FineTec: Fine-Grained Action Recognition Under Temporal Corruption via Skeleton Decomposition and Sequence Completion

Dian Shao[1]*, **Mingfei Shi**[1], **Like Liu**[2]

[1]Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China
[2]School of Software, Northwestern Polytechnical University, Xi'an, China
shaodian@nwpu.edu.cn, {mingfeishi5, like.liu}@mail.nwpu.edu.cn

Abstract

Recognizing fine-grained actions from temporally corrupted skeleton sequences remains a significant challenge, particularly in real-world scenarios where online pose estimation often yields substantial missing data. Existing methods often struggle to accurately recover temporal dynamics and fine-grained spatial structures, resulting in the loss of subtle motion cues crucial for distinguishing similar actions. To address this, we propose **FineTec**, a unified framework for **Fine**-grained action recognition under **Te**mporal **C**orruption. FineTec first restores a base skeleton sequence from corrupted input using context-aware completion with diverse temporal masking. Next, a skeleton-based spatial decomposition module partitions the skeleton into five semantic regions, further divides them into dynamic and static subgroups based on motion variance, and generates two augmented skeleton sequences via targeted perturbation. These, along with the base sequence, are then processed by a physics-driven estimation module, which utilizes Lagrangian dynamics to estimate joint accelerations. Finally, both the fused skeleton position sequence and the fused acceleration sequence are jointly fed into a GCN-based action recognition head. Extensive experiments on both coarse-grained (NTU-60, NTU-120) and fine-grained (Gym99, Gym288) benchmarks show that FineTec significantly outperforms previous methods under various levels of temporal corruption. Specifically, FineTec achieves top-1 accuracies of 89.1% and 78.1% on the challenging Gym99-severe and Gym288-severe settings, respectively, demonstrating its robustness and generalizability.

Code — https://github.com/SmartDianLab/FineTec
Website — https://smartdianlab.github.io/projects-FineTec

Introduction

Fine-grained action recognition (FAR) aims to identify human actions characterized by slight temporal variations and subtle semantic differences, making it a particularly challenging problem (Shao et al. 2020a). Skeleton-based representations have emerged as an effective modality for FAR due to their compactness and explicit focus on motion cues (Duan et al. 2022b). However, in complex scenarios such as gymnastics, e.g., *"salto forward stretched with 2 twists"* (Shao et al. 2020a), online pose estimation can suffer from severe frame dropping, reaching up to 69.6% dropping rate during rapid motion (Zheng et al. 2024). This results in temporally corrupted skeletal sequences and substantial performance degradation, as shown in Figure 1. The problem is especially critical for FAR, which depends on subtle, continuous motion cues (Huang et al. 2025; Myung et al. 2024) and is thus highly sensitive to temporal discontinuities.

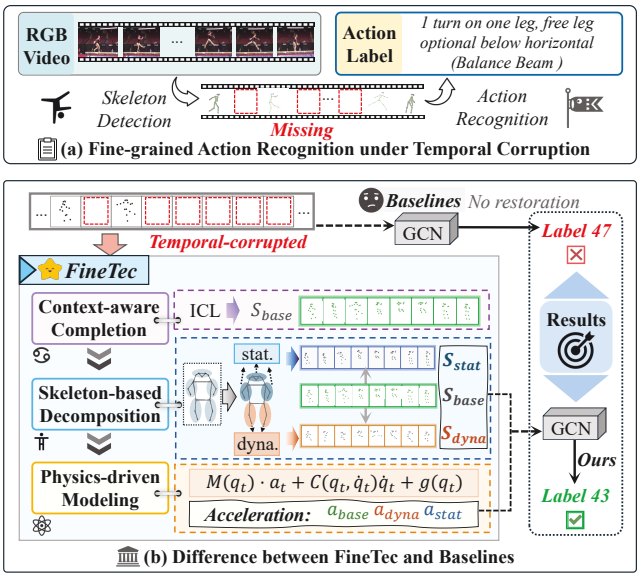

Figure 1: (a) Illustration of the challenging task: Fine-Grained Action Recognition under Temporal Corruption. (b) Compared to other GCN-based methods, the proposed FineTec framework can restore corrupted skeleton sequences and extract more discriminative features for recognition through context-aware completion, skeleton-based decomposition, and physics-driven modeling.

Nevertheless, current skeleton-based approaches mainly face two limitations when handling FAR under temporal corruption: (1) inadequate temporal recovery, as most models are trained on clean, offline-annotated skeleton sequences and lack mechanisms to handle online detection artifacts (Liu et al. 2025; Jiang and Deng 2024; Xie et al. 2024); and (2) insufficient spatiotemporal modeling, as they

*Corresponding author.

often overlook the inherent biological structure of the human body, focusing primarily on point-wise positional features while neglecting continuous kinematic constraints (Li et al. 2022; Leong et al. 2022; Chi et al. 2022).

