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

Supplementary Material

This supplementary material provides a comprehensive exposition of the theoretical underpinnings, implementation specifics, and extended empirical validation of the proposed Lagrangian-Dynamic Informed Network (LaDy). We commence by establishing the rigorous physical framework in **Sec. 6**, detailing the Euler-Lagrange derivation and the axiomatic basis of the Energy Consistency Loss. **Sec. 7** elaborates on the architectural specifications of the backbone modules and loss formulations, followed by **Sec. 8**, which provides an exhaustive account of experimental protocols to ensure reproducibility. To further substantiate our claims, **Sec. 9** presents granular comparative analyses, encompassing performance-efficiency trade-offs, boundary precision, per-class breakdowns, extended qualitative evaluations, and robustness to input perturbations. **Sec. 10** then offers extensive ablation studies scrutinizing the internal mechanisms of the dynamics synthesis, energy-based loss, and modulation modules. Finally, **Sec. 11** concludes with a broader discussion exploring the framework's generalization capabilities, alongside a critical assessment of failure cases, intrinsic limitations, and potential avenues for future research.

6. Theoretical Foundations of Dynamic Model

In this section, we provide the rigorous physical framework underpinning our model (LaDy). Moving beyond black-box approximations, we first derive the governing dynamic equations via the Euler-Lagrange formulation (Sec. 6.1), establishing a structurally explicit basis for force synthesis. We then articulate the Work-Energy Theorem (Sec. 6.2), which serves as the axiomatic foundation for our energy consistency supervision. Subsequently, we formalize the fundamental physical constraints—specifically the positive definiteness of inertia and the skew-symmetry of the Coriolis matrix—that our network architecture is strictly constrained to satisfy (Sec. 6.3). Finally, we define the kinematic mapping that theoretically bridges the observable Cartesian inputs with the generalized coordinates utilized in our dynamic formulation (Sec. 6.4).

6.1. Derivation of Lagrange Dynamic Equation

The dynamics of a complex, articulated system (like a human skeleton) can be elegantly described using the Euler-Lagrange formulation. This approach is based on the system's scalar energy functions rather than vector-based Newtonian forces.

Definition 1 (The Lagrangian)**.** *The Lagrangian $\mathcal{L}$ of a mechanical system is defined as the difference between its total kinetic energy (E_K) and its total potential energy (E_P).*

$$\mathcal{L}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = E_K(\boldsymbol{q}, \dot{\boldsymbol{q}}) - E_P(\boldsymbol{q}), \tag{17}$$

where $\boldsymbol{q} \in \mathbb{R}^D$ are the generalized coordinates, and $\dot{\boldsymbol{q}} \in \mathbb{R}^D$ are the generalized velocities.

For a rigid-body system with D degrees of freedom, the kinetic and potential energies are given by:

$$E_K(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \frac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}, \tag{18}$$

$$E_P(\boldsymbol{q}) = E_g(\boldsymbol{q}), \tag{19}$$

where E_g is defined as the gravitational potential energy, and $\boldsymbol{M}(\boldsymbol{q}) \in \mathbb{R}^{D\times D}$ is the symmetric, positive-definite mass-inertia matrix.

The **Euler-Lagrange Equation** provides the equations of motion by stating that the path taken by the system minimizes the "action" (the integral of the Lagrangian over time). For a system subject to non-conservative generalized forces $\boldsymbol{\tau}_{nc}$ (which include actuation torques $\boldsymbol{\tau}$ and friction/external forces $\boldsymbol{F}$), the equation is:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{q}}}\right) - \frac{\partial \mathcal{L}}{\partial \boldsymbol{q}} = \boldsymbol{\tau}_{nc} = \boldsymbol{\tau} - \boldsymbol{F}. \tag{20}$$

We derive the equation of motion by computing each term in Eq. (20):

Momentum Term: The derivative of $\mathcal{L}$ with respect to velocity $\dot{\boldsymbol{q}}$:

$$\frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{q}}} = \frac{\partial}{\partial \dot{\boldsymbol{q}}}\left(\frac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}} - E_P(\boldsymbol{q})\right) = \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}. \tag{21}$$

Time Derivative of Momentum: Taking the total time derivative:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{q}}}\right) = \frac{d}{dt}(\boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}) = \dot{\boldsymbol{M}}(\boldsymbol{q})\dot{\boldsymbol{q}} + \boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}}. \tag{22}$$

Lagrangian Gradient Term: The derivative of $\mathcal{L}$ with respect to position $\boldsymbol{q}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{q}} = \frac{\partial E_K}{\partial \boldsymbol{q}} - \frac{\partial E_P}{\partial \boldsymbol{q}} = \frac{\partial}{\partial \boldsymbol{q}}\left(\frac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}\right) - \boldsymbol{G}(\boldsymbol{q}), \tag{23}$$

where we define the gravitational vector $\boldsymbol{G}(\boldsymbol{q}) = \frac{\partial E_P(\boldsymbol{q})}{\partial \boldsymbol{q}}$. Substituting these back into Eq. (20):

$$\left(\dot{\boldsymbol{M}}(\boldsymbol{q})\dot{\boldsymbol{q}} + \boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}}\right) - \left(\frac{\partial E_K}{\partial \boldsymbol{q}} - \boldsymbol{G}(\boldsymbol{q})\right) = \boldsymbol{\tau} - \boldsymbol{F}. \tag{24}$$

Rearranging the terms yields:

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \left(\dot{\boldsymbol{M}}(\boldsymbol{q})\dot{\boldsymbol{q}} - \frac{\partial E_K}{\partial \boldsymbol{q}}\right) + \boldsymbol{G}(\boldsymbol{q}) = \boldsymbol{\tau} - \boldsymbol{F}. \quad (25)$$

It is a standard result in robotics to define the **Coriolis and Centrifugal Matrix** $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ such that:

$$\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} = \dot{\boldsymbol{M}}(\boldsymbol{q})\dot{\boldsymbol{q}} - \frac{\partial E_K}{\partial \boldsymbol{q}}. \quad (26)$$

This allows us to write the dynamics in the canonical form (here $\boldsymbol{F}$ represents all non-conservative forces like friction and external forces, which our model learns as $F(q, \dot{q})$):

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{G}(\boldsymbol{q}) + \boldsymbol{F} = \boldsymbol{\tau}. \quad (27)$$

This derivation formally establishes the origins of the dynamic equation $\boldsymbol{\tau} = \boldsymbol{M}\ddot{\boldsymbol{q}} + \boldsymbol{C}\dot{\boldsymbol{q}} + \boldsymbol{G} + \boldsymbol{F}$ that our LaDy model synthesizes, where $\boldsymbol{\tau}$ represents the generalized actuation forces.

6.2. The Work-Energy Theorem and Power Balance

Following the derivation of the dynamic equation (Eq. (27)), we introduce the fundamental physical principle the system must obey: the Work-Energy Theorem.

Theorem 1 (The Work-Energy Theorem)**.** *The change in a system's kinetic energy, ΔE_K, over a time interval $[t_1, t_2]$ is equal to the total work, W, done on the system by the net generalized forces $\boldsymbol{\tau}_{net}$ during that interval.*

$$\Delta E_K = E_K(t_2) - E_K(t_1) = \int_{t_1}^{t_2} P_{net}(t)dt = W, \quad (28)$$

where $P_{net}(t) = \dot{\boldsymbol{q}}(t)^T \boldsymbol{\tau}_{net}(t)$ is the instantaneous power.

In our system, this net generalized force $\boldsymbol{\tau}_{net}$ comprises all forces that alter the kinetic state: $\boldsymbol{\tau}_{net} = \boldsymbol{\tau} - \boldsymbol{G} - \boldsymbol{F}$. Per the dynamics equation (Eq. (27)), this is equivalent to:

$$\boldsymbol{\tau}_{net} = \boldsymbol{\tau} - \boldsymbol{G} - \boldsymbol{F} = \boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}}. \quad (29)$$

Therefore, the differential form of the Work-Energy Theorem, which represents the system's **Power Balance**, is the fundamental principle:

$$P_{net}(t) = \dot{E}_K(t). \quad (30)$$

Implication for LaDy: This theorem provides the physical justification for the Energy Consistency Loss $\mathcal{L}_{EC}$ (Sec. 3.3). Our loss function is built upon this axiom. By minimizing the residual between the discrete-time computations of ΔE_K and W, $\mathcal{L}_{EC}$ enforces that all synthesized dynamic components $(\boldsymbol{M}, \boldsymbol{C}, \boldsymbol{G}, \boldsymbol{F}, \boldsymbol{\tau})$ are mutually consistent and collectively obey this fundamental law of physics.

6.3. Fundamental Properties of the Dynamic Model

The estimators in our LDS module (Sec. 3.2.2) are constrained to respect fundamental physical properties. We formalize these properties below.

6.3.1. Positive Definiteness of Inertia Matrix

Proposition 1. *The inertia matrix $\boldsymbol{M}(\boldsymbol{q})$ is symmetric and positive definite (SPD), i.e., $\boldsymbol{M}(\boldsymbol{q}) = \boldsymbol{M}(\boldsymbol{q})^T$ and $\boldsymbol{x}^T\boldsymbol{M}(\boldsymbol{q})\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$.*

Proof. Symmetry $\boldsymbol{M}(\boldsymbol{q}) = \boldsymbol{M}(\boldsymbol{q})^T$ arises from the definition of the kinetic energy quadratic form (Eq. (18)). Positive definiteness is a direct physical constraint. The kinetic energy E_K of a physical system must be non-negative, and it can only be zero if the system is at rest (zero velocity).

From Eq. (18), we have $E_K = \frac{1}{2}\dot{\boldsymbol{q}}^T\boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}$. By physical principle, $E_K \geq 0$ for any possible velocity $\dot{\boldsymbol{q}}$. Furthermore, $E_K = 0$ if and only if $\dot{\boldsymbol{q}} = \boldsymbol{0}$. This is the mathematical definition of the matrix $\boldsymbol{M}(\boldsymbol{q})$ being positive definite. This property is crucial as it guarantees that any non-zero motion $\dot{\boldsymbol{q}}$ corresponds to positive kinetic energy. □

Implication for LaDy: Our model (Sec. 3.2.2) enforces this physical axiom by parameterizing $\boldsymbol{M}(\boldsymbol{q})$ via its Cholesky decomposition, $\boldsymbol{M}(\boldsymbol{q}) = \boldsymbol{L}(\boldsymbol{q})\boldsymbol{L}(\boldsymbol{q})^T$. The $\mathcal{F}_M$ estimator predicts the elements of $\boldsymbol{L}(\boldsymbol{q})$ and ensures its diagonal entries are strictly positive using a softplus function and a small constant ϵ. This construction mathematically guarantees that the synthesized $\boldsymbol{M}(\boldsymbol{q})$ is SPD.

6.3.2. Passivity and the Skew-Symmetric Property

A core property of rigid-body dynamics is passivity: the internal forces (inertia and Coriolis/centrifugal) do not generate or dissipate energy. This is captured by a key relationship between $\boldsymbol{M}$ and $\boldsymbol{C}$.

