2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Causal Transformer-Based Online Action Recognition for High-Level Control of a Unitree Go1 Robot

Chaitanya Bandi*, Kristof Kitz and Ulrike Thomas

***Abstract*— We present a new causal transformer system consisting of Spatial-Attention Tokenization (SAT) with Multi-Resolution Causal Temporal Mixing (MRCTM) to perform online skeleton-based action recognition during human–robot interaction. The novel architecture uses Spatial-Attention Tokenization (SAT) to generate soft tokens from human joint groups. MRCTM performs causal convolutions and self-attention operations to detect both detailed motion patterns and extended temporal relationships. We introduce GoHAR-12 dataset as an evaluation tool as it contains 12 gesture and posture classes which are recorded in human-robot interaction (HRI) settings and directly translate to high-level commands for the Unitree Go1 quadruped. The proposed model reaches 98.4% accuracy on the GoHAR-12 dataset and it shows superior performance in distinguishing between actions that are quite similar in motion, maintains strong results on public benchmarks such as NTU-RGB+D and NW-UCLA. We demonstrate how causal transformer performs for reliable real-time skeleton-based control of the Unitree Go1 robot.**

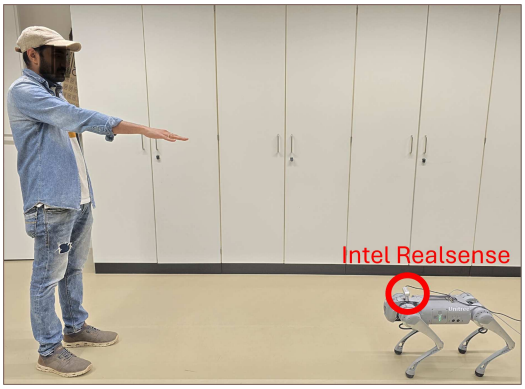

Fig. 1. Illustration of human–robot interaction scenario: a participant executes predefined actions while engaging with the Unitree Go1 robot, forming part of the dataset collection for action recognition and HRI analysis.

I. INTRODUCTION

Human action recognition stands as a fundamental capability which enables robots to interact with humans through natural and intuitive ways. The interpretation of human skeleton sequences through gesture and posture recognition delivers a strong and understandable way for robots to execute seamless responses during collaborative work. The combination of inexpensive RGB+D sensors with human pose estimation progress has made skeleton-based action recognition an operational solution for real-time human-robot interaction systems [1], [2], [3].

The initial skeleton-based recognition methods use recurrent architectures including LSTMs and GRUs [4], [5] which processed sequential data well yet failed to handle extended temporal patterns and real-time processing requirements. The implementation of spatial–temporal graph convolutional networks (ST-GCN) by Yan et al. [1] and their following developments [2], [3], [6], [7] brought substantial performance growth through the utilization of human skeleton structural connections. The latest research demonstrates that transformer-based models [8], [9] together with hybrid models that combine convolutional and attention mechanisms [10], [11] achieve top performance on NTU-RGB+D [12] and NW-UCLA [13] datasets. The models demonstrate how attention mechanisms enable the detection of intricate spatio-temporal patterns and enable action recognition across multiple viewpoints and various actions.

All authors are with Department of Robotics and Human Machine Interaction Lab, Technical university of Chemnitz, Germany Ulrike.thomas@etit.tu-chemnitz.de

The current methods for skeleton-based action recognition primarily work on benchmark datasets without solving the specific requirements for robot control of quadruped robots. The natural operation of legged robots such as Unitree Go1 [14] demands precise gesture recognition during real-time operations because small action variations between Pointing and Waiting affect both safety and task completion. The majority of existing research on skeleton-based recognition has not tested their methods in closed-loop control systems that convert recognition outputs into robot locomotion commands.

Our research presents an innovative system for gesture recognition from skeleton data which focuses on robot control operations. Our method implements a causal transformer system that incorporates two essential components: (i) Spatial-Attention Tokenization (SAT) which creates flexible joint groupings to focus on meaningful body parts including arms, torso, hands, legs and (ii) Multi-Resolution Causal Temporal Mixing (MRCTM) which unites simple causal convolutions with self-attention to detect detailed movements across time sequences while maintaining real-time causality. The combined system produces an efficient yet powerful model for real-time action detection.

