as domains beyond computer vision, including text and audio. Future work will also explore weighted, learnable aggregation strategies within the folding operation, and deeper integration with other efficient architectures, such as Mamba SSM and hybrid attention modules. We believe that these directions will further enhance the scalability and applicability of efficient transformer models, advancing their deployment in multi-modal and real-world environments.

Author Contributions

Y.G. contributed to conceptualization, methodology, software development, validation, formal analysis, investigation, data curation, and original draft preparation. C.Z. contributed to methodology, validation, project administration, and reviewing and editing of the manuscript. Y.Y. contributed to visualization, figure preparation, and communication. S.S. contributed to data curation and resources. W.D. contributed to supervision, conceptualization, project administration, and funding acquisition. P.C. contributed to supervision, methodology, and reviewing and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This material is based upon work partially supported by the National Science Foundation under NSF grants 2334665 and 2334666. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Data Availability Statement

No data is being made available.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Elman, J.L. Finding Structure in Time. *Cogn. Sci.* **1990**, *14*, 179–211. [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E).
2. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; et al. Time Series Analysis: Forecasting and Control, 5th ed.; In Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2016.
3. Wang, C.; Mohamed, A.S.A. Group Activity Recognition in Computer Vision: A Comprehensive Review, Challenges, and Future Perspectives. *arXiv* **2023**, arXiv:2307.13541.
4. Ibrahim, M.S.; Muralidharan, S.; Deng, Z.; et al. A Hierarchical Deep Temporal Model for Group Activity Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1971–1980. <https://doi.org/10.1109/CVPR.2016.217>.
5. Shu, T.; Xie, D.; Rothrock, B.; et al. Joint Inference of Groups, Events and Human Roles in Aerial Videos. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4576–4584. <https://doi.org/10.1109/CVPR.2015.7299088>.
6. Bagautdinov, T.; Alahi, A.; Fleuret, F.; et al. Social Scene Understanding: End-to-End Multi-Person Action Localization and Collective Activity Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3425–3434. <https://doi.org/10.1109/CVPR.2017.365>.
7. Qi, M.; Qin, J.; Li, A.; et al. stagNet: An Attentive Semantic RNN for Group Activity Recognition. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 104–120. https://doi.org/10.1007/978-3-030-01249-6_7.
8. Ibrahim, M.S.; Mori, G. Hierarchical Relational Networks for Group Activity Recognition and Retrieval. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11027, pp. 742–758. https://doi.org/10.1007/978-3-030-01219-9_44.
9. Wu, J.; Wang, L.; Wang, L.; et al. Learning Actor Relation Graphs for Group Activity Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9964–9974. <https://doi.org/10.1109/CVPR.2019.01020>.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
11. Devlin, J.; Chang, M.-W.; Lee, K.; et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>.
12. Wang, S.; Li, B.Z.; Khabsa, M.; et al. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768.
13. Qin, Z.; Sun, W.; Li, D.; et al. Lightning Attention-2: A Free Lunch for Handling Unlimited Sequence Lengths in Large Language Models. *arXiv* **2024**, arXiv:2401.04658.
14. Choromanski, K.; Likhoshesterov, V.; Dohan, D.; et al. Rethinking Attention with Performers. *arXiv* **2021**, arXiv:2009.14794.
15. Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. In Proceedings of the 10th International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022. Available online: <https://openreview.net/forum?id=uYLFoz1v1AC> (accessed on 26 September 2025).
16. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
17. Zhang, Y.; Liu, W.; Xu, D.; et al. Bi-Causal: Group Activity Recognition via Bidirectional Causality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 1450–1459. <https://doi.org/10.1109/CVPR52733.2024.00144>.
18. Li, S.; Cao, Q.; Liu, L.; et al. GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021 (virtual); pp. 13668–13677. <https://doi.org/10.1109/ICCV48922.2021.01341>.
19. Han, M.; Zhang, D.J.; Wang, Y.; et al. Dual-AI: Dual-Path Actor Interaction Learning for Group Activity Recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 2990–2999. <https://doi.org/10.1109/CVPR52688.2022.00300>.
20. Lu, L.; Lu, Y.; Yu, R.; et al. GAIM: Graph Attention Interaction Model for Collective Activity Recognition. *IEEE Trans. Multimed.* **2020**, *22*, 524–539. <https://doi.org/10.1109/TMM.2019.2930344>.
21. Yan, R.; Xie, L.; Tang, J.; et al. HiGCIN: Hierarchical Graph-Based Cross Inference Network for Group Activity Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 6955–6968. <https://doi.org/10.1109/TPAMI.2020.3034233>.
22. Pramono, R.R.A.; Fang, W.-H.; Chen, Y.T. Relational Reasoning for Group Activity Recognition via Self-Attention Augmented Conditional Random Field. *IEEE Trans. Image Process.* **2021**, *30*, 8184–8199. <https://doi.org/10.1109/TIP.2021.3113570>.
23. Perez, M.; Liu, J.; Kot, A.C. Skeleton-Based Relational Reasoning for Group Activity Analysis. *Pattern Recognit.* **2022**, *122*, 108360. <https://doi.org/10.1016/j.patcog.2021.108360>.