Proposition 2. *For a valid Coriolis matrix $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ derived from the Lagrangian formulation, the matrix $N = \dot{\boldsymbol{M}}(\boldsymbol{q}) - 2\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is skew-symmetric, i.e., $\boldsymbol{N} = -\boldsymbol{N}^T$.*

Proof. We analyze the power balance of the system. The rate of change of kinetic energy, $\dot{E}_K$, must be equal to the net power, P_{net}, delivered by the net generalized forces responsible for that change. As established in Eq. (29), this net generalized force is $\tau_{net} = \boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}}$. The instantaneous net power P_{net} is:

$$P_{net} = \dot{\boldsymbol{q}}^T\tau_{net} = \dot{\boldsymbol{q}}^T(\boldsymbol{M}\ddot{\boldsymbol{q}} + \boldsymbol{C}\dot{\boldsymbol{q}}). \quad (31)$$

Separately, we can find the rate of change of kinetic energy, $\dot{E}_K$, by taking the time derivative of Eq. (18):

$$\begin{aligned} \dot{E}_K &= \frac{d}{dt}\left(\frac{1}{2}\dot{\boldsymbol{q}}^T\boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{q}}\right) \\ &= \frac{1}{2}\left(\ddot{\boldsymbol{q}}^T\boldsymbol{M}\dot{\boldsymbol{q}} + \dot{\boldsymbol{q}}^T\dot{\boldsymbol{M}}\dot{\boldsymbol{q}} + \dot{\boldsymbol{q}}^T\boldsymbol{M}\ddot{\boldsymbol{q}}\right). \end{aligned} \quad (32)$$

Since $\boldsymbol{M}$ is symmetric, $\ddot{\boldsymbol{q}}^T \boldsymbol{M} \dot{\boldsymbol{q}} = (\boldsymbol{M}\ddot{\boldsymbol{q}})^T \dot{\boldsymbol{q}} = \dot{\boldsymbol{q}}^T (\boldsymbol{M}\ddot{\boldsymbol{q}})$.

$$\dot{E}_K = \dot{\boldsymbol{q}}^T (\boldsymbol{M}\ddot{\boldsymbol{q}}) + \frac{1}{2}\dot{\boldsymbol{q}}^T \dot{\boldsymbol{M}} \dot{\boldsymbol{q}}. \tag{33}$$

From the power-balance principle (Eq. (30)), the net power P_{net} equals the rate of change of kinetic energy $\dot{E}_K$:

$$\begin{aligned} P_{net} &= \dot{E}_K, \\ \dot{\boldsymbol{q}}^T (\boldsymbol{M}\ddot{\boldsymbol{q}} + \boldsymbol{C}\dot{\boldsymbol{q}}) &= \dot{\boldsymbol{q}}^T (\boldsymbol{M}\ddot{\boldsymbol{q}}) + \frac{1}{2}\dot{\boldsymbol{q}}^T \dot{\boldsymbol{M}} \dot{\boldsymbol{q}}. \end{aligned} \tag{34}$$

Canceling the $\dot{\boldsymbol{q}}^T (\boldsymbol{M}\ddot{\boldsymbol{q}})$ terms, we are left with:

$$\dot{\boldsymbol{q}}^T \boldsymbol{C} \dot{\boldsymbol{q}} = \frac{1}{2}\dot{\boldsymbol{q}}^T \dot{\boldsymbol{M}} \dot{\boldsymbol{q}}. \tag{35}$$

This can be rewritten as:

$$\dot{\boldsymbol{q}}^T \left(\dot{\boldsymbol{M}} - 2\boldsymbol{C} \right) \dot{\boldsymbol{q}} = 0. \tag{36}$$

For this quadratic form to be zero for *any* velocity vector $\dot{\boldsymbol{q}}$, the matrix $\boldsymbol{N} = (\dot{\boldsymbol{M}} - 2\boldsymbol{C})$ must be skew-symmetric ($\boldsymbol{N} = -\boldsymbol{N}^T$). □

Implication for LaDy: This property is fundamental. It ensures that the Coriolis forces are non-dissipative. As detailed in Sec. 3.2.2, our model enforces this by parameterizing $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})$ directly. We define $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = 0.5(\dot{\boldsymbol{M}}(\boldsymbol{q}) - N(\boldsymbol{q}, \dot{\boldsymbol{q}}))$, where $N(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is an explicitly learned skew-symmetric matrix. $\dot{\boldsymbol{M}}(\boldsymbol{q})$ is itself approximated via finite differences. This construction mathematically guarantees the passivity property is satisfied.

6.3.3. Supplement: Christoffel Construction of C

While our LaDy model uses a direct parameterization to enforce passivity (Prop. 2), it is theoretically insightful to know that $\boldsymbol{C}$ is not unique. A standard (but not the only) method for constructing a $\boldsymbol{C}$ matrix that satisfies the passivity property is by using the **Christoffel symbols of the first kind**, Γ_{ijk}:

$$\Gamma_{ijk}(\boldsymbol{q}) = \frac{1}{2}\left(\frac{\partial M_{ij}}{\partial q_k} + \frac{\partial M_{ik}}{\partial q_j} - \frac{\partial M_{jk}}{\partial q_i} \right). \tag{37}$$

The elements of the $\boldsymbol{C}$ matrix can then be defined as:

$$C_{ij}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \sum_{k=1}^{D} \Gamma_{ijk}(\boldsymbol{q}) \dot{q}_k. \tag{38}$$

This specific construction of $\boldsymbol{C}$ can be shown to satisfy $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} = \dot{\boldsymbol{M}}(\boldsymbol{q})\dot{\boldsymbol{q}} - \frac{\partial E_K}{\partial \boldsymbol{q}}$ and also guarantees the skew-symmetric property of $\boldsymbol{N} = \dot{\boldsymbol{M}} - 2\boldsymbol{C}$. Our model's direct parameterization of $\boldsymbol{N}$ is a more computationally tractable approach for a deep learning context, while arriving at the same fundamental physical constraint.

6.4. Cartesian to Generalized Coordinates

While the dynamics (Sec. 6.1) are formulated in generalized coordinates $\boldsymbol{q}$, the input to our system (Sec. 3.1) is in Cartesian (world) coordinates $\boldsymbol{p} \in \mathbb{R}^{3V}$. The mapping between these spaces is defined by the system's kinematics.

Definition 2 (Forward Kinematics)**.** *Forward kinematics (FK) is the function f that maps the D generalized coordinates $\boldsymbol{q}$ to the Cartesian positions of all V points in the system.*

$$\boldsymbol{p} = f(\boldsymbol{q}). \tag{39}$$

Implication for LaDy: This theoretical framework provides the justification for our "Generalized Coordinates Computation" module (Sec. 3.2.1). That module implements a direct and differentiable geometric computation based on the system's open kinematic chain. It analytically computes the generalized coordinates $q(t)$ (internal joint angles) from the Cartesian positions $\boldsymbol{p}(t)$ (external world positions). This computation is, in effect, a precise analytical solution to the Inverse Kinematics (IK) problem ($q = f^{-1}(\boldsymbol{p})$) tailored for the defined skeletal structure. Subsequently, by applying finite differences to the $q(t)$ sequence (as detailed in Sec. 3.2.1), we obtain a practical approximation of the generalized velocities $\dot{\boldsymbol{q}}(t)$ and accelerations $\ddot{\boldsymbol{q}}(t)$. This entire process grounds our abstract Lagrangian dynamics in the observable Cartesian input data.

7. Method Supplement

This supplement provides a detailed exposition of the baseline architectural components employed in LaDy, which serve as the foundation for our novel dynamics-informed modules. We elaborate on the spatial modeling pipeline (Sec. 7.1), the temporal modeling architecture (Sec. 7.2), the multi-stage prediction refinement branches (Sec. 7.3), and the formulation of the standard loss functions (Sec. 7.4).

7.1. Spatial Modeling Supplement

In the Spatial Model, the input skeleton sequence $X \in \mathbb{R}^{C_0 \times T \times V}$ is processed via multi-scale and adaptive Graph Convolutional Networks (GCNs) to extract the final kinematic feature $F_{kin} \in \mathbb{R}^{C \times T \times V}$.

Multi-Scale GCN. We adopt the multi-scale GCN mechanism [9] to capture multi-hop skeletal connectivity. We first define a set of k-adjacency matrices $A^k \in \{0,1\}^{V \times V}$, where $A^k_{ij} = 1$ if the shortest path distance between joints i and j is k, or if $i = j$. A unified multi-scale adjacency matrix $A^{MS} \in \{0,1\}^{V \times KV}$ (for K scales) is constructed by concatenating all normalized k-adjacency matrices:

$$\begin{aligned} A^{MS} = &[(\tilde{D}^1)^{-\frac{1}{2}} A^1 (\tilde{D}^1)^{-\frac{1}{2}}] \oplus \cdots \\ &\oplus [(\tilde{D}^K)^{-\frac{1}{2}} A^K (\tilde{D}^K)^{-\frac{1}{2}}], \end{aligned} \tag{40}$$

where $\oplus$ denotes concatenation and $\tilde{D}^k$ is the diagonal degree matrix for normalization. An initial spatial feature F_{ms} is produced via:

$$F_{ms} = \text{ReLU}\left(\text{reshape}\left[(A^{MS} + B) \cdot X\right] \cdot W_{ms}\right), \quad (41)$$

where $B \in \mathbb{R}^{V \times KV}$ is a learnable matrix encoding non-local correlations, $\text{reshape}(\cdot)$ transforms the feature to $\mathbb{R}^{KC_0 \times T \times V}$, and W_{ms} is a point-wise convolution for channel projection to C.

Adaptive GCNs. To further capture fine-grained, dynamic inter-joint dependencies, we employ adaptive GCNs [1], which compute feature-adaptive graph structures. Given F_{ms}, two parallel convolutional heads produce intermediate embeddings $P, Q \in \mathbb{R}^{C_1 \times T \times V}$. Temporal-wise ($G^T \in \mathbb{R}^{T \times V \times V}$) and channel-wise ($G^C \in \mathbb{R}^{C \times V \times V}$) adaptive graphs are computed by pooling P, Q and measuring pairwise joint dissimilarities:

$$G^T_{t,i,j} = P^T_{t,i} - Q^T_{t,j}, \quad G^C_{c,i,j} = P^C_{c,i} - Q^C_{c,j}. \quad (42)$$

Furthermore, inspired by [7], we construct a static Text-derived Joint Graph $G^{txt} \in \mathbb{R}^{V \times V}$ using the inverse-normalized L2 distances of BERT [3]-encoded joint descriptions. This semantic prior G^{txt} is broadcasted and added element-wise to both G^T and G^C. The final kinematic feature $F_{kin} \in \mathbb{R}^{C \times T \times V}$ is computed by modulating a refined input feature $F_{as} \in \mathbb{R}^{C \times T \times V}$ (derived from F_{ms}) with these enriched graphs:

$$F_{kin} = \text{ReLU}\left(\text{BN}\left[F_{as}G^T + F_{as}G^C\right]\right) + F_{ms}, \quad (43)$$

where the products (e.g., $F_{as}G^T$ and $F_{as}G^C$) denote batched matrix multiplication over the joint axis, enabling the model to capture frame- and channel-specific spatial dependencies.