The GoHAR-12 dataset introduced in this research is a new collection of 12 gesture and posture classes used to map human actions to high-level Unitree Go1 commands. The dataset contains fundamental actions for quadruped control including Waving, Follow Me, Come Here, and so on. The dataset provides a wide range of control primitives for human-robot interaction with quadruped robots. Fig. 1 shows

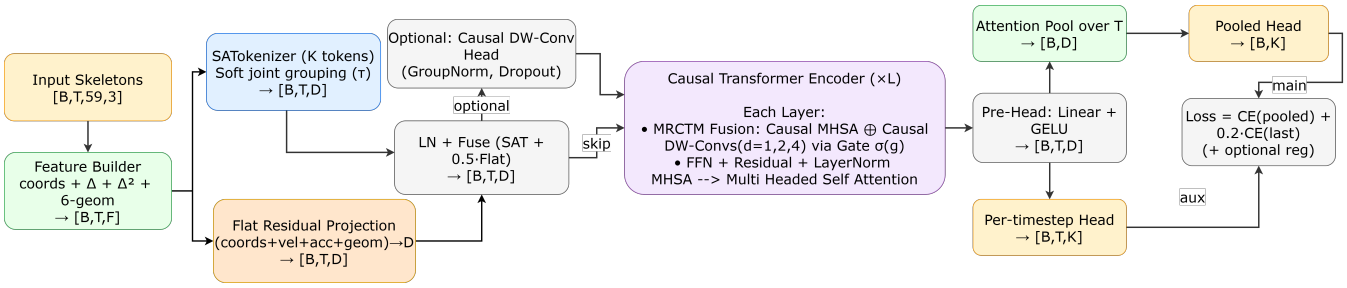

Fig. 2. The proposed system architecture where the 3D skeleton sequence including coordinates and their velocity and acceleration values and gesture-related information between joints are processed. The Spatial Attention Tokenization (SAT) module transforms joint features into a minimal set of learnable soft tokens which maintain essential spatial information while eliminating redundant data. The MRCTM block uses parallel attention-based and convolutional branches to extract both local time-dependent patterns and global contextual relationships through a gating mechanism.The encoded sequence passes through multiple Causal Transformer Encoder layers which maintain temporal order while achieving efficient long-term sequence processing.The classification head generates per-frame output logits which enable real-time action detection.

the illustration of human interacting with Unitree Go1. Our model reaches a performance level of 98.4% accuracy when tested on this dataset which surpasses all baseline causal TCN and transformer variants. The proposed method shows comparable performance on NTU-RGB+D 60/120 [12], [15] and NW-UCLA [13] benchmark datasets which proves its ability to work effectively in different domains.

Our research presents three main contributions to the field.

- The proposed causal transformer system uses SAT and MRCTM to achieve robust online skeleton-based action recognition.
- The GoHAR-12 dataset serves as a new dataset for human gesture recognition in quadruped robot control which contains 12 gesture/posture classes that correspond to high-level Go1 commands.
- The research presents thorough assessments of our system through both the GoHAR-12 dataset and public benchmarks to show its peak performance and real-time human-robot interaction capabilities.

II. RELATED WORKS

In this section, we discuss the most relevant works on skeleton-based action recognition on the NTU-RGB+D 60/120, NW-UCLA datasets using graph convolution networks, transformers and online skeleton action recognition models.

A. Skeleton-Based Action Recognition

The field of skeleton-based recognition has experienced rapid development through the transition from recurrent models to graph and transformer architectures. The Spatial–temporal GCNs brought topology-aware joint and bone reasoning to the forefront through their strong baseline performance and established training protocols which researchers widely adopted [1]. The CTR-GCN model achieved better adaptability through its channel-wise topology refinement method which learned separate graph structures for each channel to achieve high performance on NTU RGB+D 60/120 [12], [15] datasets [16]. The InfoGCN and its online-oriented successor InfoGCN++ focus on information preservation and predictive supervision to achieve competitive results on NTU60 and NTU120 and NW-UCLA datasets [17], [18]. The current state-of-the-art models achieve 93-98% top-1 accuracy on NTU60 Cross-Subject and 90-92% accuracy on NTU120 Cross-Subject across GCN transformer and hybrid model architectures [8], [10], [11]. The current state-of-the-art skeleton and RGB+pose pipelines for NW-UCLA reach or surpass 97% accuracy in standard evaluation settings [19], [10], [11]. The latest surveys in the field analyze three main trends which include transformer-based semantic enhancement and multi-representational fusion and robustness against view variations and noisy data [20], [21].

B. Transformer-Style and Hybrid Models

The global attention mechanism in Transformer-based skeleton models enables them to detect distant patterns while maintaining efficient processing for brief online time intervals. The research presents two categories of pure-transformer and hybrid conv-attention models which focus on skeleton data processing and temporal sequence analysis and tokenization methods according to the study in [20], [8], [22]. The latest transformer models achieve state-of-the-art results or near-state-of-the-art performance on NTU60/120 and NW-UCLA datasets while providing better scalability according to research findings in [23]. The work [11], propose autoregressive hypergraph generation with adaptive hyperedge learning to create a novel transformer–hypergraph hybrid for skeleton-based action recognition. The model achieves better long-range dependency and higher-order correlation detection through vector-quantized priors and joint supervised–unsupervised training. The experimental results on NTU RGB+D, NTU RGB+D 120 and NW-UCLA datasets demonstrate that AutoregAd-HGformer achieves superior performance than current hypergraph-based methods.