To address these challenges, we introduce **FineTec**, a unified framework for **Fine**-grained action recognition under **Te**mporal **C**orruption, as shown in Figure 2. Specifically, FineTec consists of three key modules: (1) The *Context-aware Sequence Completion* module restores severely corrupted skeleton sequences via diverse temporal masking and in-context learning strategies, enabling an approximate recovery of missing frames and temporal continuity. (2) The *Skeleton-based Spatial Decomposition* module partitions skeleton joints into five semantic regions based on biological priors, and further divides them into dynamic and static subgroups by motion variance. Targeted augmentation techniques are then applied within each subgroup to generate two sequences, amplifying fine-grained action distinctions. (3) The *Physics-driven Acceleration Modeling* module re-estimates joint accelerations at each time step using Lagrangian dynamics and pseudo-acceleration (computed as temporal differences between frames). The resulting acceleration sequences effectively capture discriminative motion cues essential for FAR. The whole process of how the skeleton sequence is restored, processed and utilized is illustrated in Figure 1. Finally, the fused skeleton sequence and its corresponding acceleration cues are integrated for action recognition via a GCN-based network.

To enable comprehensive evaluation, we construct the Gym288-skeleton dataset by extending the open-source Gym99. We manually annotate 11,000 initial-frame bounding boxes, apply OSTrack for athlete tracking, and perform pose estimation within the tracked boxes. The resulting dataset provides skeleton annotations for 288 fine-grained action classes, offering a more challenging benchmark for future research. To validate the effectiveness of FineTec, we conduct comprehensive experiments on both coarse-grained (NTU-60, NTU-120) and fine-grained (Gym99-skeleton, Gym288-skeleton) datasets, systematically simulating varying levels of temporal corruption, including minor (25% frame drop), moderate (50%), and severe (75%). Experimental results demonstrate that FineTec consistently outperforms previous methods across all settings, and maintains strong recognition accuracy even under severe frame-dropping scenarios.

Our contributions are summarized as follows:

- We formalize and benchmark fine-grained action recognition under temporal corruption, constructing a large-scale dataset, *Gym288-skeleton*, for comprehensive evaluation;
- We propose **FineTec**, a unified framework that integrates context-aware sequence completion, biologically-aware spatial decomposition, and physics-driven temporal refinement to recover corrupted temporal continuity and enhance skeleton sequence quality for improved recognition;
- Through extensive experiments on both coarse- and fine-grained datasets, we demonstrate that FineTec achieves state-of-the-art performance, especially in the presence of severe temporal corruption.

Related Work

Skeleton-based Fine-grained Action Recognition

Fine-grained Action Recognition (FAR) aims to distinguish subtle action differences (*e.g.*, "pike sole circle backward with 0.5 turn to handstand"), enabling specialized analysis beyond coarse-grained categories (Zhang, Gupta, and Zisserman 2021; Yang et al. 2020; Wang et al. 2018; Chen et al. 2025b; Shao et al. 2020b; Rajendran et al. 2024). Among modalities, skeletons provide an effective FAR representation by capturing human dynamics while avoiding background noise. And the key to FAR is enhancing distinctions in subtle motion details: MDR-GCN (Liu et al. 2023) and Sparse (Xie et al. 2025) enhance skeletal features multi-dimensionally. BlockGCN (Zhou et al. 2024) optimizes topological structure of GCNs to obtain more discriminative features. PoseConv3D (Duan et al. 2022b) employs a heatmap for spatio-temporal dynamics. PGVT (Zhang et al. 2024a) and SCoPLe (Zhu et al. 2025) integrates keypoint with multi-modal features. However, they predominantly emphasize "displacement" information, relying on data-driven methods to implicitly learn complex temporal dynamics, and lacking guidance from physical realism. While ActCLR (Lin, Zhang, and Liu 2023) targets fine-grained resolution in skeletal space, its contrastive learning framework limits comprehensive exploitation. Diverging from prior work, FineTec enables physically interpretable modeling of fine-grained actions by concurrently addressing skeletal spatial granularity and physics-constrained temporal dynamics.

