This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LaDy: Lagrangian-Dynamic Informed Network for Skeleton-based Action Segmentation via Spatial-Temporal Modulation

Haoyu Ji[1,†], Xueting Liu[2,†], Yu Gao[1], Wenze Huang[1], Zhihao Yang[1],
Weihong Ren[1], Zhiyong Wang[1,*], Honghai Liu[1,3]
[1]Harbin Institute of Technology, Shenzhen
[2]Southern University of Science and Technology [3]Southeast University

Abstract

Skeleton-based Temporal Action Segmentation (STAS) aims to densely parse untrimmed skeletal sequences into frame-level action categories. However, existing methods, while proficient at capturing spatio-temporal kinematics, neglect the underlying physical dynamics that govern human motion. This oversight limits inter-class discriminability between actions with similar kinematics but distinct dynamic intents, and hinders precise boundary localization where dynamic force profiles shift. To address these, we propose the Lagrangian-Dynamic Informed Network (LaDy), a framework integrating principles of Lagrangian dynamics into the segmentation process. Specifically, LaDy first computes generalized coordinates from joint positions and then estimates Lagrangian terms under physical constraints to explicitly synthesize the generalized forces. To further ensure physical coherence, our Energy Consistency Loss enforces the work-energy theorem, aligning kinetic energy change with the work done by the net force. The learned dynamics then drive a Spatio-Temporal Modulation module: Spatially, generalized forces are fused with spatial representations to provide more discriminative semantics. Temporally, salient dynamic signals are constructed for temporal gating, thereby significantly enhancing boundary awareness. Experiments on challenging datasets show that LaDy achieves state-of-the-art performance, validating the integration of physical dynamics for action segmentation. Code is available at https://github.com/HaoyuJi/LaDy.

1. Introduction

Skeleton-based Temporal Action Segmentation (STAS) has emerged as a critical research frontier, aiming to parse untrimmed skeletal sequences into dense, frame-level actions. The skeleton modality offers a compelling balance of computational efficiency and intrinsic robustness to environmental variations [24, 52], making it ideal for applications from human-computer interaction [47] to autonomous driving [19, 22] and clinical surveillance [5, 13].

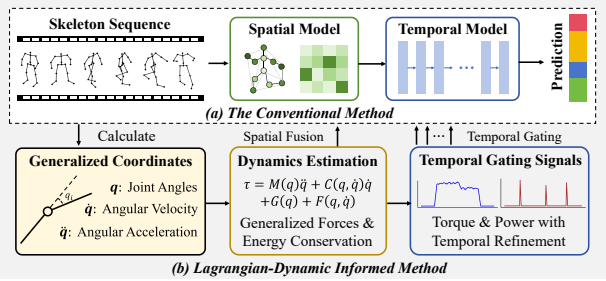


Figure 1. Conceptual comparison of STAS frameworks. (a) Conventional models capture only spatio-temporal kinematic patterns. (b) Our LaDy framework introduces Lagrangian dynamics to derive generalized forces that modulate the spatio-temporal features, leading to improved discriminability and boundary localization.

Prevailing STAS methods employ deep architectures to capture intricate spatio-temporal dependencies, as shown in Fig. 1(a). Spatially, Graph Convolutional Networks (GCNs) [14, 32, 55] and attention mechanisms [36, 45] model the non-Euclidean topology of the body, while temporally, Temporal Convolutional Networks (TCNs) [12, 14, 28] and Transformers [25, 31, 57] capture long-range dependencies. These methods excel at learning complex spatio-temporal kinematic patterns—the "what" and "how" of motion—achieving considerable success.

However, a fundamental limitation underlies current methods: they operate almost exclusively in the domain of kinematics, remaining largely agnostic to the underlying dynamics that govern the motion. Human movement is not merely a sequence of poses; it is the physical manifestation of forces dictated by the principles of classical mechanics. Neglecting this dynamic foundation discards a rich source of information defining the intent and causality of an action. This "dynamics-agnostic" approach leads

[†]Equal contribution [*]Corresponding author

to two significant, unresolved challenges. First, it compromises inter-class discriminability for actions that share similar kinematic profiles but arise from disparate dynamic intents (e.g., "Moving Cart" and "Walking", which are kinematically similar but dynamically disparate). Second, it hinders precise boundary localization, because action transitions are not merely kinematic shifts but are fundamentally defined by abrupt changes in the dynamic force profile—changes that kinematic-only models often blur.

To bridge this "kinematics-dynamics" gap, we introduce the Lagrangian-Dynamic Informed Network (LaDy), which integrates the principles of Lagrangian dynamics directly into action segmentation, as illustrated in Fig. 1(b). Instead of solely observing how a person moves, LaDy reasons about why they move by explicitly estimating the generalized forces (joint torques) driving the skeletal system. Specifically, our Lagrangian Dynamics Synthesis (LDS) module computes generalized coordinates from joint configurations and then utilizes physically-constrained modules to estimate the constituent Lagrangian terms (e.g., Inertia, Coriolis, Gravity) for synthesizing the generalized forces. To ensure these forces are physically coherent, we introduce an Energy Consistency Loss (ECLoss). This physics-based regularizer enforces the work-energy theorem, supervising the network by aligning the system's kinetic energy change with the net work done by all forces. Finally, the learned dynamic features are leveraged by a Spatio-Temporal Modulation (STM) mechanism, which fuses forces with spatial representations and uses salient dynamic signals for hierarchical temporal gating.

We validate LaDy through extensive experiments on six challenging STAS datasets [33, 37, 40, 50]. Our method achieves new state-of-the-art performance, validating our central hypothesis that integrating physical dynamics provides a more discriminative and precise foundation for action segmentation. Our main contributions are threefold: (i) We propose LaDy, the first framework to introduce Lagrangian dynamics into STAS. Its core is the Lagrangian Dynamics Synthesis (LDS) module, which estimates physics-informed generalized forces to assist segmentation. (ii) We introduce an Energy Consistency Loss (ECLoss), a physics-based regularizer that enforces the work-energy theorem to ensure the physical coherence of the forces. (iii) We design a dynamics-driven Spatio-Temporal Modulation (STM) that leverages the forces via spatial fusion and hierarchical temporal gating, enhancing action discriminability and boundary localization.

2. Related Work

Temporal Action Segmentation. Temporal Action Segmentation (TAS) research is largely differentiated by input modality. Video-based TAS methods focus on learning from RGB or flow features. The field has evolved by enhancing core temporal architectures, including TCNs [12, 15, 16, 28–30, 41, 43], Graph-based models [21, 46, 59], Transformers [2, 3, 11, 38, 48, 57], diffusion models [1, 34, 35] and boundary refinement techniques [23, 49]. Skeleton-based TAS, our focus, models spatio-temporal dependencies from skeleton sequences. The dominant paradigm, established by MS-GCN [14], combines GCNs for spatial topology with TCNs for temporal modeling. This pipeline has been the subject of numerous refinements, including decoupled architectures [31], motion-aware enhancements [4], discriminative graph networks [32], contrastive learning [7], multi-scale pyramids [8] and data processing strategies [53, 54, 56]. Alternative approaches leverage spatio-temporal attention [25, 36, 45] and language-assisted modeling [6, 24, 26]. However, these methods are fundamentally kinematic, modeling motion appearance while ignoring its underlying dynamics, thus limiting inter-class discriminability and boundary precision. To our knowledge, LaDy is the first framework to introduce Lagrangian dynamics into STAS, directly addressing this gap.