C. Online Recognition

The requirement for HRI and safety-critical applications demands both causal (future-free) inference and early prediction capabilities. The InfoGCN++ model adds online supervision and causal masks to its training process which leads to better early action prediction results on NTU60/120 and NW-UCLA datasets [18]. The evaluation of online/streaming protocols requires separate benchmarks which differentiate

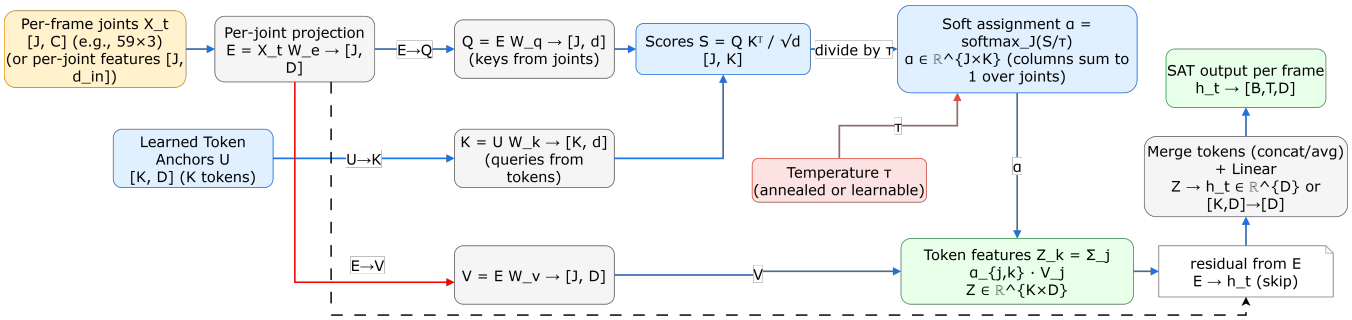

Fig. 3. Spatial Attention Tokenization (SAT) module. The model receives per-joint embeddings from each frame before it uses a soft partitioning method to divide joints into K tokens.The attention weights α receive joint normalization across all body parts to produce soft joint-token assignments. The temperature parameter τ determines how soft or hard the assignment process will be because high values produce smooth attention distribution and low values result in distinct partitions. The causal transformer encoder receives tokenized embeddings from each frame which create a condensed and meaningful representation.

them from traditional offline assessment methods according to research in [24]. The research now focuses on developing methods to handle noisy and sparse streaming skeleton data according to [25].

The number of Human-Robot Interaction systems that use skeleton and hand-gesture recognition continues to expand. The general HRC frameworks achieve reliable body-action and gesture recognition which enables collaborative task execution according to research by Terreran et al. in 2023 [26]. There exist not many works, that demonstrate the online skeleton-based action recognition in combination with controlling the Unitree Go1 robot.

D. Research Positioning

The design we present addresses the common deployment issues of offline-optimized models in robot perception systems by preventing instability and overfitting of convolutional paths and reducing sensitivity to small gesture variations. The model achieves transformer-level generalization through staged training which includes gate bias learning, conv freezing, GroupNorm application in the conv branch, and SAT temperature annealing to enhance online stability on to GoHAR-12 and public dataset results.

III. METHODOLOGY

The main objective of our work involves creating reliable online skeleton-based action recognition systems for human-quadruped interaction. The proposed model uses a causal transformer backbone structure which includes two new components: (i) Spatial-Attention Tokenization (SAT) for adaptive joint feature grouping into meaningful tokens and (ii) Multi-Resolution Causal Temporal Mixing (MRCTM) for combining causal convolutions with self-attention at different dilation rates to detect short-term and extended temporal patterns. The skeleton recognition architecture is illustrated in Fig. 2.

A. Input Representation

We operate on 3D skeleton sequences with $J = 59$ joints and Cartesian coordinates (x, y, z). Each sequence has temporal length T, producing an input tensor of shape $[B, T, J, 3]$ for a batch size B.

To enrich the representation, we compute first- and second-order temporal derivatives (velocity and acceleration) and pairwise geometric features between hands, elbows, shoulders, and torso (six distances and projections). The resulting feature vector per frame has dimension:

$$F = (J \times 3) \times 3 + 6 = 531$$

Thus the input to the network is $[B, T, F]$.

B. Spatial-Attention Tokenization (SAT)

The traditional skeleton-based action recognition methods send all joints directly to temporal encoders which results in redundant and noisy information from joints that do not provide much value. The proposed Spatial Attention Tokenization (SAT) system reduces the full joint representation into a few soft tokens which represent meaningful body regions. The model uses temperature-controlled soft assignments in its lightweight attention mechanism to determine which joints contribute to each token while preserving differentiability. The SAT module is clearly illustrated with similar notation as text in Fig. 3.