7.2. Temporal Modeling Supplement

As outlined in Sec. 3.1 and Fig. 2, each of the L temporal stages comprises three main components: a Linear Transformer, an Adaptive Fusion module, and our novel Temporal Modulation (detailed in Sec. 3.2).

Linear Transformer for Temporal Interaction. To efficiently model global temporal context, we adopt the Linear Transformer [8], which reduces attention complexity from $\mathcal{O}(T^2)$ to $\mathcal{O}(T)$. At stage l, given the input feature $\tilde{H}^{(l-1)} \in \mathbb{R}^{C \times T}$ from the previous stage's modulation (or $H^{(0)}$ for $l = 1$), the Linear Transformer computes the feature $H^{(l)}_{LT}$:

$$Q^l_t = W^l_Q \tilde{H}^{(l-1)}, \; K^l_t = W^l_K \tilde{H}^{(l-1)}, \; V^l_t = W^l_V \tilde{H}^{(l-1)},$$
$$H^{(l)}_{LT} = \text{ReLU}[\phi(Q^l_t)\left(\phi(K^l_t)^\top V^l_t\right) \cdot W^l_t + \tilde{H}^{(l-1)}], \quad (44)$$

where W_Q, W_K, W_V, W_t are linear layers and $\phi(\cdot)$ is the sigmoid activation.

Adaptive Feature Fusion. To ensure core spatial information is retained at each temporal scale, each stage l also receives input from a Spatial-Channel Fusion head. This head processes the main spatial representation $F_{sp} \in \mathbb{R}^{2C \times T \times V}$ using a point-wise convolution, a reshape operation (merging V and C dimensions), and another point-wise convolution to produce a skip-connection feature $F^l_T \in \mathbb{R}^{C \times T}$. The Adaptive Fusion module then integrates this feature with the Linear Transformer output $H^{(l)}_{LT}$ to produce the fused temporal feature $H^{(l)}_T$:

$$H^{(l)}_T = \text{GELU}\left[\left(F^l_T \oplus H^{(l)}_{LT}\right) \cdot W_f \cdot W_l\right] + H^{(l)}_{LT}, \quad (45)$$

where $\oplus$ denotes channel-wise concatenation and W_f, W_l are point-wise convolutions. This $H^{(l)}_T$ is the feature subsequently passed to our Temporal Modulation module (Sec. 3.4).

7.3. Prediction Refinement Supplement

LaDy follows the standard prediction refinement paradigm, comprising two complementary branches.

Classification Prediction Branch. The Classification Head produces an initial frame-wise class prediction $Y^0_c \in \mathbb{R}^{Q \times T}$ from the final temporal representation F_R. This is refined through S_c stages. At stage h, the previous prediction Y^{h-1}_c is processed by a stack of Linear Transformer layers (using cross-attention, where Q, K are from Y^{h-1}_c and V is from F_R) to produce a more precise prediction Y^h_c. The final stage yields the refined class prediction Y^F_c.

Boundary Prediction Branch. Similarly, the Boundary Head produces an initial boundary confidence map $Y^0_b \in \mathbb{R}^{1 \times T}$ from F_R. This is refined across S_b stages. At each stage h, the previous output Y^{h-1}_b is processed by a stack of dilated 1D TCNs to yield the refined boundary prediction Y^h_b. The final output is Y^F_b.

7.4. Loss Function Supplement

The overall training objective $\mathcal{L}_{total}$ is a composite loss: $\mathcal{L}_{total} = \mathcal{L}_{as} + \lambda_1\mathcal{L}_{br} + \lambda_2\mathcal{L}_{atc} + \lambda_3\mathcal{L}_{EC}$. Our proposed Energy Consistency Loss $\mathcal{L}_{EC}$ is detailed in Sec. 3.3. The formulations for the other three standard loss components are detailed below.

Action-Text Contrastive Loss ($\mathcal{L}_{atc}$). Following [6], we align visual features with textual embeddings. The final representation F_R is segmented based on ground-truth, and action-level visual features A^F are obtained via temporal mean pooling. Corresponding textual embeddings A^E are obtained using a pre-trained BERT [3]. We compute a pairwise cosine similarity matrix $S^A = \text{sim}(A^F, A^E)$ and apply bidirectional KL divergence loss against the ground-truth identity matrix S^{GT}:

$$\mathcal{L}_{atc} = \frac{1}{2}[\mathcal{D}_{KL}(S^{GT}|S^A_f) + \mathcal{D}_{KL}(S^{GT}|S^A_e)], \quad (46)$$

where S_f^A and S_e^A are row- and column-normalized similarity matrices, respectively. $\mathcal{D}_{KL}$ is the KL divergence.

Action Segmentation Loss ($\mathcal{L}_{as}$). This loss supervises the classification predictions Y_c at all stages. It combines a standard frame-wise cross-entropy loss $\mathcal{L}_{ce}$ and a Gaussian Similarity-weighted Truncated Mean Squared Error (GS-TMSE) smoothing loss $\mathcal{L}_{gs\text{-}tmse}$ [5, 9] to penalize over-segmentation:

$$\begin{aligned}\mathcal{L}_{as} &= \mathcal{L}_{ce} + \mathcal{L}_{gs\text{-}tmse} = -\frac{1}{T}\sum_{t}\log(\hat{y}_{t,\hat{c}}) \\ &+\frac{1}{TC}\sum_{t,c} e^{-\frac{\|x_t - x_{t-1}\|^2}{2\sigma^2}} \min(|\log(\frac{\hat{y}_{t,c}}{\hat{y}_{t-1,c}})|^2, \kappa),\end{aligned} \tag{47}$$

where $\hat{y}_t \in \mathbb{R}^C$ represents the predicted class probabilities at frame t, $\hat{y}_{t,\hat{c}}$ is the predicted probability for the ground-truth class $\hat{c}$, and x_t is the input feature at frame t. The GS-TMSE term dynamically scales the temporal smoothing penalty based on the Gaussian similarity of adjacent features. The parameter σ controls the similarity sensitivity, and κ is a predefined threshold that truncates excessively large gradients.

Boundary Prediction Loss ($\mathcal{L}_{br}$). This loss supervises the boundary predictions Y_b at all stages. It is a standard binary cross-entropy loss:

$$\mathcal{L}_{br} = -\frac{1}{T}\sum_{t}\Big(b_t\log(\hat{b}_t) + (1-b_t)\log(1-\hat{b}_t)\Big), \tag{48}$$

where $b_t \in \{0,1\}$ is the binary ground-truth boundary label at frame t, and $\hat{b}_t \in [0,1]$ is the predicted boundary probability.

8. Detailed Experimental Setup

This section provides a comprehensive account of the experimental configuration to ensure reproducibility. Supplementing the primary implementation details outlined in the main paper, we elaborate on dataset specifications, preprocessing protocols, evaluation metrics, training strategy, and specific hyperparameter settings adopted in the LaDy framework.

8.1. Datasets

We conduct extensive evaluations on six challenging benchmarks, covering diverse domains from daily activities to specialized sports and gestures.

PKU-MMD v2 [10] is a large-scale dataset captured via Kinect v2, containing approx. 50 hours of data with 52 action categories. It provides 3-axis spatial coordinates for 25 body joints. Following standard protocols, we evaluate on two splits: (1) *Cross-Subject (X-sub)*, with 775 training and 234 testing videos; and (2) *Cross-View (X-view)*, training on middle/right views and testing on the left.

LARa [12] focuses on logistics activities (e.g., picking, packing) recorded by an optical MoCap system. The high-fidelity markers provide 19 joint positions and orientations. It comprises 377 sequences across 8 classes (758 mins).

MCFS-22 & MCFS-130 [11] are sourced from the same figure skating repository but annotated at different granularities. Collected via OpenPose (30 Hz), this dataset contains 271 videos and provides 2D joint coordinates for 25 body joints. MCFS-130 provides fine-grained annotations for 130 atomic actions, while MCFS-22 aggregates them into 22 coarse categories. This dual-setting tests the model's capability in handling both semantic hierarchies.

TCG-15 [14] is a traffic control gesture dataset captured by IMU sensors at 100 Hz. It records 3-axis positions for 17 joints. It includes 550 sequences across 15 distinct gesture classes, totaling approx. 140 minutes of data.

8.2. Data Preprocessing

To ensure fair comparison and input consistency, we adhere to the standard preprocessing protocols established in prior studies [4, 6, 9].

- **Resampling:** To unify temporal resolution, LARa (originally 200 Hz) is downsampled to 50 Hz. PKU-MMD (50 Hz), MCFS (30 Hz), and TCG-15 (100 Hz) retain their native frame rates to preserve motion fidelity specific to their domains.
- **Feature Extraction:** We construct frame-level input vectors primarily using joint-wise relative coordinates and temporal displacements to ensure translation invariance.
  - For **PKU-MMD** and **TCG-15**, we extract 6-channel features (3D relative positions + 3D temporal differences) for each joint.
  - For **LARa**, we utilize 12-channel features, incorporating both 3D positions and 3D orientations along with their respective temporal derivatives.
  - For **MCFS**, given the 2D nature of OpenPose data, we construct 2-channel features using 2D relative coordinates.

8.3. Evaluation Metrics

Segmentation Metrics. We employ three standard metrics to assess segmentation quality:

(i) **Frame-wise Accuracy (Acc):** The percentage of frames correctly classified. While fundamental, it is insensitive to over-segmentation errors.

(ii) **Segmental Edit Score (Edit):** The normalized Levenshtein distance between predicted and ground-truth segment sequences. This metric explicitly penalizes ordering errors and over-segmentation.

(iii) **Segmental F1 Score (F1@k):** The harmonic mean of precision and recall for predicted segments that overlap with ground truth by an Intersection-over-Union (IoU) threshold k. We report F1 scores at

thresholds $k \in \{0.10, 0.25, 0.50\}$ to evaluate temporal boundary precision at varying strictness levels.

Clustering Metrics. To quantify the discriminability of the learned latent space (as visualized in Fig. 5 of the main paper), we utilize three intrinsic clustering indicators:

(i) **Silhouette Coefficient (SC):** Measures how similar a sample is to its own cluster (cohesion) compared to other clusters (separation). Values range from -1 to +1, where higher values indicate better-defined clusters.

(ii) **Calinski-Harabasz Index (CH):** The ratio of the sum of between-clusters dispersion to within-cluster dispersion. A higher CH score signifies dense and well-separated clusters.

(iii) **Davies-Bouldin Index (DB):** The average similarity measure of each cluster with its most similar cluster. Unlike SC and CH, a lower DB index indicates better separation and compactness.