Physics-Informed Dynamics Modeling. Integrating physical principles into deep learning is a significant research area, often utilizing Lagrangian [10, 39, 64] or Hamiltonian [17] mechanics. In the context of human motion, these principles are leveraged in three main ways: (1) As Feature Enhancers: Physics-derived quantities (e.g., forces, torques) are used as auxiliary inputs or features to deep networks for tasks like motion prediction [61, 62] or action generation [42]. (2) As Supervisory Signals: Physics is used as a loss function or regularizer to enforce physical plausibility, often for trajectory refinement [44, 60], pose correction [51], or guiding representation learning via physics-augmented decoders [18, 58]. (3) As Control Signals: Dynamic properties are used to modulate the information flow within a network, such as using velocity to gate LSTMs [20] or fusing torques in a Graph-GRU [63]. Despite this progress, leveraging explicit, complete skeletal dynamics for spatio-temporal feature processing remains underexplored. Our work presents the first physics-informed network for STAS, leveraging the Lagrangian dynamics equation and the work-energy theorem to estimate forces for spatio-temporal modulation.

3. Method

The goal of STAS is to densely parse an untrimmed skeletal sequence. We denote the input sequence as $X \in \mathbb{R}^{C_0 \times T \times V}$, with T frames, V joints and C_0 input channels (e.g., coordinates). STAS maps X to a sequence of frame-level one-hot labels $Y \in \mathbb{R}^{Q \times T}$, where Q is the number of action classes.

3.1. Overall Architecture

As shown in Fig. 2, LaDy achieves action segmentation through two parallel streams: a standard and conventional

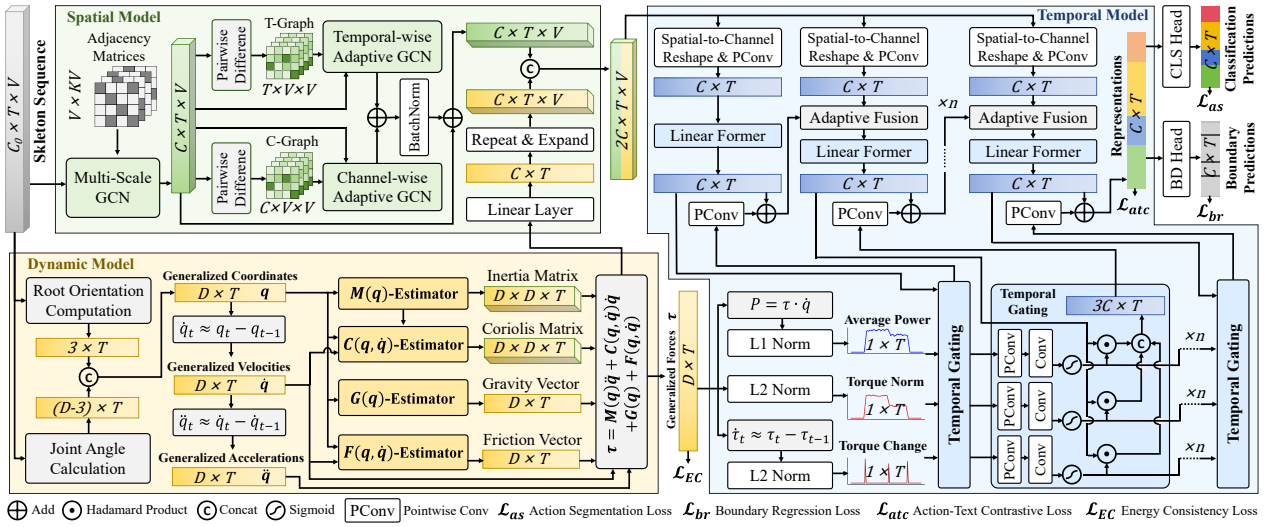


Figure 2. Overview of the LaDy framework. The Lagrangian Dynamic Model (lower-left) synthesizes generalized forces from coordinates, constrained by an Energy Consistency Loss for physical coherence. Concurrently, the Spatial Model (upper-left) extracts kinematic features (GCNs) and fuses them with these physics-aware dynamics. The multi-stage Temporal Model (right) processes the fused spatial representation, where each stage is hierarchically gated by force-driven signals to enhance boundary awareness.

Spatio-Temporal Model as main branch (top) and our novel Lagrangian Dynamic Model as auxiliary branch (bottom), which interact via Spatio-Temporal Modulation (STM).

The Lagrangian Dynamic Model, the core of our innovation, first computes generalized coordinates, velocities, and accelerations ($q, \dot{q}, \ddot{q} \in \mathbb{R}^{T\times D}$) from input coordinates X. These are fed into the physically-constrained Lagrangian Dynamics Synthesis (LDS) module to estimate the generalized forces $\tau \in \mathbb{R}^{T\times D}$ (Sec. 3.2). These forces are supervised by our Energy Consistency Loss (ECloss) to ensure physical rationality (Sec. 3.3). This branch outputs two signals for modulation: (1) a dynamic feature $F_{dyn} \in \mathbb{R}^{C\times T\times V}$ (via a Linear layer) and (2) three salient dynamic signals (Power, Torque, Torque Change) $\in \mathbb{R}^{1\times T}$.

In the main branch, the Spatial Model first extracts kinematic features F_{kin} from X using Multi-Scale [31] and Adaptive GCNs [4]. Spatial modulation (Sec. 3.4) then fuses F_{kin} with the dynamic feature F_{dyn} to produce an enriched representation $F_{sp} \in \mathbb{R}^{2C\times T\times V}$. F_{sp} is then fed into the L-stage Temporal Model. Each stage comprises a Linear Transformer [27], an adaptive fusion module, and a temporal modulation (Sec. 3.4), which performs hierarchical gating driven by three salient dynamic signals.

The final L-th stage representation F_R is fed to Classification and Boundary Heads, yielding initial classification predictions Y_c^0 and boundary predictions Y_b^0. F_R is also supervised by an action-text contrastive loss ($\mathcal{L}_{atc}$), following [24, 26]. These initial predictions are refined by a multi-stage refinement module to produce final outputs Y_c^F and Y_b^F. The model is trained end-to-end with a composite objective: standard segmentation losses [23, 31], $\mathcal{L}_{atc}$, and our $\mathcal{L}_{EC}$. Inference employs the post-processing [23] scheme.

3.2. Lagrangian Dynamics Synthesis

The Lagrangian Dynamics Synthesis (LDS) module moves beyond black-box estimation, synthesizing generalized forces from kinematics by structurally parameterizing the Lagrangian equation with embedded physical constraints.

3.2.1. Generalized Coordinates Computation

The skeleton is modeled as an open kinematic chain $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of V joints and $\mathcal{E}$ defines the bone connectivity. The generalized coordinates q comprise the root orientation q^{root} and local joint rotations q^{local}.

Root Orientation (q^{root}): We define a time-varying local body frame $\{\vec{x}_{axis}(t), \vec{y}_{axis}(t), \vec{z}_{axis}(t)\}$ at the root joint j_{ro}-SpineBase using the Cartesian positions $p^j \in \mathbb{R}^3$ of j_{ro} and its neighbors (j_m-SpineMid, j_{rh}-RightHip, j_{lh}-LeftHip). The primary axes are:

$$\vec{v}_y = p^{j_m}(t) - p^{j_{ro}}(t), \quad \vec{v}_{x,raw} = p^{j_{rh}}(t) - p^{j_{lh}}(t). \quad (1)$$

An orthonormal basis is obtained via Gram-Schmidt orthogonalization:

$$\begin{aligned} \vec{y}_{axis} &= \mathcal{N}(\vec{v}_y), \quad \vec{z}_{axis} = \mathcal{N}(\vec{v}_{x,raw} \times \vec{y}_{axis}), \\ \vec{x}_{axis} &= \mathcal{N}(\vec{y}_{axis} \times \vec{z}_{axis}), \end{aligned} \quad (2)$$

where $\mathcal{N}(\cdot)$ is L2 normalization. These axes form the rotation matrix $R_{root}(t) = [\vec{x}_{axis}|\vec{y}_{axis}|\vec{z}_{axis}] \in SO(3)$, which is converted to an axis-angle representation $q^{root}(t) \in \mathbb{R}^3$.