The per-joint features at frame t after linear embedding are denoted as $X_t \in \mathbb{R}^{J \times d}$. The system includes M learnable token anchors which function as joint prototypes to generate queries. The system transforms each joint feature X_j into a query vector and each token anchor into a key vector.

$$q_j = W_q X_j, \quad k_m = W_k u_m, \quad v_j = W_v X_j,$$

where W_q, W_k, W_v are learned linear projections. The attention score between joint j and token m is given by

$$\alpha_{jm} = \frac{\exp\left(\frac{q_j \cdot k_m}{\tau}\right)}{\sum_{j'=1}^{J} \exp\left(\frac{q_{j'} \cdot k_m}{\tau}\right)}$$

where τ is a temperature parameter that controls the sharpness of the assignment. The training process begins with high τ values which produce one-hot token assignments before

transitioning to smoother joint-to-token distributions when τ reaches its lower values. The training process includes a gradual reduction of τ values to enhance stability.

Each token $Z_m \in \mathbb{R}^d$ is then obtained as a weighted sum of joint values:

$$Z_m = \sum_{j=1}^{J} \alpha_{jm} v_j$$

The causal transformer encoder receives the compact token set M which contains Z_1 through Z_M to represent the skeleton at frame t. The final residual pathway sends the pooled joint features together with the tokenized representation to maintain global information when tokens become lost.

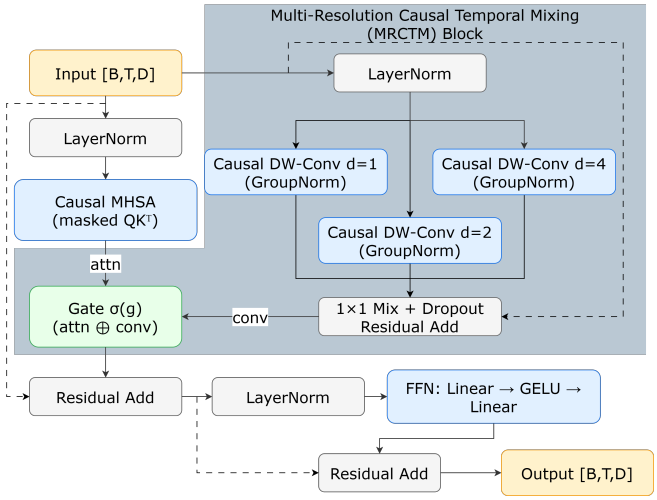

Fig. 4. Multi Resolution Causal Temporal Mixing module (MRCTM) with transformer encoder and gating. The gray region is MRCTM block with gating.

C. Multi-Resolution Causal Temporal Mixing (MRCTM)

The temporal encoder requires skeleton stream processing through causal methods which prevent it from accessing any future frames. The system implements two separate branches which work together to achieve its functionality. The system uses depthwise separable convolutions with dilations set to 1, 2, and 4 to detect various time-based patterns which range from small local movements to extensive motion patterns. The system implements standard transformer attention followed by a mask operation to stop information from leaking into future time steps for modeling distant relationships. The system uses an attention-biased gate to combine the outputs from both branches through learned weights. The output of this process is calculated as

$$h = g \odot h_{\text{conv}} + (1 - g) \odot h_{\text{attn}}, \quad g = \sigma(W[h_{\text{conv}}; h_{\text{attn}}])$$

The design achieves training stability through its combination of motion pattern inductive bias from the conv branch and global dependency detection from the attention branch. The MRCTM block in combination with causal multi headed attention transformer and gate functionality is presented in the Fig. 4.

D. Transformer Encoder

The model contains L encoder layers which perform MRCTM fusion followed by feed-forward sublayers and residual connections and layer normalization. The absolute positional embeddings function as an addition to maintain the correct sequence of frames.The encoder contains multiple causal transformer layers which maintain temporal order while processing both short-term and long-term joint relationships.

The input sequence undergoes LayerNorm [27] normalization at the start of each layer before the model applies two parallel processing paths. The first branch contains a causal multi-head self-attention (MHSA) module [28] which uses masked queries keys and values to prevent access to future frames. The first branch uses masked queries keys and values to detect long-range temporal connections between joints throughout the sequence. The second branch contains multiple depthwise separable 1D convolutional layers with different dilation rates (1, 2, 4) which operate as the MRCTM. The network uses these dilated convolutions to detect motion patterns at various time scales while maintaining causal processing. The attention output and convolutional output merge through a learnable gating system which controls their weighted combination. The output of the fusion process goes through a residual connection before another LayerNorm operation. The feed-forward network (FFN) applies two linear layers with GELU [29] activation and dropout followed by a residual connection [30]. The encoder processes streaming skeleton sequences through a combination of causal attention and dilated convolutions and gated fusion to detect both detailed local movements and distant relationships.

E. Classification Head

The system performs a linear transformation to obtain penultimate embeddings from each time step. The final classification head receives a pooled representation from the last timestep through mean pooling or last timestep pooling for its linear classification output to K action classes. The system includes an auxiliary classification module to identify between two specific classes (i.e., classes that closely resemble each other) through binary classification supervision.