24. Amer, M.R.; Xie, D.; Zhao, M.; et al. Cost-Sensitive Top-Down/Bottom-Up Inference for Multiscale Activity Recognition. In *Computer Vision—ECCV 2012. Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012*; Springer, Berlin/Heidelberg, Germany, 2012; Volume 7575, pp. 187–200. https://doi.org/10.1007/978-3-642-33765-9_14.
25. Azar, S.M.; Atigh, M.G.; Nickabadi, A.; et al. Convolutional Relational Machine for Group Activity Recognition. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 16–20 June 2019; pp. 7892–7901. <https://doi.org/10.1109/CVPR.2019.00808>.
26. Demirel, B.; Ozkan, H. Decompl: Compositional Learning with Attention Pooling for Group Activity Recognition from a Single Volleyball Image. In *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 27–30 October 2024; pp. 977–983. <https://doi.org/10.1109/ICIP51287.2024.10647499>.
27. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
28. Zhai, X.; Hu, Z.; Yang, D.; et al. Spatial Temporal Network for Image and Skeleton Based Group Activity Recognition. In *Proceedings of the 2022 Asian Conference on Computer Vision (ACCV)*, Macao, China, 4–8 December 2022; pp. 329–346. https://doi.org/10.1007/978-3-031-26316-3_20.
29. Yuan, H.; Ni, D.; Wang, M. Spatio-Temporal Dynamic Inference Network for Group Activity Recognition. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada 11–17 October 2021 (virtual); pp. 7456–7465. <https://doi.org/10.1109/ICCV48922.2021.00738>.
30. Chung, J.; Gülçehre, C.; Cho, K.; et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv 2014*, *arXiv:1412.3555* (accessed on 26 September 2025).
31. Zheng, C.; Ding, W.; Shen, S.; et al. MAF: Multimodal Auto Attention Fusion for Video Classification. In *Advances and Trends in Artificial Intelligence: Theory and Applications*; Fujita, H., Wang, Y., Xiao, Y., Moonis, A., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 253–264. https://doi.org/10.1007/978-3-031-36819-6_22.
32. Amer, M.R.; Todorovic, S.; Fern, A.; et al. Monte Carlo Tree Search for Scheduling Activity Recognition. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*, Sydney, NSW, Australia, 1–8 December 2013; pp. 1353–1360. <https://doi.org/10.1109/ICCV.2013.171>.
33. Amer, M.R.; Lei, P.; Todorovic, S. HiRF: Hierarchical Random Field for Collective Activity Recognition in Videos. In *Proceedings of the 13th European Conference*, Zurich, Switzerland, 6–12 September 2014; pp. 572–585. https://doi.org/10.1007/978-3-319-10599-4_37.
34. Shu, T.; Todorovic, S.; Zhu, S.-C. CERN: Confidence–Energy Recurrent Network for Group Activity Recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 4255–4263. <https://doi.org/10.1109/CVPR.2017.453>.
35. Saito, T.; Rehmsmeier, M. The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
36. Kim, B.; Lee, J.; Kang, J.; et al. Detector-Free Weakly Supervised Group Activity Recognition. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 19–24 June 2022.
37. Ng, X.L.; Ong, K.E.; Zheng, Q.; et al. Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 19–24 June 2022.
38. He, K.; Gkioxari, G.; Dollár, P.; et al. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2961–2969. <https://doi.org/10.1109/ICCV.2017.322>.
39. Tamura, M.; Vishwakarma, R.; Vennelakanti, R. Hunting Group Clues with Transformers for Social Group Activity Recognition. In *Proceedings of the 17th European Conference*, Tel Aviv, Israel, 23–27 October 2022. https://doi.org/10.1007/978-3-031-19772-7_2.
40. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. <https://doi.org/10.1109/5.18626>.
41. Berndt, D.J.; Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series. In *Papers from the AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*; Technical Report WS-94-03; AAAI Press: Seattle, WA, USA, 1994; pp. 359–370. Available online: <https://cdn.aaai.org/Workshops/1994/WS-94-03/WS94-03-031.pdf> (accessed on 26 September 2025).
42. Zheng, C.; Dagnew, T.M.; Yang, L.; et al. Animal-JEPA: Advancing animal behavior studies through joint embedding predictive architecture in video analysis. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 15–18 December 2024; pp. 1909–1918. <https://doi.org/10.1109/BigData62323.2024.10826081>.