Physics-aware Video Understanding

Traditional video understanding is fundamentally limited by the agnostic nature of models and weakly interpretable feature spaces (Lin et al. 2025; Wang et al. 2025; Shao et al. 2025). And physics-aware manners (Gärtner et al. 2022a,b) offers a promising alternative by embedding physical principles. Some works leverage simulation environments for phenomena like rigid body collision or fluid dynamics (Andriluka et al. 2025; Liu et al. 2024; Chen et al. 2025a), but adjusting physical parameters within engines remains challenging. Others directly integrate mathematical physics equations into model design (Zhang et al. 2024b; Ugrinovic et al. 2024). Among them, PIMNet (Zhang et al. 2022) uses Newtonian for motion prediction, InfoGCN++ (Chi et al. 2025) utilizes Neural-ODEs for action recognition, and LieGroupHamDL (Thai Duong and Atanasov 2023) combines Lie groups with Hamiltonian for robot control. In contrast, our FineTec: (1) focuses on fine-grained temporal-corrupted action recognition tasks; (2) introduces the ICL mechanism and masking strategies for temporal completion; (3) and integrates biological priors and Lagrangians for fine-grained analysis.

Methodology

Preliminary

❑ **Physics of Rigid-body Dynamics.** Rigid-body dynamics are typically formulated using Lagrangian or Hamiltonian methods, with the Lagrangian formulation (Zhang,

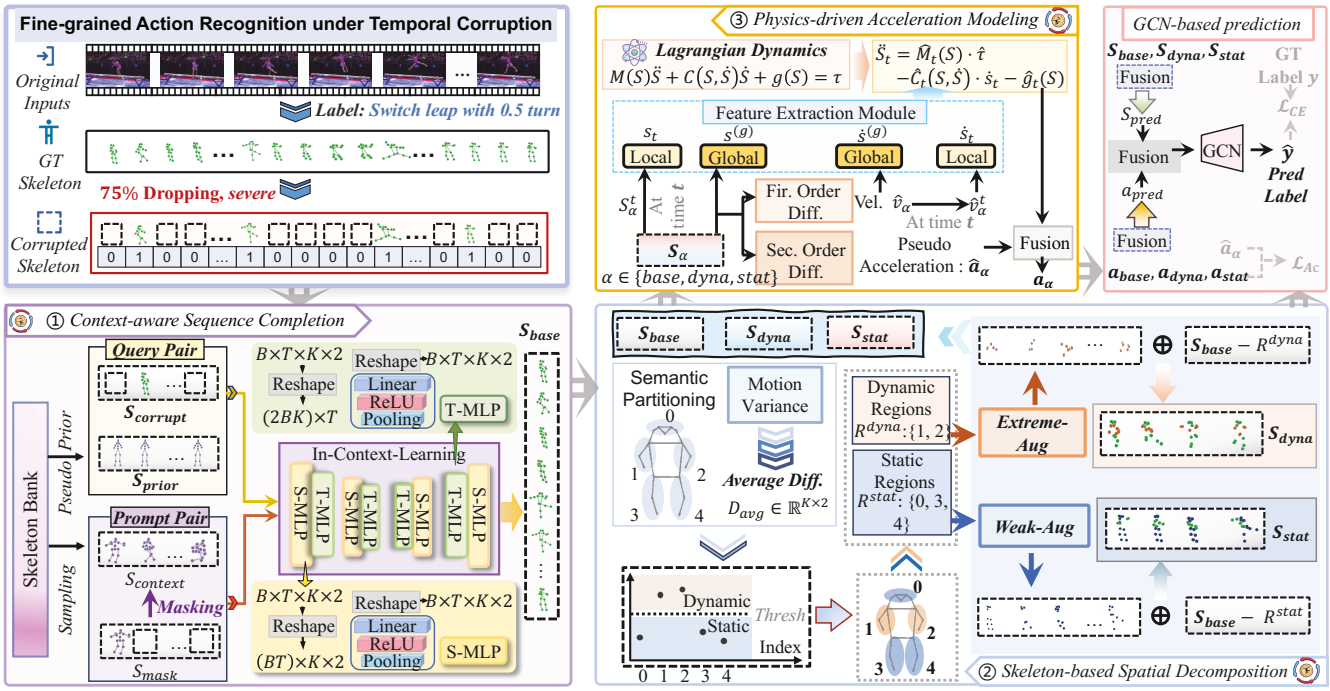

Figure 2: Overview of the Pipeline. FineTec consists of three core modules: ① Context-aware Sequence Completion restores missing or corrupted skeleton frames using in-context learning, producing S_{base}; ② Skeleton-based Spatial Decomposition partitions S_{base} into anatomical regions by motion intensity, generating dynamic (S_{dyna}) and static (S_{stat}) variants, which are fused into S_{pred}; ③ Physics-driven Acceleration Modeling infers joint accelerations via Lagrangian dynamics and data-driven finite differences, producing fused temporal dynamics features $\mathbf{a}$. The resulting positional (S_{pred}) and dynamic (a_{pred}) features are used for downstream fine-grained action recognition.