F. Loss Function

The proposed SAT+MRCTM causal transformer is trained in a supervised fashion with a combination of classification and auxiliary regularization terms. The main objective is the **cross-entropy loss** [31] between the predicted logits and the ground-truth class label. Given a batch of B sequences with labels $\{y_i\}_{i=1}^{B}$ and model outputs $\{\hat{y}_i\}_{i=1}^{B}$, the cross-entropy is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{K} \mathbf{1}(y_i = c) \log \hat{y}_{i,c}$$

where K is the number of action classes and $\hat{y}_{i,c}$ is the predicted probability of class c for sample i.

In addition, SAT introduces a set of M learnable spatial tokens. To prevent trivial token collapse (all joints mapped

to a single token), we apply an **entropy regularization** term on the attention distributions:

$$\mathcal{L}_{\text{SAT}} = -\frac{1}{BTM} \sum_{i=1}^{B} \sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{j=1}^{J} \alpha_{ijm}(t) \log \alpha_{ijm}(t),$$

where $\alpha_{ijm}(t)$ is the attention weight of joint j for token m at time t. This term encourages a balanced utilization of tokens across joints.

For MRCTM, which fuses convolutional and transformer branches, we employ a **feature consistency loss** to stabilize the gating mechanism. Let h^{conv} and h^{attn} denote the intermediate features from the convolutional and transformer paths, respectively. A simple ℓ_2 penalty enforces alignment:

$$\mathcal{L}_{\text{MRCTM}} = \|h^{\text{conv}} - h^{\text{attn}}\|_2^2.$$

The final training loss is a weighted sum of the above objectives:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{SAT}} \mathcal{L}_{\text{SAT}} + \lambda_{\text{MRCTM}} \mathcal{L}_{\text{MRCTM}}$$

where λ_{SAT} and λ_{MRCTM} control the strength of auxiliary regularization. In our experiments, we set $\lambda_{\text{SAT}} = 0.01$ and $\lambda_{\text{MRCTM}} = 0.05$, which provided a good trade-off between classification accuracy and stable tokenization.

In addition to this, we also consider last value loss for online recognition case because it helps in stable output classes. This is applied for the final loss of GoHar-12 online training and evaluation. The final loss is a weighted sum of both terms:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{last}} + \mathcal{L}_{\text{pool}}$$

where λ balances stability and online responsiveness which is set to 0.2.

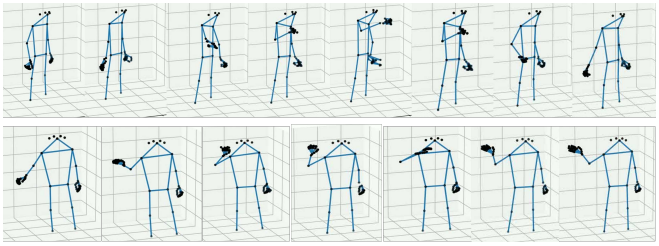

Fig. 5. Two samples from the GoHar-12 dataset. Top row belongs to Come_here class and bottom row is Waving.

IV. EXPERIMENTS

A. GoHAR-12 Dataset Collection

The GoHAR-12 dataset (Gesture-oriented Human Action Recognition, 12 classes) serves as our evaluation tool to assess online human-robot interaction performance. We introduce a new GoHAR-12 dataset for high level control of the Unitree Go1 quadruped robot. We record the RGB+D videos of 10 participants performing twelve robot-controllable actions using two realsense cameras. Once the dataset is captured, we process through the RTMW-3D pose [32] system which generated accurate 3D whole-body coco [33] joint data representation.

TABLE I

ACTION CLASSES IN THE GOHAR-12 DATASET. EACH ACTION CORRESPONDS TO A DISTINCT HUMAN COMMAND FOR CONTROLLING THE UNITREE GO1 QUADRUPED ROBOT.

ID	Action Class	No. of Seqs.
0	Waving	330
1	Stop	340
2	Come_here	330
3	Follow_me	330
4	Pointing_to_sit	340
5	Turn_right	330
6	Turn_left	330
7	Move_back	330
8	Stay_here	340
9	Rotate_clockwise	330
10	Rotate_anticlock	330
11	Idle_background	340

Then we manually label all recorded data to achieve high accuracy in their annotations. The dataset received multiple rounds of data augmentation through spatial jittering, temporal scaling, mirroring, random occlusion, random viewpoint rotation ($\pm 60°$), Gaussian noise application. The dataset contains diverse content which replicates natural human-robot interactions through different viewing angles and human physical characteristics and movement speeds. The dataset contains 4000 sequences with frames ranging from 60 to 150 in each clip which distribute evenly across the twelve action categories. From the dataset, we randomly split 785 sequences for evaluation and 3215 sequences for training. We plan to release GoHAR-12 as a public resource to enable additional studies about robot-oriented action recognition. The samples skeleton sequences of two different classes can be observed in Fig 5.

B. Standard Benchmark Evaluation

The model receives benchmark testing through evaluation on three prominent skeleton-based action recognition datasets which include NTU RGB+D 60 [12] and NTU RGB+D 120 [15] and NW-UCLA [13] Multiview Action3D.