Local Joint Rotations (q^{local})**:** For each joint j with parent $p = pa(j)$ and grandparent $gp = pa(p)$, we define

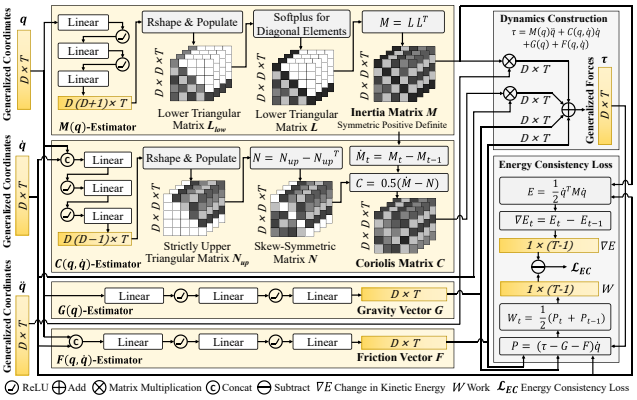


Figure 3. Physics-constrained dynamics synthesis and energy-based supervision. Generalized states $(q, \dot{q}, \ddot{q})$ are fed into physics-constrained estimators to derive the Lagrangian terms (M, C, G, F). These terms are then assembled via the Lagrangian equation to synthesize the generalized forces (τ). Concurrently, the Energy Consistency Loss $(\mathcal{L}_{EC})$ regularizes the forces.

the parent bone $v_{pa}(t) = p^p(t) - p^{gp}(t)$ and child bone $v_{child}(t) = p^j(t) - p^p(t)$. The local rotation $q^j(t) \in \mathbb{R}^3$ is the axis-angle vector that aligns the direction of $v_{pa}(t)$ with $v_{child}(t)$:

$$\begin{aligned} \vec{a}(t) &= \mathcal{N}(\mathcal{N}(v_{pa}(t)) \times \mathcal{N}(v_{child}(t))), \\ \theta(t) &= \arccos(\mathcal{N}(v_{pa}(t)) \cdot \mathcal{N}(v_{child}(t))), \\ q^j(t) &= \theta(t) \cdot \vec{a}(t). \end{aligned} \tag{3}$$

Note: For 2D data, $q^{root} \in \mathbb{R}^1$ is the root bone angle and $q^j \in \mathbb{R}^1$ is the signed angle between v_{pa} and v_{child}.

State Variables Assembly: The generalized coordinate vector $q(t) \in \mathbb{R}^D$ is the concatenation of $q^{root}(t)$ and all $q^j(t)$ for $j \in \mathcal{V}' \subset \mathcal{V}$. We approximate the generalized velocity $\dot{q}$ and acceleration $\ddot{q}$ using first-order backward finite differences (assuming $\Delta t = 1$ and $q(-1) = 0$):

$$\dot{q}(t) = q(t) - q(t-1), \quad \ddot{q}(t) = \dot{q}(t) - \dot{q}(t-1). \tag{4}$$

The resulting state variables $\{q, \dot{q}, \ddot{q}\} \in \mathbb{R}^{D \times T}$ serve as the input to the dynamics estimation.

3.2.2. Physics-Constrained Dynamics Estimation

We estimate the terms of the Lagrangian dynamics equation, $\tau = M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) + F(q, \dot{q})$, using four neural networks $(\mathcal{F}_M, \mathcal{F}_C, \mathcal{F}_G, \mathcal{F}_F)$ that embed physical constraints, as illustrated in Fig. 3.

Inertia Matrix $M(q)$: The inertia matrix $M(q) \in \mathbb{R}^{D \times D \times T}$ must be symmetric positive definite. We enforce this by parameterizing $M(q)$ via Cholesky decomposition, $M(q) = L(q)L(q)^T$. An MLP, $\mathcal{F}_M$, predicts a vector of $D(D+1)/2$ elements from q, which populates a lower-triangular matrix $L^{low}(q)$. To ensure $M(q)$ is positive definite, the diagonal entries of $L(q)$ are enforced to be positive:

$$L_{ii}(q) = \text{softplus}(L^{low}_{ii}(q)) + \epsilon, \tag{5}$$

where $L'_{ii}(q)$ is the i-th diagonal entry of $L(q)$. The small constant $\epsilon > 0$. The matrix is then $M(q) = L(q)L(q)^T$.

Coriolis Matrix $C(q, \dot{q})$: The Coriolis and centrifugal matrix $C(q, \dot{q}) \in \mathbb{R}^{D \times D \times T}$ is constrained by the passivity property, requiring $\dot{M}(q) - 2C(q, \dot{q})$ to be skew-symmetric. We enforce this by defining $N = \dot{M} - 2C$, where $N(q, \dot{q}) = -N(q, \dot{q})^T$ is a skew-symmetric matrix. An MLP, $\mathcal{F}_C$, takes the channel-wise concatenation $\text{Concat}(q, \dot{q})$ as input and outputs a vector of $D(D-1)/2$ elements, which populates a strictly upper-triangular matrix $N^{up}(q, \dot{q})$. The full matrices N and C are then constructed as:

$$\begin{aligned} N(q, \dot{q}) &= N^{up}(q, \dot{q}) - N^{up}(q, \dot{q})^T, \\ C(q, \dot{q}) &= 0.5(\dot{M}(q) - N(q, \dot{q})), \end{aligned} \tag{6}$$

$\dot{M}(q)$ is approximated via finite differences, $\dot{M}(q(t)) \approx M(q(t)) - M(q(t-1))$. This construction mathematically guarantees the passivity constraint.

Gravity Vector $G(q)$ and Other Forces $F(q, \dot{q})$: The gravitational torque $G(q) \in \mathbb{R}^{D \times T}$ and non-conservative forces $F(q, \dot{q}) \in \mathbb{R}^{D \times T}$ (e.g., friction, external interactions) are estimated directly using two separate MLPs:

$$G(q) = \mathcal{F}_G(q), \quad F(q, \dot{q}) = \mathcal{F}_F(\text{Concat}(q, \dot{q})). \tag{7}$$

Generalized Force Synthesis: The estimated generalized force $\tau \in \mathbb{R}^{D \times T}$ is synthesized by assembling the learned components according to the Lagrangian equation:

$$\tau = M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) + F(q, \dot{q}), \tag{8}$$

where the products (e.g., $M(q)\ddot{q}$ and $C(q, \dot{q})\dot{q}$) denote batched matrix-vector multiplications performed independently at each time step t. This force τ, along with M, G, F, and $\dot{q}$, is then passed to the energy consistency loss and spatio-temporal modulation module.

3.3. Energy Consistency Loss

To ensure physically coherent force synthesis, we introduce the Energy Consistency Loss $(\mathcal{L}_{EC})$, a physics-based regularizer that enforces the Work-Energy Theorem. This theorem equates the change in kinetic energy (ΔE_K) to the work (W) done by the net force. In the rotational skeleton system, this work is done by the net torque $\tau_{net} = \tau - G - F$, where τ is the predicted actuation and G and F are load torques (gravity, friction). The integral form over a discrete time step $\Delta t = [t-1, t]$ is:

$$E_K(t) - E_K(t-1) = \int_{t-1}^{t} P(s) ds, \tag{9}$$

where $P(s) = \tau_{net}(s) \cdot \dot{q}(s)$ is the instantaneous power.