8.4. Training Strategy

Loss Configuration. The total objective function is a weighted sum of four components: the segmentation loss $\mathcal{L}_{as}$, the boundary loss $\mathcal{L}_{br}$, the action-text contrastive loss $\mathcal{L}_{atc}$, and our proposed Energy Consistency loss $\mathcal{L}_{EC}$. The balancing hyperparameters are set as follows:

- $\lambda_1 = 1.0$ for $\mathcal{L}_{br}$, following the established settings in [5, 6, 9] to enforce boundary regression.
- $\lambda_2 = 0.8$ for $\mathcal{L}_{atc}$, consistent with language-assisted frameworks like [6, 7], ensuring semantic alignment.
- $\lambda_3 = 0.1$ for $\mathcal{L}_{EC}$, determined via empirical ablation to balance physical regularization with kinematic learning.

Delayed Physics Injection (Warmup Strategy). Applying the strict Energy Consistency constraint ($\mathcal{L}_{EC}$) from the initial iteration can destabilize training, as the dynamic estimators (e.g., Inertia estimator $\mathcal{F}_M$, Coriolis estimator $\mathcal{F}_C$) are initialized randomly and require time to converge to plausible physical manifolds. To mitigate this, we design a *Delayed and Phased Warmup* strategy:

$$\lambda_3^{(e)} = \begin{cases} 0 & e < \mathcal{Z}, \\ \frac{e-\mathcal{Z}}{\mathcal{Z}_w} \cdot \lambda_3 & \mathcal{Z} \le e < \mathcal{Z} + \mathcal{Z}_w, \\ \lambda_3 & e \ge \mathcal{Z} + \mathcal{Z}_w, \end{cases} \tag{49}$$

where e denotes the current training epoch, and $\lambda_3^{(e)}$ is the weighting coefficient for the $\mathcal{L}_{EC}$ at epoch e. The physical regularization is deactivated for the first $\mathcal{Z}$ epochs (post-start), allowing the model to first establish a coarse kinematic representation. Subsequently, the weight is linearly ramped up to $\lambda_3 = 0.1$ over a short period $\mathcal{Z}_w$ (typically spanning 300 mini-batches, equating to approx. 3-5 epochs). This curriculum ensures that strict physical laws are enforced only after the dynamical priors have stabilized, effectively preventing gradient conflict during the early training phase.

8.5. Implementation Details

Network Architecture & Hyperparameters. We incorporate specific constants to guarantee numerical stability in the physics-constrained modules: $\epsilon = 10^{-5}$ is added to the diagonal entries of $L(q)$ (Eq. (5)) to strictly enforce the positive definiteness of the inertia matrix $M(q)$, and $\delta = 0.1$ is applied to the denominator of the relative energy residual (Eq. (12)) to prevent division by zero. The warmup threshold $\mathcal{Z}$ is set adaptively based on dataset scale: $\mathcal{Z} = 20$ for LARa (trained for 60 epochs) and $\mathcal{Z} = 50$ for all other datasets (trained for 300 epochs). For the kinematic backbone, we adhere to the configurations of established baselines [5, 9]. The multi-scale GCN operates with a scale $K = 13$. The Linear Transformer employs 4 attention heads, with the channel dimensions for query, key, and value projections fixed at $C_2 = 16$. For prediction refinement, we utilize $S_c = 1$ stage for the class prediction branch and $S_b = 2$ stages for the boundary regression branch.

Text Embeddings. For the action-text contrastive loss, semantic representations are derived from the natural language descriptions of each action category. We utilize a pretrained BERT [3] encoder to extract word-level tokens, which are then averaged to produce sentence-level embeddings with a fixed dimensionality of $\mathbb{R}^{768}$. To ensure alignment across modalities, the dimensionality of the action-level visual representation is projected to $C_t = 768$.

9. Comparative Experiments

In this section, we conduct a multi-dimensional comparative analysis to comprehensively substantiate the efficacy and robustness of the proposed LaDy framework. First, we visualize the performance-efficiency trade-off to highlight LaDy's exceptional computational cost-effectiveness (Sec. 9.1). Subsequently, we quantify boundary localization precision across continuous IoU thresholds, validating the model's superior temporal alignment under stringent evaluation regimes (Sec. 9.2). This is followed by a granular per-class breakdown, which reveals how explicit dynamic priors effectively resolve specific kinematic ambiguities (Sec. 9.3). Furthermore, extended qualitative visualizations are provided to corroborate the structural coherence and boundary sharpness achieved by our approach (Sec. 9.4). Finally, we assess the model's resilience against complex input perturbations, demonstrating the powerful intrinsic regularization provided by our physics-constrained design (Sec. 9.5).

9.1. Performance vs. Efficiency Trade-off

To provide a more intuitive visualization of the performance vs. efficiency trade-off discussed in the main text (Sec. 4.2 and Tab. 1), we plot the performance scores (F1@50)

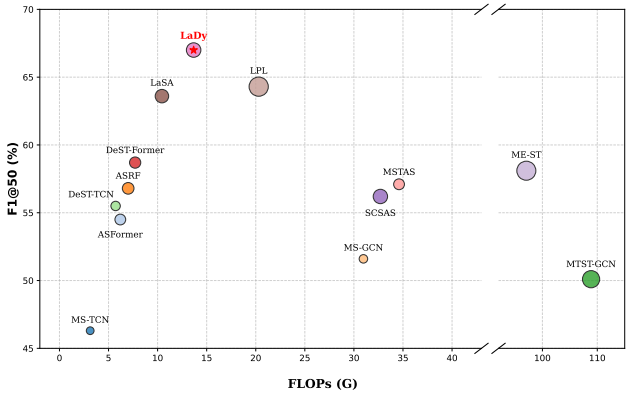
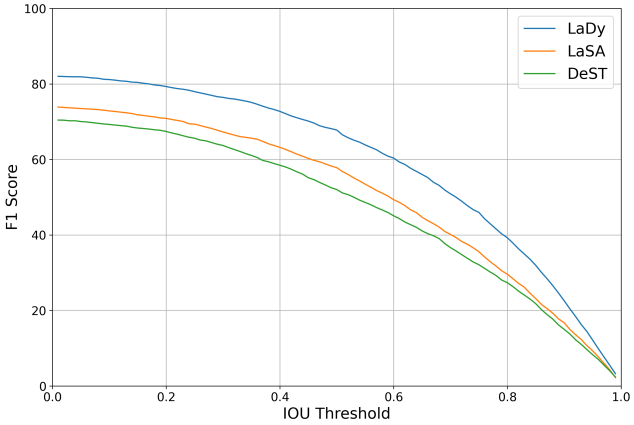


Figure 8. Performance vs. efficiency comparison on the PKU-MMD v2 dataset. The y-axis represents the F1@50 score, and the x-axis indicates computational complexity (FLOPs). Circle area is proportional to the model's parameter count. LaDy (red star) achieves state-of-the-art performance with a highly competitive computational footprint.

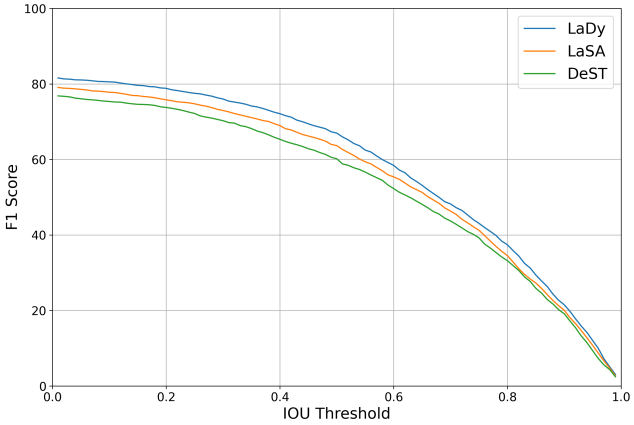
of recent state-of-the-art methods against their computational complexity (FLOPs) on the PKU-MMD v2 dataset (Fig. 8). The area of each circle reflects the corresponding model's parameter count. As illustrated, our LaDy framework establishes a new state-of-the-art in segmentation accuracy while maintaining a remarkably lightweight footprint (13.67G FLOPs). Compared to recent computationally heavy models such as LPL (20.3G), MSTAS (34.6G), and ME-ST (97.07G), LaDy achieves superior performance with significantly lower costs. This exceptional cost-to-performance ratio stems from our architectural design: the proposed physical branch avoids heavy computational overhead by relying solely on efficient MLPs and structured matrix operations. Specifically, integrating the Lagrangian Dynamics Synthesis (LDS) and Spatio-Temporal Modulation (STM) modules—coupled with the Energy Consistency Loss (ECLoss) during training—introduces a marginal overhead of merely 3.5G FLOPs and 0.51M parameters over the baseline, yet yields a substantial 2.7% absolute gain in F1@50. This confirms that explicitly leveraging physical priors is a highly efficient strategy for action segmentation, circumventing the need for excessive parametric scaling.

9.2. Boundary Precision Analysis

We evaluate the temporal precision of LaDy by analyzing F1 scores across a continuous spectrum of IoU thresholds ($\in [0, 1]$). As shown in Fig. 9, LaDy consistently outperforms the previous state-of-the-art kinematic models (LaSA [6], DeST [9]) across all thresholds on PKU-MMD v2. Crucially, LaDy exhibits superior robustness against localization strictness. While all methods display a monotonic performance decay as the IoU threshold tightens,



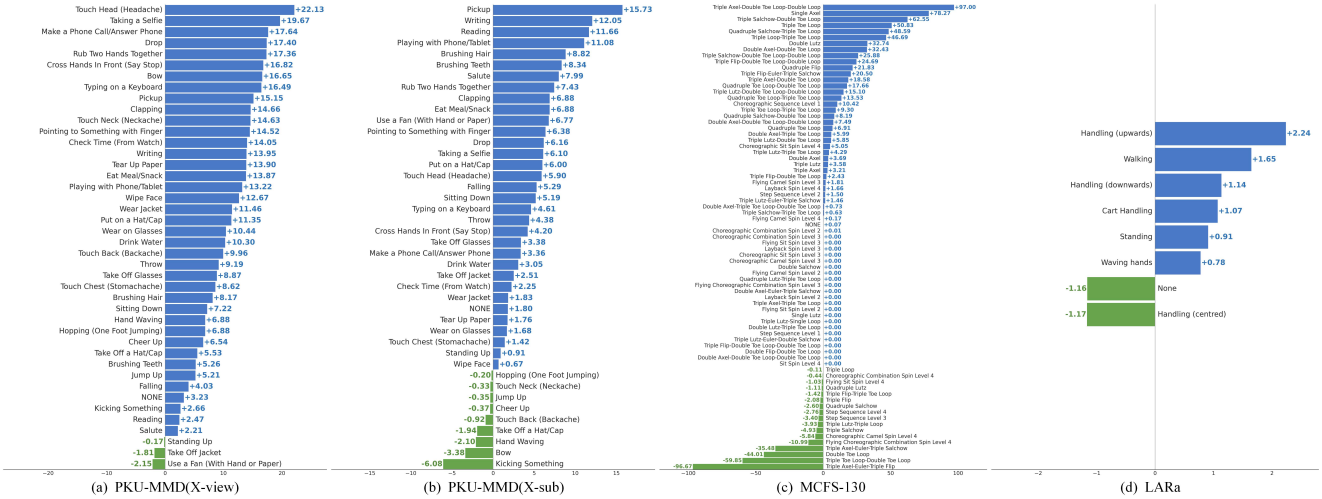
(a) F1 vs. IoU on PKU-MMD v2 (X-view)



(b) F1 vs. IoU on PKU-MMD v2 (X-sub)

Figure 9. F1 scores vs. IoU thresholds on PKU-MMD v2. LaDy (blue) maintains a distinct performance margin over LaSA (orange) and DeST (green), particularly at strict thresholds (IoU $>$ 0.7). This demonstrates that integrating Lagrangian dynamics (salient dynamic signals) for temporal modulation yields significantly sharper boundary localization.