The NTU RGB+D 60 dataset contains 60 actions from 40 subjects across 56,000+ sequences while following two evaluation protocols: X-Sub for participant separation and X-View for 120 dataset contains 120 action categories with 106 participants and more than 114,000 sequences while camera-1 testing against cameras 2 and 3 training. The NTU RGB+D 120 using X-Sub and X-Set evaluation protocols for testing and training. The NW-UCLA Multiview Action3D dataset contains 1494 sequences of 10 action classes which Kinect cameras record from three different viewpoints. The training process uses two camera views before testing with the third camera view under the common protocol.

C. Training Details

The AdamW optimizer [38] trains the model with 3×10^{-4} learning rate and 1×10^{-4} weight decay under cosine scheduling with warmup. The NTU-60/120 and GoHAR-12 datasets use batch sizes of 64 while NW-UCLA requires 32

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON NTU RGB+D 60, NTU RGB+D 120, AND NW-UCLA DATASETS.

Arch.	Method	NTU-60		NTU-120		NW-UCLA	Params
		X-Sub	X-View	X-Sub	X-Set	Acc.	Millions (M)
Graph Conv.	ST-GCN [1]	81.5	88.3	70.7	73.2	-	3.08
	2S-AGCN [2]	88.5	95.1	82.5	84.2	-	-
	Shift-GCN [34]	90.7	96.5	85.9	87.6	94.6	2.76
	Info-GCN [17]	93.0	97.1	89.8	91.2	97.0	6.28
	HD-GCN [19]	93.4	97.2	90.1	91.6	97.2	6.72
Transformer	ST-TR [8]	89.9	96.1	81.9	84.1	-	
	IIP-TR [35]	92.3	96.4	88.4	89.7	-	
	FG-STFormer [36]	92.6	96.7	89.0	90.6	97.0	-
	IGFormer [37]	93.4	96.5	85.4	86.5	-	-
	Hyperformer [10]	92.9	96.5	89.9	91.3	96.9	2.60
	Skateformer [23]	93.5	97.8	89.8	91.4	98.3	-
	AutoregAd-HGformer [11]	94.15	97.83	91.02	92.42	97.98	3.20
Ours	SAT+MRCTM-Transformer	**93.85**	**97.3**	**89.9**	**91.42**	**98.3**	**2.01**

due to its limited size. The evaluation process requires two testing conditions which include Causal training with 30-frame sliding windows for online recognition and Non-causal training on complete sequences for offline evaluation against current best results. The NTU datasets use 25-joint data while NW-UCLA and GoHAR-12 depend on the COCO-WholeBody 59-joint subset. The NTU model has 6 transformer layers (i.e, L) but NW-UCLA and GoHAR-12 models use 2 layers each during training for 50 and 100 epochs with early stopping based on peak validation accuracy. The model processes full sequences directly for NTU and NW-UCLA benchmarks to match previous research while maintaining real-time GoHAR-12 robot control through causal streaming.

D. Evaluation Metrics

The evaluation process for all datasets uses top-1 classification accuracy on their test data. The evaluation of GoHAR-12 includes precision and recall measurements for each class to assess the reliability of essential high level commands between *Pointing to sit* and *Stay here*. The results from NTU-60/120 and NW-UCLA Multiview Action3D receive evaluation against current state-of-the-art methods in the field.

E. Offline Evaluation

As we do not have more than two views in the GoHar-12 dataset, we perform only the cross subject evaluation on our dataset. The Spatial Attention Tokenization (SAT) and Multi-Resolution Causal Temporal Mixing (MRCTM) modules in our model achieve a top-1 accuracy of 98.4% for cross subject evaluation.

The proposed method undergoes evaluation on NTU-RGB+D 60 and 120 datasets which represent the leading large-scale skeleton action recognition benchmarks. The model demonstrates performance of **93.85%** accuracy when tested on NTU-60 data through both X-Sub and **97.3%** X-View evaluation protocols.

The model demonstrates good performance on NTU-120 by achieving **89.9%** accuracy in the X-Sub protocol and **91.42%** accuracy in the X-Set protocol which indicates its ability to handle extensive class sets and intricate body movements.

The NW-UCLA dataset contains 10 action classes recorded from three camera angles where our model reaches **98.3%** X-View accuracy. The causal transformer backbone provides reliable predictions.

F. Comparison with the State-of-the-Art Methods

The comparison results against state-of-the-art methods include high impact works from very early stages including GCN and transformer-based models are presented in Table II.

The SAT and MRCTM integration enables our system to handle both inter-subject variations and small pose uncertainties which previous transformer-based models struggled to address. Although, the model does not outperform the state-of-the-art works in NTU-RGB+D datasets, we can still see that it is inline with the recent models. On the NW-UCLA dataset, our network outperforms the current state-of-the-art works.