Discrete-Time Work-Energy Calculation: We first compute the kinetic energy at each time step t:

$$E_K(t) = 0.5 \cdot \dot{q}(t)^T M(t) \dot{q}(t), \tag{10}$$

where $\dot{q}(t) \in \mathbb{R}^{D\times 1}$ and $M(t) \in \mathbb{R}^{D\times D}$. The change in kinetic energy is $\Delta E_K(t) = E_K(t) - E_K(t-1)$ during the interval $[t-1, t]$. Next, the net power $P(t)$ is:

$$P(t) = \sum_{i=1}^{D} [(\tau(t) - G(t) - F(t)) \cdot \dot{q}(t)], \tag{11}$$

where $\cdot$ denotes the Hadamard product. The work $W(t)$ over the interval $[t-1, t]$ is approximated as $W(t) = 0.5 \cdot (P(t) + P(t-1))$ using the trapezoidal rule.

Relative Energy Residual Loss: A naive L1 loss between ΔE_K and W would be highly sensitive to the magnitude of motion. We therefore design a scale-invariant relative energy residual r_E:

$$r_E(t) = \frac{\Delta E_K(t) - W(t)}{|\Delta E_K(t)| + |W(t)| + \delta} \cdot \mathcal{M}(t). \tag{12}$$

The denominator normalizes the residual for robustness to varying energy scales, while δ ensures numerical stability. The mask $\mathcal{M}(t)$ directly zeros out the residual $r_E(t)$ where the denominator is near-zero, preventing noise amplification during static or micro-movement phases. The final loss $\mathcal{L}_{EC}$ is the Huber loss computed between the residual $r_E(t)$ and $\mathbf{0}$ and averaged over all time steps, which provides robustness to potential outliers in the dynamic estimates.

3.4. Spatio-Temporal Modulation

The dynamics offer causal insights crucial for resolving kinematic ambiguities and localizing boundaries. We leverage this via the Spatio-Temporal Modulation (STM), which injects these insights into the spatial and temporal streams.

3.4.1. Spatial Modulation

Spatial modulation enriches the spatial representation by fusing force information, enhancing inter-class discriminability. Given the kinematic features $F_{kin} \in \mathbb{R}^{C\times T\times V}$ from the GCN and the generalized forces $\tau \in \mathbb{R}^{D\times T}$, we first align their dimensions. τ is projected into the kinematic channel dimension C using a linear layer, yielding the dynamic feature $F'_{dyn} \in \mathbb{R}^{C\times T}$. This feature is then expanded V times along the joint dimension to create $F_{dyn} \in \mathbb{R}^{C\times T\times V}$. Finally, F_{dyn} and F_{kin} are concatenated along the channel dimension, producing the final spatial representation $F_{sp} \in \mathbb{R}^{2C\times T\times V}$. This integrated feature, encoding both geometric pose and its causal dynamics, serves as the input to the temporal model.

3.4.2. Temporal Modulation

Temporal modulation employs salient dynamic signals to hierarchically gate the temporal feature flow, enhancing sensitivity to high-energy events and abrupt dynamic shifts characteristic of action boundaries.

Salient Dynamic Signal Construction: We first distill three salient 1D signals from the dynamic state variables to quantify the motion's instantaneous energy and transients: (1) Instantaneous Power (g_P): The L1 norm of the element-wise power ($\tau \cdot \dot{q}$), capturing total energy expenditure: $g_P(t) = \|\tau(t) \cdot \dot{q}(t)\|_1$. (2) Torque Norm ($g_\tau$): The L2 norm of the generalized forces, indicating the overall actuation magnitude: $g_\tau(t) = \|\tau(t)\|_2$. (3) Torque Change ($g_{\dot{\tau}}$): The L2 norm of the first-order temporal difference of τ (via finite difference), quantifying dynamic transients that often signal boundary shifts: $g_{\dot{\tau}}(t) = \|\tau(t) - \tau(t-1)\|_2$. These three signals ($g_P, g_\tau, g_{\dot{\tau}} \in \mathbb{R}^T$) are concatenated, forming the initial dynamic gate $G^{(0)}_{dyn} \in \mathbb{R}^{3\times T}$ for the first stage.

Multi-Stage Gating: Our temporal model features a parallel gating branch of L stages, mirroring the main temporal backbone to enable hierarchical modulation. At each stage $l \in \{1, \dots, L\}$, the module takes two inputs: the main temporal features $H^{(l)}_T \in \mathbb{R}^{C\times T}$ from stage l and the gate signal $G^{(l-1)}_{dyn} \in \mathbb{R}^{3\times T}$ from the previous gating stage. The process at stage l unfolds in two steps:

First, the incoming gate signal $G^{(l-1)}_{dyn}$ is refined. A key design choice is to process the three dynamic signals in parallel, independent streams (Fig. 2) to preserve their distinct physical interpretations (Power, Torque, Torque Change) without channel cross-talk. Specifically, each channel $g^{(l-1)}_k$ (where $k \in \{P, \tau, \dot{\tau}\}$) is processed by a separate 1D temporal convolution block $\mathcal{G}^{(l)}_k$, followed by a sigmoid activation σ to produce normalized gate weights:

$$g^{(l)}_k = \sigma(\mathcal{G}^{(l)}_k(g^{(l-1)}_k)), \quad \text{for } k \in \{P, \tau, \dot{\tau}\}. \tag{13}$$

The resulting single-channel gates are concatenated to form the updated gate $G^{(l)}_{dyn} = \text{Concat}(g^{(l)}_P, g^{(l)}_\tau, g^{(l)}_{\dot{\tau}}) \in \mathbb{R}^{3\times T}$, which is propagated to stage $l+1$.

Second, the temporal features $H^{(l)}_T$ are dynamically modulated by the new gate $G^{(l)}_{dyn}$. Each of the three refined gate signals $g^{(l)}_k$ individually modulates the feature map $H^{(l)}_T$ via broadcasted Hadamard product ($\odot$):

$$H^{(l)}_k = H^{(l)}_T \odot g^{(l)}_k, \quad \text{for } k \in \{P, \tau, \dot{\tau}\}. \tag{14}$$

These three modulated features are concatenated along the channel dimension and fused via a 1x1 convolution ($\mathcal{F}_{fuse}$). A residual connection with the original $H^{(l)}_T$ is added for stable training and information preservation:

$$\tilde{H}^{(l)} = \mathcal{F}_{fuse}(\text{Concat}(H^{(l)}_P, H^{(l)}_\tau, H^{(l)}_{\dot{\tau}})) + H^{(l)}_T. \tag{15}$$

The gated feature $\tilde{H}^{(l)}$ is passed to the next stage, forming a hierarchical mechanism that adaptively modulates the feature flow at each temporal scale via dynamic cues, progressively highlighting boundary-sensitive features.

3.5. Overall Training Objective

The LaDy is trained end-to-end by minimizing a composite loss function $\mathcal{L}_{total}$, defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{as} + \lambda_1 \mathcal{L}_{br} + \lambda_2 \mathcal{L}_{atc} + \lambda_3 \mathcal{L}_{EC}, \quad (16)$$

where $\mathcal{L}_{as}$ is a standard action segmentation loss [23, 31] applied to all stage outputs Y_c. $\mathcal{L}_{br}$ is a binary cross-entropy loss [23] supervising Y_b. $\mathcal{L}_{atc}$ denotes the action-text contrastive loss [24, 26]. Finally, $\mathcal{L}_{EC}$ is our ECLoss. The values of λ_1, λ_2, and λ_3 are set to 1.0, 0.8, and 0.1.