LaDy demonstrates a significantly slower decay rate. Consequently, the relative performance margin between LaDy and competing methods widens considerably in the high-precision regime (IoU $\geq$ 0.7). This sustained accuracy indicates that our predicted segments possess higher structural alignment with the ground truth, effectively mitigating the boundary jitter common in kinematic-only models. This precision stems from the Spatio-Temporal Modulation (STM) mechanism. Unlike kinematic transitions, which are often smoothed and ambiguous, action boundaries are physically defined by abrupt shifts in force profiles. By explicitly leveraging the salient dynamic signals (e.g., torque change) for temporal gating, LaDy captures these sharp dynamic transients, transforming physical force variations into precise cues for boundary localization.



(a) PKU-MMD(X-view) (b) PKU-MMD(X-sub) (c) MCFS-130 (d) LARa

Figure 10. Per-class frame-wise F1 improvement of LaDy over the previous method (LaSA) on the PKU-MMD v2 (X-view and X-sub), MCFS-130, and LARa datasets. Blue bars indicate performance gains, while green bars denote declines. The results demonstrate that LaDy yields significant improvements in complex, physically distinct actions, validating that explicitly modeling dynamics effectively resolves kinematic ambiguities.

9.3. Per-Class Performance Analysis

To provide a granular understanding of how physical dynamics contribute to segmentation performance, we present a class-wise comparison between LaDy and the state-of-the-art kinematic method (LaSA) across four datasets in Fig. 10. The results reveal a clear pattern: LaDy achieves broad and significant improvements across a diverse spectrum of actions, validating the universality of the proposed dynamics-informed modeling.

Resolving Kinematic Ambiguities via Dynamic Signatures. The most substantial gains are observed in actions that are kinematically similar but dynamically distinct. On PKU-MMD (Fig. 10(a), (b)), subtle interactions such as *Touch Head*, *Make a Phone Call*, *Brush Hair*, and *Brush Teeth* see improvements in both X-sub and X-view settings. Kinematically, these actions all involve bringing the hand near the head, creating severe inter-class confusion for conventional models. However, LaDy effectively distinguishes them by capturing their unique "dynamic signatures"—the specific torque profiles required to stabilize the limb in these varying postures and the distinct force transients during the initiation of the movement. Similarly, on MCFS-130 (Fig. 10(c)), our method shows substantial gains in recognizing and distinguishing highly complex figure skating maneuvers, particularly the entire family of *Toe Loop*. These actions are defined by precise physical constraints (e.g., angular momentum conservation). While kinematic models struggle to differentiate the rapid, blurred rotations of a *Triple* vs. *Double* jump, LaDy's LDS module explicitly infers the magnitude of the driving torques, providing a decisive cue for classifying these high-energy, physics-dominated motions.

Analysis of Performance Degradation. Despite the overall superiority, we observe performance drops in a few specific categories, such as *Use a fan* (PKU-MMD X-view), *Kicking Something* (PKU-MMD X-sub) and *Handling (centred)* (LARa). We attribute this to two physical factors: (1) **Inaccurate External Force Estimation:** Actions like *Kicking* involve significant impact forces with external objects. Although the generalized force term $F(q, \dot{q})$ accounts for non-conservative and external forces in our Lagrangian equation, it is estimated solely from the input kinematics. Without explicit external contact sensing, the sudden velocity change upon impact cannot be accurately modeled, resulting in a transient violation of the estimated dynamics and confusing the network. (2) **Low-Energy Dynamics:** For passive or low-energy motions like *Use a fan*, the signal-to-noise ratio of the estimated joint torques is lower compared to high-energy actions. In these cases, the strong kinematic periodicity dominates, and the auxiliary dynamic features may introduce minor interference.

Nevertheless, the overwhelming prevalence of positive gains (Blue bars) confirms that for the vast majority of human actions, the integration of physical dynamics serves as a powerful and complementary prior.

9.4. Extended Qualitative Analysis

To provide a comprehensive assessment of the model's segmentation capability, we present extended qualitative comparisons across four benchmarks: PKU-MMD v2 (X-view/X-sub), MCFS-22, and LARa, as illustrated in Fig. 11.

Segmentation Consistency and Boundary Sharpness. Consistent with the main paper, LaDy generates segmen-

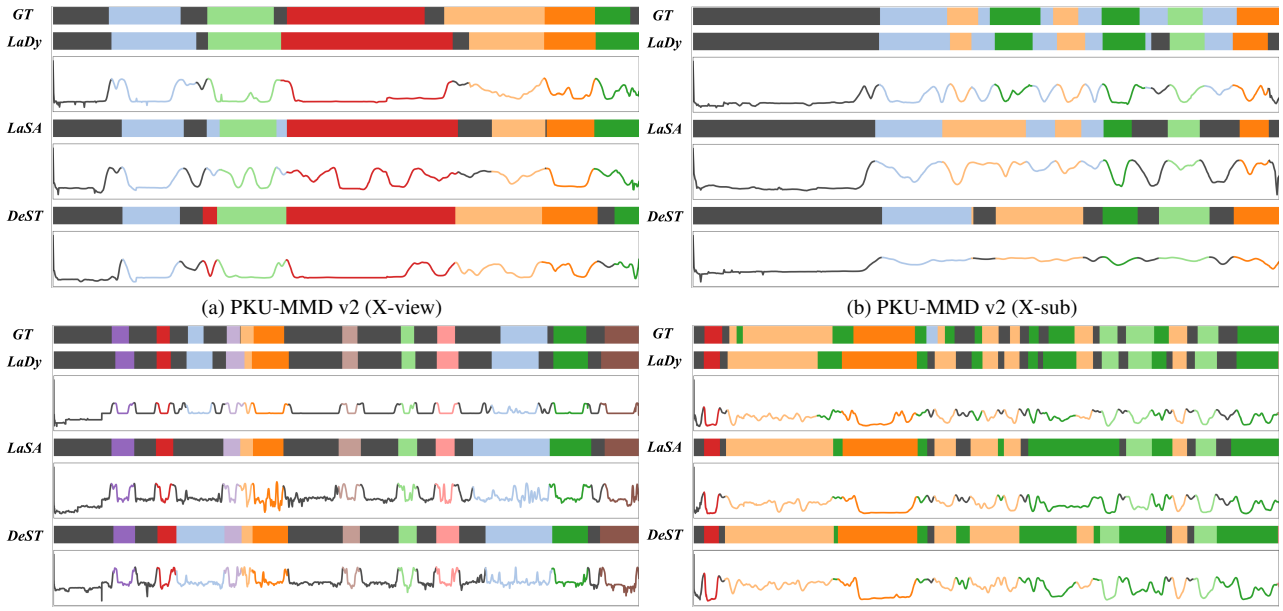


Figure 11. Qualitative results on PKU-MMD v2 (X-view and X-sub), MCFS-22, and LARa. The top row is the Ground Truth, followed by the segmentation results (bars) and boundary confidence scores (curves) for LaDy, LaSA, and DeST. Different colors denote distinct action classes.

tation predictions that are topologically more congruent with the Ground Truth (GT) compared to LaSA [6] and DeST [9]. The improvements are physically interpretable through the boundary confidence curves (plotted below the segmentation bars). LaDy's curves exhibit two distinct characteristics: (1) **Intra-segment Stability:** Within action segments, the confidence scores remain flat and low, effectively suppressing the "over-segmentation" noise often triggered by kinematic jitter in previous methods (e.g., the fluctuating curves in LaSA). (2) **Inter-segment Sharpness:** At action transitions, LaDy produces prominent, narrow peaks. This confirms that the salient dynamic (e.g., torque change) signal serves as a decisive cue, enabling the model to "lock on" to the physical onset of motion rather than drifting with visual ambiguity.

Analysis of Residual Errors and Failure Modes. While LaDy significantly reduces errors, it is not immune to failure. As observed in the dense transitions of LaDy's prediction bars, minor temporal boundary shifts persist. We attribute these deviations to "soft transitions"—scenarios where the dynamic force profile changes gradually rather than abruptly (e.g., slowly transitioning from a stance to a walk), making the precise definition of a "boundary" inherently ambiguous even for physics-informed models. Furthermore, sporadic classification errors remain in semantically overlapping classes (e.g., the middle segment in Fig. 11a). This suggests that while Lagrangian dynamics resolve kinematic ambiguities, future work could further benefit from integrating high-level semantic context to address these residual categorical confusions.

9.5. Robustness to Input Perturbations

To further evaluate the model's resilience against complex environmental noise, we conduct a robustness analysis under three synthetic noise conditions: Gaussian noise, joint occlusion, and missing frames. All evaluated models are trained solely on the clean training set and subsequently tested on the corrupted versions of the testing set. To ensure a strictly fair comparison, identical random seeds are applied across all methods to generate the exact same noise patterns. As reported in Tab. 4, LaDy consistently achieves the highest absolute performance across all corruption scenarios. More importantly, it exhibits significantly lower relative degradation rates compared to the previous state-of-the-art models, DeST and LaSA. This exceptional resilience substantiates the critical role of our Lagrangian Dynamics Synthesis (LDS) module. The embedded physical constraints, coupled with the Energy Consistency Loss, act as powerful intrinsic regularizers. By enforcing global dynamic consistency and work-energy principles, the physical branch effectively mitigates kinematic noise and fills observational gaps, enabling the model to maintain robust representations even when the raw data is heavily compromised.

Table 4. Robustness evaluation against various input perturbations (Gaussian noise, joint occlusion, and missing frames) on the PKU-MMD v2 (X-sub) dataset. All models are trained on clean data and evaluated on corrupted test sets. For each method, the first row reports the absolute performance, and the second row indicates the relative performance drop compared to the clean setting.