1) Discussion: Our model achieves results that match recent high-performing architectures on NTU-RGB+D 60 and 120 benchmarks although it does not reach absolute state-of-the-art levels. Our proposed design achieves better results than all previous state-of-the-art methods on the NW-UCLA dataset. The majority of previous research on NTU-RGB+D and NW-UCLA used offline recognition models that require complete sequence observation before classification. The methods lack real-time deployment capability for robotic interaction because they need complete sequence data access. Our architecture combines causal modeling with SAT tokenization and MRCTM mixing to perform online streaming recognition which enables real-time decision-making for controlling quadruped robots.

G. Online Evaluation Setup

The proposed action recognition framework is tested on the Unitree Go1 quadruped robot for online evaluation. The Intel RealSense camera installed on the robot recorded RGB video streams of people who interacted with it. The RTMW-3D [32] pose estimation pipeline processed incoming video frames to obtain normalized 3D skeletal joints. We use depth

TABLE III

ABLATION RESULTS (% TOP-1 ACCURACY) ON GOHAR-12 (CROSS-SUBJECT) AND NW-UCLA (CROSS-VIEW). SAT = SPATIAL ATTENTION TOKENIZATION, MRCTM = MULTI-RESOLUTION CAUSAL TEMPORAL MIXING, DWCONV = DEPTHWISE CAUSAL CONVOLUTION.

Model Variant	GoHAR-12	NW-UCLA
Baseline Transformer	90.3	88.5
+ SAT only	95.2	92.7
+ MRCTM only	94.1	94.3
+ DWConv only	92.8	90.5
+ SAT + MRCTM (no DWConv)	96.7	96.2
+ SAT + DWConv (no MRCTM)	96.0	94.8
+ MRCTM + DWConv (no SAT)	94.9	95.1
Full Model	**98.4**	**98.3**

data with skeleton information to produce camera-based 3D coordinates with whole-body representation that use 59 joints according to the COCO-WholeBody format.

The proposed SAT+MRCTM Causal Transformer architecture processed the 3D keypoints for real-time action recognition with DeepSORT+ReID [39], [40] network to track subjects and avoid misclassifications. The model processed each 30-frame temporal clip to generate recognition logits throughout the entire sequence. The per-window accuracy calculation used the following the final decision for that specific clip.

The system measured online recognition accuracy through clip-based evaluation by comparing formula:

$$\text{Acc}_{\text{win}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i],$$

The total number of evaluated clips receives the symbol N in this formula.

The evaluation process followed these steps. The test participants completed five randomly chosen actions five times each during the evaluation process. We recorded and performed manual verification to determine the correct error rate measurement.

The proposed framework achieved a 95.3% average online recognition rate while maintaining a 34 ms (29 fps) per-window latency (for end-to-end architecture) during real-time operation on a single NVIDIA RTX 3060 6GB GPU. The system achieved above 90% accuracy for all actions while showing exceptional performance in detecting locomotion-based gestures including walk, turn, and stay. The system achieved 91.8% accuracy for pointing to sit gestures but struggled with these gestures because of partial body coverage and similar movement patterns between actions. We achieve this with just 2.01 million parameters compared to the existing works with over 2.6 million parameters. We plan to release the dataset via OrchidID after the blind review process to avoid any complications.

V. ABLATION STUDY

The ablation study of proposed modules on GoHAR-12 data reveals their individual contributions while the NW-UCLA dataset demonstrates these trends because of its small size and architectural sensitivity. The evaluation assesses three components separately which include SA) and MRCTM and depthwise causal convolution (DWConv) branch. The complete model implements all three components inside a causal transformer architecture.

1) Setup: The training process for each variant follows the exact parameters which the main experiments used (AdamW optimizer and 300 epochs and cosine learning rate schedule). The evaluation of top-1 accuracy uses three random seeds to minimize experimental results variability. The evaluation of GoHAR-12 uses cross-subject testing while NW-UCLA requires cross-view testing.

2) Results and Discussions: The results are presented in Table III. The accuracy drops substantially when SAT is removed because the model needs tokenized attention to distinguish between the similar actions of *pointing-to-sit* and *stay-here* on GoHAR-12. The removal of MRCTM results in reduced robustness on NW-UCLA because the multi-resolution dilations help maintain recognition stability when dealing with motions of different lengths. The removal of DWConv results in a minor decrease in performance which indicates that local temporal filtering enhances the transformer's global modeling capabilities. The complete model achieves the best results in both datasets which demonstrates that each component enhances the network's ability to recognize and maintain stability in its output.

The ablation test shows SAT stands as the essential element for learning discriminative spatial information yet MRCTM improves temporal generalization and DWConv adds local refinement capabilities. The combined effect of these components results in substantial performance improvements with better stability across different datasets.

VI. CONCLUSIONS

This research introduces a new causal transformer system which combines Spatial-Attention Tokenization (SAT) with the Multi-Resolution Causal-Temporal Module (MRCTM) to perform skeleton-based online action recognition. A dual-head output system that produces last-timestep logits for online recognition and pooled logits for offline stability. The model design enables both strong causal prediction performance and reliable sequence-level accuracy through its combined architecture.