4. Experiment

4.1. Setup

Datasets. We evaluate LaDy on six public benchmarks. PKU-MMD v2 [33] is a 52-class daily activity dataset; we report on its cross-subject (X-sub) and cross-view (X-view) splits. MCFS-22 & MCFS-130 [37] are figure skating datasets with 22 and 130 action classes. LARa [40] is a warehouse activity dataset with 8 classes. TCG-15 [50] is a traffic control gesture dataset with 15 classes.

Evaluation Metrics. Following standard TAS protocols [12], we use frame-wise accuracy (Acc), segmental edit distance (Edit), and segmental F1 scores at IoU thresholds of 10%, 25%, and 50% (F1@{10, 25, 50}).

Implementation Details. The feature dimension is $C = 64$. The linear layers for Lagrangian term estimation use 128 channels. We optimize models using Adam with a learning rate of 10^{-3}. Consistent with prior works [31], batch sizes are 8 for PKU-MMD v2/LARa, 2 for TCG-15, and 1 for MCFS-22/130. We train for 300 epochs, except for LARa (60 epochs) due to faster convergence. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.2. Comparisons with the State-of-the-Arts

Quantitative Comparison. We compare LaDy with recent STAS methods on six public datasets, with detailed results in Tab. 1 and Tab. 2. LaDy consistently achieves state-of-the-art performance across nearly all metrics. Its superiority is particularly evident on PKU-MMD v2, where it surpasses the prior best-performing method by significant margins: **+5.2%** in F1@50, **+4.2%** in Edit, and **+2.9%** in Acc on the X-view setting; and **+2.7%** in F1@50, **+1.7%** in Edit, and **+1.5%** in Acc on the X-sub setting. LaDy achieves these substantial gains while remaining computationally efficient (13.67G FLOPs, 1.83M params). This blend of high accuracy and efficiency validates our physics-informed Lagrangian dynamics approach for spatio-temporal modulation in action segmentation.

Qualitative Analysis. As illustrated in Fig. 4, LaDy's segmentation results are visibly more accurate, showing fewer classification errors and smaller boundary deviations than its counterparts. Furthermore, LaDy's boundary confidence scores (curves) are significantly more stable and less volatile. While other methods exhibit noisy fluctuations and spurious peaks within action segments, LaDy effectively suppresses this internal noise. Consequently, LaDy produces sharper and more distinct confidence peaks that align precisely with the true action boundaries. This underscores our model's effectiveness in enhancing class discrimination and precise temporal localization.

Feature Space Analysis. To further assess the learned representations, we visualize the feature space on PKU-MMD v2 using t-SNE (Fig. 5). Visually, LaDy's representations exhibit significantly more compact intra-class aggregation and clearer inter-class separation than previous methods, indicating a more structured semantic distribution. This observation is confirmed by quantitative clustering metrics in Fig. 5: Silhouette Coefficient (SC ↑), Calinski-Harabasz Index (CH ↑), and Davies-Bouldin Index (DB ↓), where LaDy achieves the best scores across all three metrics. This demonstrates that LaDy learns a more discriminative feature space, enhancing inter-class separability.

4.3. Ablation Studies and Analysis

We validate LaDy's components—Lagrangian Dynamics Synthesis (LDS), Spatio-Temporal Modulation (STM), and Energy Consistency Loss (ECLoss)—via ablations on PKU-MMD v2 (X-sub), summarized in Tab. 3. The strong Baseline is built upon DeST, equipped with CTR-GCN's adaptive GCN [9, 26] and LaSA's $\mathcal{L}_{atc}$ [24] (Fig. 2, upper). First, integrating the LDS module with either Spatial Modulation (+LDS+SM) or Temporal Modulation (+LDS+TM) markedly outperforms the dynamics-agnostic Baseline. This confirms our hypothesis that the dynamics prior is highly effective, directly enhancing spatial feature discriminability and temporal boundary perception. Next, combining both modulations (+LDS+STM) achieves a substantial, synergistic performance jump to 75.9% Acc and 66.3% F1@50. This demonstrates that spatial and temporal modulation are complementary, effectively addressing the distinct challenges of inter-class ambiguity (via SM) and precise boundary localization (via TM). Finally, we applied the ECLoss to enforce physical plausibility. This physics-based regularization provides a final, consistent performance boost. This is evident as our full model (+LDS+STM+ECLoss) improves upon the version without the loss (+LDS+STM) across all metrics, achieving the peak performance of 76.2% Acc and 67.0% F1@50. This validates that enforcing the work-energy theorem refines the dynamic representations for more precise segmentation.

Analysis of Synthesized Dynamics. To intuitively vali-

Table 1. Comparison with the latest results on PKU-MMD v2 (X-view and X-sub), and LARa. **Bold** and underline indicate the best and second-best results in each column. FLOPs denote the computational complexity when the input is $X \in \mathbb{R}^{12\times6000\times19}$ on LARa dataset.

Dataset			PKU-MMD v2 (X-view)					PKU-MMD v2 (X-sub)					LARa				
Metric	FLOPs↓	Param.↓	Acc↑	Edit↑	F1@{10,25,50}↑			Acc↑	Edit↑	F1@{10,25,50}↑			Acc↑	Edit↑	F1@{10,25,50}↑		
MS-TCN [12]	3.13G	0.52M	58.2	56.6	58.6	53.6	39.4	65.5	-	-	-	46.3	65.8	-	-	-	39.6
ASFormer [57]	6.20G	1.01M	64.8	62.6	65.1	60.0	45.8	68.3	68.1	71.9	68.2	54.5	72.2	62.2	66.1	61.9	49.2
ASRF [23]	7.00G	1.18M	60.4	59.3	62.5	58.0	46.1	67.7	67.1	72.1	68.3	56.8	71.9	63.0	68.3	65.3	53.2
MS-GCN [14]	30.97G	0.63M	65.3	58.1	61.3	56.7	44.1	68.5	-	-	-	51.6	65.6	-	-	-	43.6
MTST-GCN [4]	108.87G	2.58M	66.5	64.0	67.1	62.4	49.9	70.0	65.8	68.5	63.9	50.1	73.7	58.6	63.8	59.4	47.6
DeST-TCN [31]	5.72G	0.77M	62.4	58.2	63.2	59.2	47.6	67.6	66.3	71.7	68.0	55.5	72.6	63.7	69.7	66.7	55.8
DeST-Former [31]	7.71G	1.10M	67.3	64.7	69.3	65.6	52.0	70.3	69.3	74.5	71.0	58.7	75.1	64.2	70.3	68.0	57.7
MSTAS [8]	34.60G	1.02M	71.8	63.0	68.0	63.4	52.8	67.6	67.0	72.1	69.0	57.1	74.8	60.8	66.2	62.2	50.4
SCSAS [7]	32.70G	1.80M	67.9	64.0	70.4	66.8	55.5	70.3	68.2	72.2	68.2	56.2	74.2	61.7	67.3	65.2	53.5
ME-ST [25]	97.07G	3.16M	74.1	70.5	76.6	73.2	62.4	68.5	67.2	72.3	68.8	58.1	74.2	65.0	71.0	68.2	57.1
LaSA [24]	10.44G	1.58M	69.5	67.8	72.9	69.2	57.0	73.5	73.4	78.3	74.8	63.6	75.3	65.7	71.6	69.0	57.9
LPL [6]	20.30G	3.22M	70.0	67.7	73.3	70.0	58.5	74.7	72.7	78.6	75.8	64.3	**76.1**	65.2	72.3	70.0	58.6
LaDy	13.67G	1.83M	**77.0**	**74.7**	**80.5**	**77.3**	**67.6**	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**	75.6	**65.9**	**72.9**	**70.4**	**59.7**

Table 2. Comparison with the latest results on MCFS-22, MCFS-130, and TCG-15 datasets.