Noise	**Gaussian Noise**					**Joint Occlusion**					**Missing Frames**				
Metric	Acc	Edit	F1@{10, 25, 50}			Acc	Edit	F1@{10, 25, 50}			Acc	Edit	F1@{10, 25, 50}		
DeST	65.5	63.8	69.8	65.4	51.7	58.3	57.4	61.2	56.6	43.7	66.9	65.9	71.7	67.2	52.7
	-6.8%	-7.9%	-6.4%	-7.9%	-12.0%	-17.1%	-17.1%	-17.9%	-20.2%	-25.6%	-4.8%	-5.0%	-3.7%	-5.4%	-10.3%
LaSA	70.4	70.7	75.9	71.5	59.4	60.4	60.0	64.9	59.1	48.0	70.2	69.9	75.5	71.3	58.0
	-4.2%	-3.7%	-3.1%	-4.4%	-6.6%	-17.8%	-18.2%	-17.1%	-21.0%	-24.5%	-4.5%	-4.8%	**-3.6%**	-4.7%	-8.8%
LaDy	**73.6**	**73.1**	**78.2**	**75.4**	**64.2**	**70.1**	**68.6**	**73.6**	**70.1**	**59.5**	**73.4**	**71.7**	**77.1**	**74.5**	**62.1**
	-3.4%	**-2.6%**	**-2.4%**	**-2.5%**	**-4.2%**	**-8.0%**	**-8.6%**	**-8.1%**	**-9.3%**	**-11.2%**	**-3.7%**	**-4.5%**	-3.7%	**-3.6%**	**-7.3%**

10. Detailed Ablation Studies

In this section, we present a comprehensive ablation analysis to scrutinize the internal mechanisms of LaDy. Our evaluation is organized into three logical parts corresponding to the model's core contributions: (1) validating the structural necessity of the Lagrangian Dynamics Synthesis (LDS) (Sec. 10.1); (2) optimizing the formulation and stability of the Energy Consistency Loss (ECLoss) (Sec. 10.2); and (3) identifying the optimal structure for the Spatio-Temporal Modulation (STM) module (Sec. 10.3).

10.1. Ablation on Lagrangian Dynamics Synthesis

We systematically validate the design of the LDS module through four focused experiments. First, we investigate the impact of kinematic smoothing strategies on the computation of generalized kinematics. Second, we verify the superiority of our physics-constrained synthesis over black-box and unconstrained baselines. Third, we dissect the impact of specific geometric constraints imposed on the Inertia (M), Coriolis (C), and Gravity (G) terms. Finally, we assess the completeness of the dynamic formulation by evaluating the contribution of each individual Lagrangian component to the segmentation performance.

Table 5. Ablation on the calculation of generalized velocities and accelerations on the PKU-MMD v2 (X-sub) dataset. We compare our simple finite difference approach against explicit pre-smoothing techniques (Gaussian and Savitzky-Golay filtering).

Loss Formulation	**Acc**	**Edit**	**F1@{10, 25, 50}**		
Finite difference (Ours)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**
Gaussian smoothing	75.8	74.6	79.4	76.8	66.5
Savitzky-Golay	75.9	74.9	79.8	76.8	66.4

Impact of Kinematic Smoothing Strategies. To compute the generalized velocities $\dot{q}$ and accelerations $\ddot{q}$, our framework employs a straightforward first-order finite difference (Eq. (4)). To address concerns that discrete differentiation might amplify input noise, we compare this simple approximation against explicit pre-smoothing techniques, specifically Gaussian and Savitzky-Golay filtering. As shown in Tab. 5, introducing these filters does not improve performance; rather, it causes a slight degradation. We attribute this to two primary factors. First, the downstream spatio-temporal backbone, coupled with our physics-constrained modules and Energy Consistency Loss ($\mathcal{L}_{EC}$), inherently functions as a powerful intrinsic regularizer. The network effectively learns to handle input jitter implicitly by optimizing the generalized forces, rendering heuristic pre-denoising redundant. Second, explicit smoothing inevitably suppresses high-frequency kinematic signals. In action segmentation, these high-frequency transients are not merely noise; they are critical dynamic cues that represent abrupt force shifts at genuine action boundaries. Consequently, pre-smoothing artificially blurs these physical transitions, thereby diminishing the temporal gating module's sensitivity to boundary shifts and leading to suboptimal segmentation performance.

Table 6. Ablation on Dynamics Modeling Strategies on PKU-MMD v2 (X-sub) dataset. We compare direct estimation against equation-based synthesis with and without physical constraints and energy supervision.

Modeling Strategy	$\mathcal{L}_{EC}$	**Acc**	**Edit**	**F1@{10, 25, 50}**		
Baseline (w/o LDS)	-	73.6	73.0	78.2	74.6	64.3
Direct MLP Mapping	-	75.0	74.1	79.0	76.2	65.8
Unconstrained Eq.	-	75.2	74.5	79.7	76.5	66.2
Unconstrained Eq.	✓	75.3	74.7	79.3	76.5	66.7
Physics-Constrained Eq.	-	75.9	74.6	79.6	76.9	66.3
Physics-Constrained Eq.	✓	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Impact of Force Modeling Strategies. We further investigate how the generalized forces are modeled within the LDS module, comparing three paradigms: black-box estimation, unconstrained equation synthesis, and our physics-constrained synthesis. As shown in Tab. 6, the *Direct Mapping* (a simple MLP taking concatenated states as input) yields limited gains over the baseline, suggesting that implicit black-box modeling struggles to capture the underlying physical causality. Explicitly structuring the es-

timation via the Lagrangian formulation (*Unconstrained Eq.*) improves performance, yet remains suboptimal without geometric guarantees. Crucially, enforcing intrinsic physical constraints—specifically the positive definiteness of the inertia matrix and the skew-symmetry of the Coriolis term (*Physics-Constrained Eq.*)—provides a significant boost, confirming that structural validity is essential for learning robust dynamics. Finally, the Energy Consistency Loss ($\mathcal{L}_{EC}$) consistently enhances all equation-based models, with our full LaDy framework achieving the best results. This validates that combining structural physical constraints with energy-based supervision ensures the synthesized forces are not only semantically discriminative but also physically coherent.

Table 7. Ablation on Physical Constraints for Lagrangian Terms on PKU-MMD v2 (X-sub) dataset. We evaluate different constraint formulations for Inertia (M), Coriolis (C), and Gravity (G) estimation.

Term	Constraint Formulation	Acc	Edit	F1@{10, 25, 50}		
$\boldsymbol{M}$	Unconstrained	75.4	74.6	79.6	76.6	66.0
	Symmetric Only	75.3	74.1	79.3	76.3	66.0
	Sym. + Semi-Positive Definite	75.9	74.2	78.9	76.5	65.9
	Sym. + Positive Definite (SPD)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**
$\boldsymbol{C}$	Unconstrained C	75.5	74.2	79.6	76.7	66.2
	Unconstrained $C = 0.5(\dot{M} - N)$	75.9	74.8	79.7	76.6	66.6
	Passivity ($\dot{M} - 2C$ skew-sym.)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**
$\boldsymbol{G}$	Conservative (∇ Potential Energy)	75.6	74.5	79.4	76.8	65.7
	Unconstrained (Direct MLP)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Effectiveness of Physical Constraints. We further dissect the impact of imposing specific physical constraints on the constituent Lagrangian terms (M, C, and G). As detailed in Tab. 7, for the Inertia Matrix M, strictly enforcing **Symmetric Positive Definiteness (SPD)** is crucial, yielding the peak performance. Relaxing this to semi-definiteness or mere symmetry degrades performance, confirming that maintaining the valid Riemannian geometry of the mass matrix is fundamental for stable force synthesis. For the Coriolis Matrix C, the **Passivity Constraint** (ensuring $\dot{M} - 2C$ is skew-symmetric) outperforms unconstrained estimation. This structural coupling ensures the synthesized dynamics respect energy conservation principles, preventing artificial energy generation that would contradict our ECLoss. Interestingly, for the Gravity vector G, the direct unconstrained mapping proves superior to the strictly conservative modeling via potential energy gradients (∇E_P). We hypothesize that while theoretically rigorous, learning a scalar potential field to derive forces introduces a challenging optimization landscape with unstable high-order derivatives, whereas the direct mapping offers a more flexible and trainable approximation for the complex latent gravity manifold.

Table 8. Ablation on the necessity of each Lagrangian term on PKU-MMD v2 (X-sub) dataset. We report results when removing the Inertia (M), Coriolis (C), Gravity (G), or Non-conservative (F) terms individually.

Model Variant	Acc	Edit	F1@{10, 25, 50}		
w/o Inertia (M)	74.5	73.2	78.5	75.9	65.9
w/o Coriolis (C)	75.5	74.4	78.9	76.2	66.2
w/o Gravity (G)	75.0	73.9	79.1	76.0	65.9
w/o Non-Conservative (F)	75.2	73.5	78.8	76.5	65.7
Full Model (LaDy)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Necessity of Lagrangian Terms. We evaluate the contribution of each dynamic component by selectively removing the Inertia (M), Coriolis (C), Gravity (G), or Non-conservative (F) terms from the LDS module. As shown in Tab. 8, removing any single term leads to a distinct performance drop compared to the full model. This degradation is attributed to two factors: (1) **Dynamic Completeness**: Human motion is physically governed by the interplay of all these forces; omitting one yields a deficient dynamic representation that fails to capture the full kinematics-dynamics causality. (2) **Energy Consistency**: The efficacy of our $\mathcal{L}_{EC}$ relies on the closed-loop Work-Energy theorem. A missing term disrupts this physical equilibrium, rendering the energy supervision mathematically ill-posed and less effective. Notably, the exclusion of non-conservative forces (F) results in a significant decline (e.g., **-1.3%** in F1@50). Although explicitly inferring external interactions and friction solely from skeletal pose is theoretically challenging, this result confirms that the F branch successfully learns to approximate these critical dissipative effects, which are essential for recognizing actions involving object interactions or sudden stops.

10.2. Ablation on Energy Consistency Loss

Here, we investigate the numerical stability and optimization dynamics of $\mathcal{L}_{EC}$. Our analysis sequentially covers: (1) the internal loss formulation, validating the critical roles of relative normalization and noise masking; (2) the robustness of different regression metrics (e.g., Huber loss) against outlier noise; (3) the sensitivity analysis of the regularization weight λ_3; and (4) the necessity of the Delayed Physics Injection strategy for resolving the initialization conflict between kinematic learning and physical constraints.

Design of Energy Consistency Loss. We strictly investigate the formulation of $\mathcal{L}_{EC}$ by dissecting its two key regularization components: relative normalization and the noise mask. As reported in Tab. 9, a *Naive Formulation* (Smooth-L1 distance between work and kinetic energy change) yields suboptimal results, primarily because the op-

Table 9. Ablation on the internal design of the Energy Consistency Loss ($\mathcal{L}_{EC}$) on PKU-MMD v2 (X-sub) dataset. We compare the naive formulation against versions with Relative Normalization and Noise Masking.

Loss Formulation	Acc	Edit	F1@{10, 25, 50}		
Naive Formulation	75.7	74.3	79.7	76.9	66.6
w/ Relative Normalization	75.9	74.6	79.6	76.9	66.8
w/ Noise Masking $\mathcal{M}(t)$	76.2	74.4	79.8	77.3	66.8
Full Formulation (Ours)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

timization is biased towards high-energy actions. Incorporating *Relative Normalization* alleviates this by enforcing scale-invariant supervision, ensuring that subtle, low-energy motions contribute equally to the gradient, which notably improves boundary precision (Edit: **+0.3%**). Furthermore, the *Noise Masking* strategy $\mathcal{M}(t)$ proves critical for stability. By filtering out numerical singularities during static or micro-movement phases (where the energy denominator approaches zero), it prevents noise amplification and false penalties, significantly boosting frame-wise accuracy (Acc: **+0.5%**). The *Full Formulation*, combining both strategies, achieves the best trade-off between discriminability and boundary smoothness, validating that robust physical supervision requires both scale balance and noise resilience.