The proposed architecture received testing on our GoHAR-12 dataset and three other benchmark datasets including NTU-RGB+D 60 and NTU-RGB+D 120 and NW-UCLA. The model achieved a recognition accuracy of 98.4% when tested on GoHar-12 dataset through offline evaluation methods and 95.3% through online. The model achieved state-of-the-art recognition results on the NTU-RGB+D 60, the NTU-RGB+D 120 and the NW-UCLA benchmarks.

The proposed architecture demonstrates excellent performance for interactive robotic action recognition because it provides both fast causal inference and reliable recognition stability. The framework needs future development to include facial expressions and gaze cues and hand-object interaction features as additional modalities. The integration of

additional multimodal signals will lead to improved action understanding and intention prediction which will enhance safe and natural human-robot interaction in collaborative spaces.

ACKNOWLEDGMENT

The authors would like to acknowledge the assistance of ChatGPT, an AI-based language model, which was used to support the rewriting and refinement of text for improved readability and reasoning.

REFERENCES

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[2] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[3] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152.

[4] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[5] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," *CoRR*, vol. abs/1703.08274, 2017. [Online]. Available: http://arxiv.org/abs/1703.08274

[6] H. Cui, R. Huang, R. Zhang, and T. Hayama, "Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling," *Neurocomputing*, p. 130066, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225007386

[7] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[8] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Underst.*, vol. 208-209, p. 103219, 2020.

[9] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognition*, p. 108487, 2021.

[10] Y. Zhou, C. Li, Z.-Q. Cheng, Y. Geng, X. Xie, and M. Keuper, "Hypergraph transformer for skeleton-based action recognition," *ArXiv*, vol. abs/2211.09590, 2022.

[11] A. Ray, A. Raj, and M. H. Kolekar, "Autoregressive adaptive hypergraph transformer for skeleton-based activity recognition," *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9690–9699, 2024.

[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019.

[13] H. Zhang, Y. Li, P. Wang, Y. Liu, and C. Shen, "Rgb-d based action recognition with light-weight 3d convolutional networks," *ArXiv*, vol. abs/1811.09908, 2018.

[14] "Go1 robot dog," https://www.unitree.com/go1, accessed: 2025-09-13.

[15] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020.

[16] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 359–13 368.

[17] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 154–20 164.

[18] S. Chi, H.-g. Chi, Q. Huang, and K. Ramani, "Infogcn++: Learning representation by predicting the future for online human skeleton-based action recognition," *arXiv preprint arXiv:2310.10547*, 2023.

[19] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 10 444–10 453.

[20] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, and Q. Miao, "Transformer for skeleton-based action recognition: A review of recent advances," *Neurocomput.*, vol. 537, no. C, p. 164–186, Jun. 2023.

[21] M. Liu, H. Liu, Q. Hu, B. Ren, J. Yuan, J. Lin, and J. Wen, "3d skeleton-based action recognition: A review," *ArXiv*, vol. abs/2506.00915, 2025.

[22] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*. Springer, 2021, pp. 694–701.

[23] J. Do and M. Kim, "Skateformer: skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*. Springer, 2025, pp. 401–420.

[24] N. Heidari and A. Iosifidis, "Progressive spatio-temporal graph convolutional network for skeleton-based human action recognition," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3220–3224, 2020.

[25] Y. Xu, K. Peng, D. Wen, R. Liu, J. Zheng, Y. Chen, J. Zhang, A. Roitberg, K. Yang, and R. Stiefelhagen, "Skeleton-based human action recognition with noisy labels," *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4716–4723, 2024.

[26] M. Terreran, L. Barcellona, and S. Ghidoni, "A general skeleton-based action and gesture recognition framework for human–robot collaboration," *Robotics and Autonomous Systems*, vol. 170, p. 104523, 2023.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[29] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[31] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[32] T. Jiang, X. Xie, and Y. Li, "Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation," *ArXiv*, vol. abs/2407.08634, 2024.

[33] S. Jin, W. Xu, Y. Xu, Y. Wang, C. Lu, P. Luo, X. Wang, and C. Qian, "Whole-body human pose estimation in the wild," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 196–214.

[34] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 180–189.

[35] Q. Wang, J. Peng, S. Shi, T. Liu, J. He, and R. Weng, "Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition," *2023 IEEE International Conference on Big Data (BigData)*, pp. 936–945, 2021.

[36] Z. Gao, P. Wang, P. Lv, X. Jiang, Q. dong Liu, P. Wang, M. Xu, and W. Li, "Focal and global spatial-temporal transformer for skeleton-based action recognition," *ArXiv*, vol. abs/2210.02693, 2022.

[37] Y. Pang, Q. Ke, H. Rahmani, J. Bailey, and J. Liu, "Igformer: Interaction graph transformer for skeleton-based human interaction recognition," in *European Conference on Computer Vision*, 2022.

[38] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *ArXiv*, vol. abs/1711.05101, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3312944

[39] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[40] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 748–756.