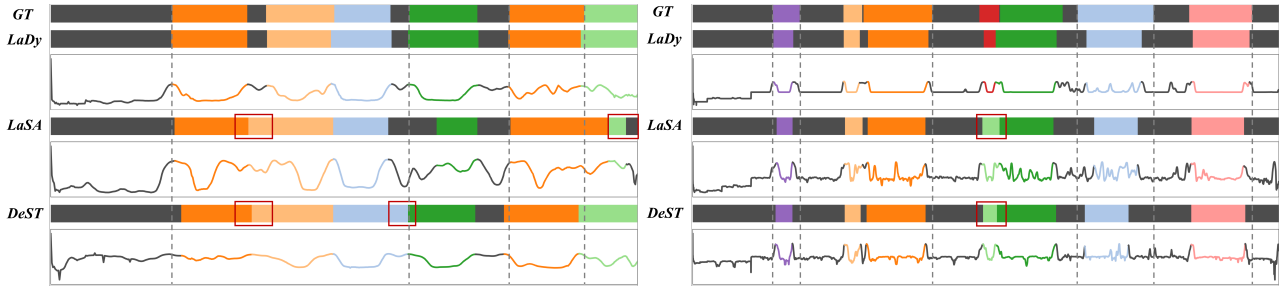
Dataset	MCFS-22					MCFS-130					TCG-15				
Metric	Acc↑	Edit↑	F1@{10, 25, 50}↑			Acc↑	Edit↑	F1@{10, 25, 50}↑			Acc↑	Edit↑	F1@{10, 25, 50}↑		
MS-TCN [12]	75.6	74.2	74.3	69.7	59.5	65.7	54.5	56.4	52.2	42.5	85.5	67.7	70.0	67.3	59.1
ASFormer [57]	78.7	82.3	82.8	77.9	66.9	67.5	69.1	68.3	64.0	55.1	82.3	69.1	71.3	67.7	57.6
ASRF [23]	75.5	77.3	83.3	80.1	69.2	65.6	65.6	66.7	62.3	51.9	83.9	65.3	71.5	69.9	62.4
MS-GCN [14]	75.5	72.6	75.7	70.5	57.9	64.9	52.6	52.4	48.8	39.1	86.6	68.7	70.9	69.7	59.4
IDT-GCN [32]	79.9	84.5	88.0	84.9	74.9	68.6	70.2	70.7	67.3	58.6	-	-	-	-	-
DeST-TCN [31]	78.7	82.3	86.6	83.5	73.2	70.5	73.8	74.0	70.7	61.8	87.1	73.7	78.7	76.9	69.6
DeST-Former [31]	80.4	85.2	87.4	84.5	75.0	71.4	75.8	75.8	72.2	63.0	88.1	75.7	80.0	78.2	71.2
MSTAS [8]	-	-	-	-	-	-	-	-	-	-	88.3	69.5	73.5	70.4	63.9
SCSAS [7]	-	-	-	-	-	67.4	66.9	65.8	62.2	53.1	87.4	69.6	72.8	71.6	67.0
ME-ST [25]	-	-	-	-	-	-	-	-	-	-	88.7	77.3	81.8	78.6	72.4
LaSA [24]	80.8	86.7	89.3	86.2	76.3	72.6	79.3	79.3	75.8	66.6	88.7	77.4	81.2	79.5	73.3
LPL [6]	-	-	-	-	-	-	-	-	-	-	88.8	77.4	81.9	80.0	73.8
LaDy	**81.5**	**89.8**	**90.8**	**86.7**	**76.7**	**72.8**	**79.7**	**80.0**	**76.8**	**66.9**	**89.3**	**77.9**	**82.1**	**80.1**	**74.3**

Table 3. Ablation study of Lagrangian Dynamics Synthesis (LDS), Spatio-Temporal Modulation (STM), and Energy Consistency Loss (ECLoss) on PKU-MMD v2 (X-sub) dataset.

Components	**Acc**	**Edit**	**F1@{10, 25, 50}**		
Baseline	73.6	73.0	78.2	74.6	64.3
+LDS+SM	74.8	74.0	78.9	76.0	65.7
+LDS+TM	74.9	73.5	78.1	75.2	65.8
+LDS+STM	75.9	74.6	79.6	76.9	66.3
+LDS+SM+ECLoss	74.7	74.3	79.1	76.3	66.0
+LDS+TM+ECLoss	75.4	74.5	79.6	76.7	66.5
+LDS+STM+ECLoss (LaDy)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

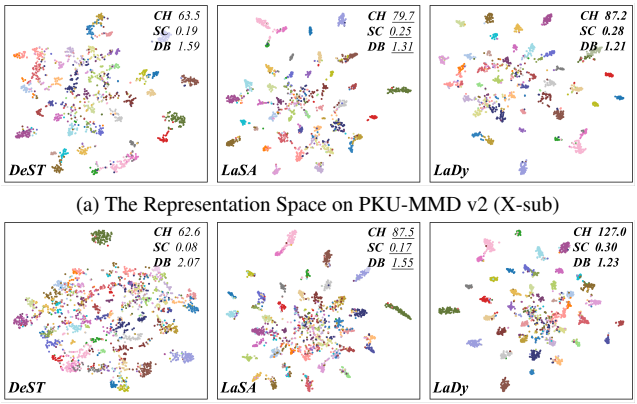
date our model's adherence to physical principles, we visualize the synthesized generalized forces, as shown in Fig. 6. We compute the norm of the 3-axis force for each joint and render it as the thickness of the corresponding skeletal limb. The visualization clearly shows that LaDy learns highly discriminative and physically plausible dynamic profiles. For instance, (1) *Salute* shows high force concentration in the saluting arm; (2) *Throw* activates the entire upper body and arms; (3) *Hopping* correctly identifies the legs as the primary force source; and (4) *Clapping* focuses on the hands and forearms. This interpretable result strongly validates that our physics-constrained LDS module, guided by the ECLoss, successfully synthesizes meaningful and physically coherent dynamics, rather than arbitrary features.

Analysis of Temporal Gating Signals. To further validate the efficacy of our Temporal Modulation, we visualize the learned gating signals. We average the L-stage gating signals and plot them against the ground-truth action segments in Fig. 7. The visualization reveals a key finding: all three signals exhibit sharp, distinct troughs (valleys) that precisely coincide with the ground-truth action boundaries (marked by gray dashed lines). This alignment is particularly pronounced in the "Torque Change" signal. This phenomenon is highly intuitive: action transitions are fundamentally defined by abrupt shifts in the body's dynamic force profile, and LaDy successfully learns to identify these physical, dynamic events as explicit cues for semantic action boundaries. This mechanism provides the temporal network with superior boundary awareness, explaining the strong performance of Temporal Modulation.



(a) PKU-MMD v2 (X-view) (b) MCFS-130

Figure 4. Qualitative results on PKU-MMD v2 and MCFS-130. The top row is the Ground Truth, followed by the segmentation results (bars) and boundary confidence scores (curves) for LaDy, LaSA, and DeST. Different colors denote distinct action classes. Red boxes and gray dashed vertical lines highlight misclassifications and larger boundary deviations observed in other methods compared to LaDy.