Table 10. Ablation on the distance metric for $\mathcal{L}_{EC}$ on PKU-MMD v2 (X-sub) dataset. We compare standard regression losses against the robust Huber loss.

Distance Function	Acc	Edit	F1@{10, 25, 50}		
ℓ_1 Loss (MAE Loss)	76.0	74.7	79.9	77.2	66.5
ℓ_2 Loss (MSE Loss)	76.0	74.9	80.0	77.4	67.0
Huber Loss (Smooth ℓ_1)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Regression Loss for Energy Residual. We ablate the distance function used to minimize the relative energy residual r_E. As shown in Tab. 10, the *Huber Loss* (Smooth ℓ_1) consistently outperforms both ℓ_1 and ℓ_2 losses. While the ℓ_2 loss (MSE) provides strong gradients for large errors, it is overly sensitive to outliers—common in instantaneous power estimates due to sensor noise—leading to training instability. Conversely, the ℓ_1 norm (MAE) is robust to outliers but suffers from optimization difficulties near zero due to its non-differentiability and constant gradient magnitude. The Huber loss strikes an optimal balance: it behaves quadratically for small residuals to ensure smooth, precise convergence towards physical equilibrium, while transitioning to a linear behavior for large deviations to maintain robustness against dynamic transients. This dual characteristic proves essential for learning physically coherent dynamics without being disrupted by transient noise.

Table 11. Ablation on the regularization weight λ_3 for $\mathcal{L}_{EC}$ on PKU-MMD v2 (X-sub) dataset.

Coefficient λ_3	Acc	Edit	F1@{10, 25, 50}		
1.0	75.9	74.2	79.8	76.9	66.3
0.1	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**
0.01	75.7	74.5	79.9	76.9	66.7
0.001	75.7	74.9	79.6	76.9	66.5
0.0001	76.1	75.0	79.8	77.1	66.6

Sensitivity to Loss Weight λ_3. We investigate the optimal balancing coefficient λ_3 for the Energy Consistency Loss ($\mathcal{L}_{EC}$). As presented in Tab. 11, the performance follows a bell-shaped trend, peaking at $\lambda_3 = 0.1$. Setting the weight too high ($\lambda_3 = 1$) leads to performance degradation (e.g., **-0.7%** in F1@50), as the strong physical regularization overshadows the primary segmentation objective, forcing the network to prioritize equation satisfaction over semantic feature learning. Conversely, reducing the weight below 0.1 diminishes the corrective impact of the physical constraints, yielding results similar to the unconstrained baseline. Thus, $\lambda_3 = 0.1$ provides the optimal trade-off, offering sufficient physical guidance to regularize the latent dynamics without disrupting the learning of discriminative spatio-temporal features.

Table 12. Ablation on the Delayed Physics Injection strategy on PKU-MMD v2 (X-sub) dataset. We evaluate the impact of Delayed Start ($\mathcal{Z}$) and Linear Warmup ($\mathcal{Z}_w$) on training stability.

Injection Strategy	Acc	Edit	F1@{10, 25, 50}		
Direct Injection	75.5	74.0	78.8	76.4	66.1
Delayed Start ($\mathcal{Z}$) only	75.9	74.7	79.7	76.9	66.8
Linear Warmup ($\mathcal{Z}_w$) only	75.7	74.9	79.2	76.5	66.4
Delayed + Phased Warmup	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Impact of Delayed Physics Injection. We strictly validate the necessity of our *Delayed and Phased Warmup* strategy for training stability. As shown in Tab. 12, directly imposing the Energy Consistency Loss from the first iteration (*Direct Injection*) compromises performance (e.g., **-0.9%** in F1@50 vs. Full). This deficit arises because the dynamic estimators ($\mathcal{F}_M$, $\mathcal{F}_C$, $\mathcal{F}_G$, $\mathcal{F}_F$) are randomly initialized; enforcing strict physical constraints on these erratic priors induces severe gradient conflicts, effectively disrupting the model's ability to learn basic kinematic representations in the early phase. Introducing a *Delayed Start* ($\mathcal{Z}$) significantly boosts performance by allowing the network to first establish a coarse kinematic manifold before physical regularization kicks in. Adding *Linear Warmup* ($\mathcal{Z}_w$) further smooths this transition. The full strategy, combining both, acts as a physical curriculum, ensuring that strict laws are enforced only after the dynamical priors have stabilized,

thus achieving the optimal convergence.

Table 13. Ablation on the fusion stage for Spatial Modulation on PKU-MMD v2 (X-sub) dataset. We compare fusing dynamics at the raw input level (Early), after spatial-channel merging (Late), and after GCN encoding (Mid).

Fusion Stage	Acc	Edit	F1@{10, 25, 50}		
Early (Input-level)	75.4	74.8	79.8	77.0	66.4
Late (Spatial Merging)	75.7	74.8	79.6	76.7	66.4
Mid (Post-GCN, Ours)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

10.3. Ablation on Spatio-Temporal Modulation

Finally, we conduct an extensive architectural search to determine the optimal paradigm for integrating dynamic priors. We first identify the most effective fusion stage and mechanism for Spatial Modulation. Subsequently, we analyze the temporal gating topology, validating the hierarchical evolving design. We then dissect the dynamic signal composition, evaluating the synergy of Torque, Change, and Power, followed by an investigation into their optimal fusion logic and interaction strategies (Decoupled vs. Coupled). Lastly, we verify the sufficiency of channel-agnostic scalar gating.

Optimal Stage for Spatial Modulation. We investigate the most effective insertion point for fusing the synthesized dynamics (F_{dyn}) with the kinematic features. As shown in Tab. 13, *Early Fusion* (concatenating forces with raw input coordinates) yields suboptimal performance, likely due to the semantic gap between raw signals and latent features. Similarly, *Late Fusion* (injecting dynamics after spatial-channel aggregation) also degrades results. By compressing the spatial dimension before fusion, this strategy aggregates local joint details into a global vector prematurely. This prevents the dynamic context from interacting with specific active joints, depriving the model of the ability to spatially localize dynamic effects. In contrast, our *Mid-Level Fusion* (post-GCN) achieves the best performance. By injecting the global dynamic context into the uncompressed GCN features, we allow the physical intent to explicitly modulate specific local joint representations before spatial information is lost. This global-to-local modulation ensures that the dynamics can selectively emphasize kinematically relevant limbs, thereby maximizing spatial discriminability.

Fusion Mechanism for Spatial Modulation. We compare different operations for integrating the global synthesized dynamics with local kinematic features. As shown in Tab. 14, naive *Node Concatenation* (appending force as an extra vertex) performs poorly, as it fails to explicitly interact with body joints. Element-wise operations (*Addition* and *Product*) also yield poor results, likely due to feature conflict or information washout. Notably, while *Cross-Attention* (using kinematics to adaptively weight dynamic features per joint) improves performance by enabling dynamic allocation, it is still outperformed by our proposed *Channel Concatenation*. We attribute this to the fact that concatenating the broadcasted global dynamics preserves the complete, uncompressed information of both modalities. This allows the subsequent spatial-channel mixing layers to learn unrestricted, high-dimensional non-linear correlations between forces and poses. This "learnable fusion" proves more robust and expressive than the constrained inductive bias of soft-attention mechanisms, effectively baking the global dynamic intent into the spatial representation.

Table 14. Ablation on fusion mechanisms for Spatial Modulation on PKU-MMD v2 (X-sub) dataset. We compare our Channel Concatenation strategy against node-level concatenation, cross-attention, and element-wise operations.

Fusion Mechanism	Acc	Edit	F1@{10, 25, 50}		
Node Concatenation	75.0	74.0	79.1	76.2	66.4
Element-wise Addition	75.6	74.7	79.7	76.9	66.0
Element-wise Product	75.9	73.9	79.2	76.3	66.2
Cross-Attention (Kin-Query)	75.7	74.7	79.8	76.8	66.7
Channel Concatenation (Ours)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Table 15. Ablation on the architecture of Temporal Modulation on PKU-MMD v2 (X-sub) dataset. We compare single-stage gating against multi-stage strategies with static, parallel, or hierarchical topologies.

Gating Architecture	Acc	Edit	F1@{10, 25, 50}		
Single-Stage (Input only)	75.5	74.4	79.4	76.7	66.5
Single-Stage (Output only)	75.3	74.2	79.1	76.5	66.3
Multi-Stage (Static/Repeated)	76.0	74.4	79.9	77.0	66.5
Multi-Stage (Parallel Projection)	75.8	74.2	79.6	76.6	66.6
Multi-Stage (Hierarchical Evolving)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Architecture of Temporal Modulation. We analyze the structural design of the temporal gating mechanism to determine the optimal topology for injecting dynamic cues across the L temporal stages. As shown in Tab. 15, limiting gating to a *Single-Stage* (Input or Output temporal modeling stage) is insufficient, as it fails to regulate the intermediate feature evolution. Extending to all stages improves results, yet the topology matters. The *Static* strategy (using the same initial signal for all L temporal gating layers) lacks adaptability. The *Parallel* strategy (using L independent heads to transform the initial signal for each gate) also falls short. We attribute this to a topological mismatch: the temporal backbone processes features serially, accumulating receptive fields and semantics layer-by-layer, whereas parallel gating treats each stage independently, failing to capture this accumulated context. Our *Hierarchical Evolving* strategy achieves the best performance by structurally

mirroring the backbone. By treating the gate as a serial stream that evolves alongside the main features, we ensure precise receptive field alignment and semantic co-evolution. This allows the dynamic signals to mature from local, sharp cues in shallow layers to abstract, boundary-aware signals in deep layers, perfectly matching the needs of the temporal hierarchy.

Table 16. Ablation on the combination of Salient Dynamic Signals for temporal gating on PKU-MMD v2 (X-sub) dataset. We evaluate the synergistic effect of Torque (g_τ), Torque Change ($g_{\dot{\tau}}$), and Power (g_P).

Dynamic Signals			Acc	Edit	F1@{10, 25, 50}		
g_τ	$g_{\dot{\tau}}$	g_P					
✓	-	-	75.2	74.2	79.2	75.9	66.2
-	✓	-	74.7	74.1	78.9	76.0	66.2
-	-	✓	75.3	74.5	79.8	76.9	66.7
✓	✓	-	75.2	74.1	79.1	76.0	66.3
✓	-	✓	75.4	74.6	79.8	76.8	66.6
-	✓	✓	76.0	74.8	79.8	76.9	66.9
✓	✓	✓	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Impact of Salient Dynamic Signals. We investigate the contribution of different dynamic cues—Torque (g_τ), Torque Change ($g_{\dot{\tau}}$), and Power (g_P)—to the temporal gating mechanism. As detailed in Tab. 16, individual signals yield limited gains, as they capture isolated physical aspects: g_τ reflects actuation magnitude, $g_{\dot{\tau}}$ detects transient shifts, and g_P measures energy intensity. Pairwise combinations provide partial improvements, but the *Full Combination* achieves the best performance (e.g., **+0.8%** F1@50 vs. single $g_{\dot{\tau}}$). This confirms that these signals are mutually complementary rather than redundant. By integrating all three, the model constructs a holistic dynamic profile that simultaneously encodes the "how much" (Torque/Power) and "when" (Torque Change) of the motion. This synergy enables the temporal hierarchy to adaptively attend to both high-energy semantic segments and subtle boundary transitions, maximizing segmentation precision.