(a) The Representation Space on PKU-MMD v2 (X-sub)

(b) The Representation Space on PKU-MMD v2 (X-view)

Figure 5. Visualization of the representation space on PKU-MMD v2. Each point represents an action segment feature, colored according to its ground-truth class. Quantitative clustering metrics are reported in the top-right corner: Silhouette Coefficient (SC ↑), Calinski-Harabasz Index (CH ↑), and Davies-Bouldin Index (DB ↓). Higher (↑) is better for SC/CH; lower (↓) is better for DB.

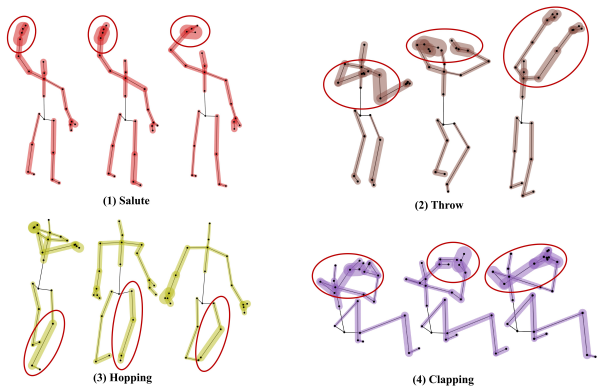


Figure 6. Visualization of synthesized generalized force magnitudes for four actions. Limb thickness is proportional to the norm of the estimated torque. Red circles highlight the primary active limbs, illustrating that the learned dynamics are both class-discriminative and physically plausible.

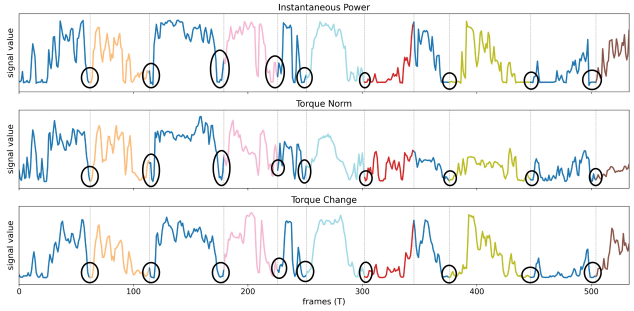


Figure 7. Visualization of layer-averaged dynamic gating signals. The time-series plots for Instantaneous Power, Torque Norm, and Torque Change show signals averaged across all L temporal layers. Note the distinct, sharp troughs (highlighted by ovals) that precisely align with the ground-truth action boundaries (dashed lines), validating their role in boundary localization.

5. Conclusion

In this paper, we addressed the dynamics-agnostic nature of existing Skeleton-based Temporal Action Segmentation (STAS) methods by proposing the Lagrangian-Dynamic Informed Network (LaDy). Our Lagrangian Dynamics Synthesis (LDS) module synthesizes generalized forces, which are made physically coherent by a novel Energy Consistency Loss (ECLoss) that enforces the work-energy theorem. The synthesized dynamic features are then leveraged via Spatio-Temporal Modulation (STM) to enhance action discriminability through direct spatial fusion and promote precise boundary localization via hierarchical temporal gating. Extensive experiments on six benchmarks demonstrated that LaDy achieves state-of-the-art performance. Ablation studies and visualizations further confirmed the efficacy and physical interpretability of our approach, demonstrating that integrating physical dynamics provides a more discriminative and precise foundation for segmentation.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2025YFE0218500; in part by the National Natural Science Foundation of China under Grant 62503139, Grant 62573163, and Grant 62521006; in part by Shenzhen Science and Technology Program under Grant JCYJ20240813105137049; and in part by the GuangDong Basic and Applied Basic Research Foundation under Grant 2024A1515012028, and Grant 2026A1515011822; and in part by the Shenzhen Medical Research Fund under Grant A2502034.

References

[1] Nicolas Aziere, Tieqiao Wang, and Sinisa Todorovic. Difformer for action segmentation. In *ICCV*, pages 2986–2995, 2025. 2

[2] Emad Bahrami, Gianpiero Francesca, and Juergen Gall. How much temporal long-term context is needed for action segmentation? In *ICCV*, pages 10351–10361, 2023. 2

[3] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *ECCV*, pages 52–68, 2022. 2

[4] Shurong Chai, Rahul Kumar Jain, Jiaqing Liu, Shiyu Teng, Tomoko Tateyama, Yinhao Li, and Yen-Wei Chen. A motion-aware and temporal-enhanced spatial–temporal graph convolutional network for skeleton-based human action segmentation. *Neurocomputing*, 580:127482, 2024. 2, 3, 7

[5] Bowen Chen, Weihong Ren, Honghai Liu, Huiping Li, Xiu Xu, Bingrui Zhou, and Qiong Xu. Autoenp: An auto rating pipeline for expressing needs via pointing protocol. In *ICPR*, pages 3280–3286, 2022. 1

[6] Bowen Chen, Haoyu Ji, Zhiyong Wang, Benjamin Filtjens, Chunzhuo Wang, Weihong Ren, Bart Vanrumste, and Honghai Liu. Language-assisted human part motion learning for skeleton-based temporal action segmentation. *arXiv preprint arXiv:2410.06353*, 2024. 2, 7

[7] Bowen Chen, Haoyu Ji, Hanwei Ma, Ruihan Lin, Wei Nie, Weihong Ren, Zhiyong Wang, and Honghai Liu. Exploring supervised contrastive learning for skeleton-based temporal action segmentation. *IEEE TCDS*, 2025. 2, 7

[8] Bowen Chen, Wei Nie, Haoyu Ji, Weihong Ren, Qiyi Tong, Zhiyong Wang, and Honghai Liu. Multiscale skeleton-based temporal action segmentation using hierarchical temporal modeling and prediction ensemble. *IEEE TCYB*, 2025. 2, 7

[9] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 6

[10] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR Workshop on Integration of Deep Neural Models and Differential Equations*, 2020. 2

[11] Dazhao Du, Bing Su, Yu Li, Zhongang Qi, Lingyu Si, and Ying Shan. Do we really need temporal convolutions in action segmentation? In *ICME*, pages 1014–1019, 2023. 2

[12] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, pages 3575–3584, 2019. 1, 2, 6, 7

[13] Benjamin Filtjens, Pieter Ginis, Alice Nieuwboer, Peter Slaets, and Bart Vanrumste. Automated freezing of gait assessment with marker-based motion capture and multi-stage spatial-temporal graph convolutional neural networks. *Journal of NeuroEngineering and Rehabilitation*, 19(1):48, 2022. 1

[14] Benjamin Filtjens, Bart Vanrumste, and Peter Slaets. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*, 2022. 1, 2, 7

[15] Shanghua Gao, Zhong-Yu Li, Qi Han, Ming-Ming Cheng, and Liang Wang. Rf-next: Efficient receptive field search for convolutional neural networks. *IEEE TPAMI*, 45(3):2984–3002, 2022. 2

[16] Shang-Hua Gao, Qi Han, Zhong-Yu Li, Pai Peng, Liang Wang, and Ming-Ming Cheng. Global2local: Efficient structure search for video action segmentation. In *CVPR*, pages 16805–16814, 2021. 2

[17] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *NeurIPS*, 32, 2019. 2

[18] Hongji Guo and Qiang Ji. Physics-augmented autoencoder for 3d skeleton-based gait recognition. In *ICCV*, pages 19627–19638, 2023. 2

[19] Xiaofeng Guo, Yang Mo, Yaonan Wang, and Qing Zhu. Skeleton-based traffic gesture recognition via a motion-guided 2s-gcn in autonomous driving: A review of challenges and opportunities. *IEEE Intelligent Systems*, 2024. 1

[20] Wenhui Huang, Yunhan Lin, Mingxin Liu, and Huasong Min. Velocity-aware spatial-temporal attention lstm model for inverse dynamic model learning of manipulators. *Frontiers in Neurorobotics*, 18:1353879, 2024. 2

[21] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, pages 14024–14034, 2020. 2

[22] Rongtian Huo, Junkang Chen, Ye Zhang, and Qing Gao. 3d skeleton aware driver behavior recognition framework for autonomous driving system. *Neurocomputing*, 613:128743, 2025. 1

[23] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *WACV*, pages 2322–2331, 2021. 2, 3, 6, 7

[24] Haoyu Ji, Bowen Chen, Xinglong Xu, Weihong Ren, Zhiyong Wang, and Honghai Liu. Language-assisted skeleton action understanding for skeleton-based temporal action segmentation. In *ECCV*, pages 400–417. Springer, 2024. 1, 2, 3, 6, 7

[25] Haoyu Ji, Bowen Chen, Wenze Huang, Weihong Ren, Zhiyong Wang, and Honghai Liu. Snippet-aware transformer with multiple action elements for skeleton-based action segmentation. *IEEE TNNLS*, 2025. 1, 2, 7

[26] Haoyu Ji, Bowen Chen, Weihong Ren, Wenze Huang, Zhihao Yang, Zhiyong Wang, and Honghai Liu. Text-derived relational graph-enhanced network for skeleton-based action segmentation. *arXiv preprint arXiv:2503.15126*, 2025. 2, 3, 6

[27] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pages 5156–5165, 2020. 3

[28] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, pages 156–165, 2017. 1, 2

[29] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE TPAMI*, 2020.

[30] Yunheng Li, Zhuben Dong, Kaiyuan Liu, Lin Feng, Lianyu Hu, Jie Zhu, Li Xu, Shenglan Liu, et al. Efficient two-step networks for temporal action segmentation. *Neurocomputing*, 454:373–381, 2021. 2

[31] Yuheng Li, Zhongyu Li, Shanghua Gao, Qilong Wang, Hou Qibin, and Cheng Mingming. A decoupled spatio-temporal framework for skeleton-based action segmentation. *arXiv preprint arXiv:2312.05830*, 2023. 1, 2, 3, 6, 7

[32] Yun-Heng Li, Kai-Yuan Liu, Sheng-Lan Liu, Lin Feng, and Hong Qiao. Involving distinguished temporal graph convolutional networks for skeleton-based temporal action segmentation. *IEEE TCSVT*, 2024. 1, 2, 7

[33] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for skeleton-based human action understanding. In *ACM VSCC*, pages 1–8, 2017. 2, 6

[34] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *ICCV*, pages 10139–10149, 2023. 2

[35] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffact++: Diffusion action segmentation. *IEEE TPAMI*, 2024. 2

[36] Kaiyuan Liu, Yunheng Li, Yuanfeng Xu, Shuai Liu, and Shenglan Liu. Spatial focus attention for fine-grained skeleton-based action tasks. *IEEE Signal Processing Letters*, 29:1883–1887, 2022. 1, 2

[37] Shenglan Liu, Aibin Zhang, Yunheng Li, Jian Zhou, Li Xu, Zhuben Dong, and Renhao Zhang. Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset. In *AAAI*, pages 2163–2171, 2021. 2, 6

[38] Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *CVPR*, pages 18175–18185, 2024. 2

[39] Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks: Using physics as model prior for deep learning. In *ICLR*, 2019. 2

[40] Friedrich Niemann, Christopher Reining, Fernando Moya Rueda, Nilah Ravi Nair, Janine Anika Steffens, Gernot A Fink, and Michael Ten Hompel. Lara: Creating a dataset for human activity recognition in logistics using semantic attributes. *Sensors*, 20(15):4083, 2020. 2, 6

[41] Junyong Park, Daekyum Kim, Sejoon Huh, and Sungho Jo. Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction. *Pattern Recognition*, 129:108764, 2022. 2

[42] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *CVPR*, pages 1905–1916, 2025. 2

[43] Dipika Singhania, Rahul Rahaman, and Angela Yao. C2f-tcn: A framework for semi-and fully-supervised temporal action segmentation. *IEEE TPAMI*, 2023. 2

[44] Hiromu Taketsugu, Takeru Oba, Takahiro Maeda, Shohei Nobuhara, and Norimichi Ukita. Physical plausibility-aware trajectory prediction via locomotion embodiment. In *CVPR*, pages 12324–12334, 2025. 2

[45] Xiaoyan Tian, Ye Jin, Zhao Zhang, Peng Liu, and Xianglong Tang. Stga-net: Spatial-temporal graph attention network for skeleton-based temporal action segmentation. In *ICMEW*, pages 218–223, 2023. 1, 2

[46] Dong Wang, Di Hu, Xingjian Li, and Dejing Dou. Temporal relational modeling with self-supervision for action segmentation. In *AAAI*, pages 2729–2737, 2021. 2

[47] Hongsong Wang, Wanjiang Weng, Junbo Wang, Fang Zhao, Guo-Sen Xie, Xin Geng, and Liang Wang. Foundation model for skeleton-based human action understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[48] Peiyao Wang, Yuewei Lin, Erik Blasch, Haibin Ling, et al. Efficient temporal action segmentation via boundary-aware query voting. *NeurIPS*, 37:37765–37790, 2024. 2

[49] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *ECCV*, pages 34–51, 2020. 2

[50] Julian Wiederer, Arij Bouazizi, Ulrich Kressel, and Vasileios Belagiannis. Traffic control gesture recognition for autonomous vehicles. In *IROS*, pages 10676–10683, 2020. 2, 6

[51] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *ICCV*, pages 11532–11541, 2021. 2

[52] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 537: 164–186, 2023. 1

[53] Leiyang Xu, Qiang Wang, Xiaotian Lin, and Lin Yuan. An efficient framework for few-shot skeleton-based temporal action segmentation. *Computer Vision and Image Understanding*, 232:103707, 2023. 2

[54] Leiyang Xu, Qiang Wang, Xiaotian Lin, Lin Yuan, and Xiang Ma. Skeleton-based tai chi action segmentation using trajectory primitives and content. *Neural Computing and Applications*, 35(13):9549–9566, 2023. 2

[55] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 1

[56] Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Lac-latent action composition for skeleton-based action segmentation. In *ICCV*, pages 13679–13690, 2023. 2

[57] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *BMVC*, 2021. 1, 2, 7

[58] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, pages 16010–16021, 2023. 2

[59] Junbin Zhang, Pei-Hsuan Tsai, and Meng-Hsun Tsai. Semantic2graph: graph-based multi-modal feature fusion for action segmentation in videos. *Applied Intelligence*, 54(2): 2084–2099, 2024. 2

[60] Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *CVPR*, pages 2305–2317, 2024. 2

[61] Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *WACV*, pages 6164–6174, 2024. 2

[62] Zhibo Zhang, Yanjun Zhu, Rahul Rai, and David Doermann. Pimnet: Physics-infused neural network for human motion prediction. *IEEE RAL*, 7(4):8949–8955, 2022. 2

[63] Rongyong Zhao, Bingyu Wei, Lingchen Han, Yuxin Cai, Yunlong Ma, and Cuiling Li. A multiscale mixed-graph neural network based on kinematic and dynamic joint features for human motion prediction. *Applied Sciences*, 15(4):1897, 2025. 2

[64] Yaofeng Desmond Zhong and Naomi Leonard. Unsupervised learning of lagrangian dynamics from images for prediction and control. *NeurIPS*, 33:10741–10752, 2020. 2