Table 17. Ablation on the feature fusion strategy within the Temporal Modulation module on PKU-MMD v2 (X-sub) dataset. We compare linear summation methods against learnable concatenation-based fusion.

Fusion Strategy	Acc	Edit	F1@{10, 25, 50}		
Naive Summation	75.4	74.7	79.6	76.6	66.1
Static Weighted Sum	75.3	74.2	79.3	76.5	66.1
Adaptive Weighted Sum	75.4	74.9	80.0	77.0	66.6
Concat + Conv Fusion (Ours)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Feature Fusion in Temporal Modulation. We analyze how to effectively recombine the three modulated feature streams (gated by Torque, Change, and Power) within each temporal stage. As shown in Tab. 17, simple *Naive Summation* or *Static Weighted Sum* (learnable weight parameters) yields lower performance, as these linear superpositions enforce a fixed interaction mode that ignores frame-specific variations. While *Adaptive Weighted Sum* (generating instance-specific weights via an MLP) improves results by introducing dynamic adaptivity, it remains restricted to a linear combination, leading to potential feature conflict where distinct dynamic cues might cancel each other out. Our *Concat + Conv Fusion* outperforms all summation-based methods. We attribute this to its ability to perform non-linear synthesis: by concatenating the features, we preserve the independent semantic subspaces of each dynamic cue without premature merging. The subsequent convolution then acts as a learnable projection, enabling the model to construct complex, high-dimensional correlations between power, torque, and transients, thus maximizing the expressive power of the fused representation.

Table 18. Ablation on the interaction strategy of dynamic signals within the hierarchical gating mechanism on PKU-MMD v2 (X-sub) dataset. We compare merging signals (Fusion), interacting channels (Coupled), and independent processing (Decoupled).

Interaction Strategy	Acc	Edit	F1@{10, 25, 50}		
Early Fusion (Input-level)	75.4	74.2	79.3	76.4	66.5
Step-wise Fusion (Gate-level)	74.8	74.3	79.1	76.0	65.9
Coupled Evolution (Inter-signal)	75.4	74.7	80.0	76.8	66.7
Decoupled Evolution (Ours)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

Interaction Strategy for Dynamic Signals. We investigate how the three salient dynamic signals (Torque, Change, and Power) should interact during the hierarchical gating process. As shown in Tab. 18, strategies that merge signals into a single gate channel, either initially (*Early Fusion*) or at each stage (*Step-wise Fusion*), yield inferior performance. This suggests that forcing distinct physical cues to compete within a single-channel bottleneck dilutes their specific semantic roles. Furthermore, allowing three signals to interact during evolution via channel concatenation and convolution (*Coupled Evolution*) also proves suboptimal. We attribute this to feature homogenization: mixing heterogeneous physical quantities (e.g., energy magnitude vs. derivative transients) blurs their distinct functional properties. Our *Decoupled Evolution* strategy achieves the best results by treating the signals as heterogeneous, independent streams. By preserving the "physical purity" of each cue throughout the hierarchy, we enable the network to learn specialized modulations—where one stream exclusively highlights high-energy frames and another isolates boundary transitions—maximizing their complementarity when fused at the high-level feature space.

Channel Dimensionality of Dynamic Signals. We

Table 19. Ablation on the channel dimensionality of dynamic gating signals on PKU-MMD v2 (X-sub) dataset. We compare channel-wise modulation (C channels) against channel-agnostic scalar modulation (1 channel).

Signal Dimensionality	Acc	Edit	F1@{10, 25, 50}		
Multi-Channel ($T \times C$)	76.0	75.0	80.1	77.0	66.8
Channel-Agnostic ($T \times 1$)	**76.2**	**75.1**	**80.1**	**77.3**	**67.0**

assess whether the dynamic gating signals should retain channel-wise granularity or be compressed into channel-agnostic scalars. As shown in Tab. 19, the *Multi-Channel Gating* strategy (mapping dynamics to C channels for element-wise modulation) yields slightly inferior performance compared to the scalar approach. While channel-wise gating offers theoretical flexibility, it increases model complexity and risks overfitting to channel-specific noise. In contrast, our *Channel-Agnostic Gating* (compressing dynamics to 1D scalars via norm aggregation) achieves the best results. By aggregating dynamic energy across all dimensions, this strategy forces the gate to focus purely on frame-level temporal salience rather than channel variations. This global perspective ensures that the modulation acts as a stable temporal attention mechanism, prioritizing key timestamps (e.g., action boundaries) where the aggregate dynamic profile shifts, thus maximizing boundary precision with lower computational overhead.

11. Discussion

In this section, we present a broader discussion to contextualize the versatility and boundaries of the proposed LaDy framework. Specifically, we first explore its generalization capabilities across different tasks and non-human subjects (Sec. 11.1), followed by a critical analysis of its failure cases, intrinsic limitations, and directions for future research (Sec. 11.2).

11.1. Generalization Capabilities

To demonstrate the versatility of the proposed framework, we analyze its generalizability across different downstream tasks and non-human subject domains.

Generalization to Other Tasks: Action Recognition. While LaDy is primarily designed for action segmentation, its core physical components are inherently task-agnostic. The Lagrangian Dynamics Synthesis (LDS) module can serve as a general-purpose, physics-informed feature enhancer for any skeleton-based architecture. Specifically, the "dynamic signatures" extracted by the LDS (e.g., distinct torque profiles for actions like "throwing" or "clapping") explicitly model the causal intent of the motion, providing a strong prior that boosts inter-class discriminability.

To validate this empirically, we conduct preliminary experiments on the skeleton-based action recognition task. We integrate the LDS module into the established CTR-GCN [2] baseline and evaluate it on the widely used NTU-RGB+D (X-Sub) [13] dataset. The implementation simply involves deriving the generalized forces supervised by the Energy Consistency Loss ($\mathcal{L}_{EC}$) and injecting them into the early stages of the CTR-GCN backbone via our Spatial Modulation (SM) mechanism. As reported in Tab. 20, equipping the baseline with our physical modules yields consistent performance improvements across all input modalities (Joint, Bone, and their temporal motions) and the final multi-stream ensemble. Notably, these gains are achieved seamlessly without dataset-specific hyperparameter tuning, validating that modeling physical dynamics provides a universally beneficial inductive bias for general action understanding.

Table 20. Preliminary evaluation of LaDy as a plug-and-play feature enhancer for skeleton-based action recognition on the NTU-RGB+D (X-Sub) dataset. The integration consistently improves the CTR-GCN baseline across all individual modalities and the multi-stream ensemble.

Method	Joint	Bone	J-Motion	B-Motion	Ensemble4
CTR-GCN	90.1	90.2	87.8	87.2	92.4
+LaDy (Ours)	**90.3**	**90.4**	**88.3**	**88.9**	**92.9**

Generalization to Non-Human Subjects. The proposed framework is not fundamentally restricted to human biomechanics. The core Lagrangian dynamic formulation (Eq. (27)) relies mathematically on the definition of an open kinematic chain, which is represented by the skeleton graph topology $\mathcal{G}$. Because the underlying physical axioms—such as energy conservation, mass-inertia positive definiteness, and torque generation—apply universally to any articulated rigid-body system, the framework is theoretically subject-agnostic. By simply redefining the joint connectivity within $\mathcal{G}$ and establishing a corresponding root orientation to match a target morphology (e.g., a quadrupedal animal or a robotic manipulator), the LDS module can be directly transferred to analyze animal behavior or robotic motion. This universal adaptability underscores the broader potential of physics-informed modeling in multi-domain motion analysis.

11.2. Failure Cases, Limitations, and Future Work

Failure Cases and Intrinsic Limitations. While LaDy establishes a robust physics-informed paradigm for action segmentation, an analysis of its failure cases (as detailed in Sec. 9.3 and Sec. 9.4) reveals specific intrinsic limitations. Performance bottlenecks primarily manifest in the following scenarios: (1) **Inaccurate External Force Estimation:** For interaction-heavy actions (e.g., *Kicking Something*), sudden kinematic shifts are dictated by unmod-

eled external objects rather than internal actuation. Inferring the non-conservative force $F(q, \dot{q})$ solely from skeletal kinematics without explicit external contact sensing remains fundamentally ill-posed, occasionally resulting in transient violations of the estimated dynamics that confuse the network. (2) **Low-Energy Dynamics:** In passive or micro-movements (e.g., *Use a fan*), the actual physical actuation is minimal, leading to a naturally low signal-to-noise ratio in the estimated joint torques. In such cases, strong kinematic periodicity dominates, and enforcing auxiliary dynamic modulation can introduce minor interference. (3) **Ambiguous "Soft" Transitions:** During gradual motion changes (e.g., slowly transitioning from a stance to a walk), the dynamic force profile shifts smoothly rather than abruptly. The absence of sharp force transients makes the precise temporal localization of a boundary inherently ambiguous, even for physics-guided architectures. (4) **Semantic Overlaps:** Sporadic classification errors persist among semantically overlapping classes, indicating that resolving kinematic ambiguities via dynamics is sometimes insufficient without broader contextual understanding.

Future Work. Future research will address these challenges by extending the current framework into a physics-grounded multimodal architecture. To resolve the ambiguity of external forces, we plan to integrate visual context or object-centric representations to explicitly model interaction dynamics, thereby grounding the non-conservative force term F in observable physical contacts. Additionally, synergizing our low-level dynamic priors with the high-level semantic reasoning capabilities of Large Multimodal Models (LMMs) will help resolve residual categorical confusions, moving towards a holistic framework that understands both the *mechanics* of how humans move and the *semantics* of why they move. To better handle low-energy actions, we aim to explore adaptive gating mechanisms that dynamically adjust the network's reliance on the physics branch based on the instantaneous energy scale of the motion. Finally, to address the inherent ambiguity of soft transitions, future iterations could transition from deterministic boundary localization to probabilistic temporal modeling, allowing the network to explicitly quantify transition uncertainties.

References

[1] 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. 4

[2] 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. 15

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 4, 6

[4] 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. 5

[5] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *WACV*, pages 2322–2331, 2021. 5, 6

[6] 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. 4, 5, 6, 7, 9

[7] 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. 4, 6

[8] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pages 5156–5165, 2020. 4

[9] 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. 3, 5, 6, 7, 9

[10] 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. 5

[11] 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. 5

[12] 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. 5

[13] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 15

[14] Julian Wiederer, Arij Bouazizi, Ulrich Kressel, and Vasileios Belagiannis. Traffic control gesture recognition for autonomous vehicles. In *IROS*, pages 10676–10683, 2020. 5