Kephart, and Ji 2024) widely adopted for its clarity and simplicity. For human kinematics (Jain and Rodriguez 1995), the Lagrangian dynamics are described by:

$$M(q_t) \cdot \ddot{q}_t + C(q_t, \dot{q}_t)\dot{q}_t + g(q_t) = \tau_t, \tag{1}$$

where q_t and $\dot{q}_t$ denote generalized coordinates, at time t. $M(q_t)$ is the configuration-dependent inertia matrix reflecting the mass distribution of body segments; $C(q_t, \dot{q}_t)\dot{q}_t$ represents Coriolis and centrifugal forces; $g(q_t)$ accounts for gravitational forces; and τ_t denotes the vector of generalized forces, including joint torques and external influences.

❍ **Task Definition** In this work, we address the challenging task of fine-grained action recognition under temporal corruption. Given a ground-truth 2D skeleton sequence $S_{gt} \in \mathbb{R}^{T\times K\times 2}$, where T is the number of frames and $K = 17$ is the number of joints, we simulate temporal corruption by randomly dropping 25% (minor), 50% (moderate), or 75% (severe) of the frames to obtain $S_{corrupt} \in \mathbb{R}^{T\times K\times D}$, with only $\hat{T} < T$ valid frames and the rest zero-padded. This setting reflects the missing data challenges frequently encountered in real-world online action detection scenarios. The goal is to predict the action category $y \in \{1, \ldots, C\}$ from the corrupted sequence $S_{corrupt}$, where C is the number of action classes.

The FineTec Framework

❍ **Overall Pipeline.** The FineTec framework processes temporally corrupted skeleton sequences through three key modules to achieve better fine-grained action recognition results, as illustrated in Figure 2. First, the *Context-aware Sequence Completion* module employs in-context learning to restore missing or corrupted frames, producing a basically completed skeleton sequence S_{base}. Next, the *Skeleton-based Spatial Decomposition* module partitions S_{base} into five anatomical regions and classifies them by motion intensity. Region-specific augmentations generate dynamic (S_{dyna}) and static (S_{stat}) variants, which are fused to yield the final sequence S_{pred}. These processed sequences are then used to extract displacement and acceleration features, which serve as inputs for downstream recognition networks. Finally, the *Physics-driven Acceleration Modeling* module infers joint accelerations via physics-based Lagrangian dynamics, and combine with data-driven finite differences to generate the fused temporal dynamics features $\mathbf{a}$. The resulting positional S_{pred} and dynamic features $\mathbf{a}$ are then utilized together for downstream fine-grained action recognition.

♣ **Context-aware Sequence Completion.** To handle temporally corrupted skeleton sequences $S_{corrupt}$, we adopt an In-Context Learning (ICL) paradigm (Wang et al. 2024; Kim et al. 2025) to approximately recover the complete sequence

at first. We first construct a skeleton bank from Human3.6M 2D skeleton data, and obtain an average prior sequence S_{prior} via temporal averaging. For each training instance, a sequence is sampled from the bank and corrupted by one of five temporal masking strategies: random, pattern-based, or contiguous block masking (prefix, suffix, in-between). The original $S_{context}$ and the masked sequence S_{mask} form a prompt pair, demonstrating recovery from corruption. The input $S_{corrupt}$ is paired with S_{prior} as a query pair. Both the prompt and query pairs are processed by lightweight spatial and temporal MLPs, enabling the network to approximately restore the base sequence S_{base} by contextually completing missing frames. Further implementation details are provided in the Appendix.

♦ **Skeleton-based Spatial Decomposition.** This module enhances fine-grained action discrimination by decomposing and augmenting the predicted skeleton sequence S_{base} based on motion analysis and anatomical priors. Specifically, leveraging the human biological structure, we first partition the K joints into five semantic regions: head (G_0), left arm (G_1), right arm (G_2), left leg (G_3), and right leg (G_4). To quantify the motion level of each joint, we compute the average frame-wise displacement as follows:

$$D_{avg}^{(i)} = \frac{1}{T-1} \sum_{t=0}^{T-2} \|S_{base}^{t+1,i} - S_{base}^{t,i}\|_2 \tag{2}$$

where i indexes the joint. The regional motion intensity for each group G_j is then calculated as:

$$\bar{D}_j = \frac{1}{|G_j|} \sum_{i \in G_j} D_{avg}^{(i)} \in \mathbb{R}, \quad j \in \{G_1, G_2, \cdots, G_5\}. \tag{3}$$

The top two regions with the highest $\bar{D}_j$ are designated as dynamic, and the remaining three as static. To introduce region-specific diversity, we apply strong spatial-temporal perturbations (such as temporal cropping, random dropping, or interpolation) to S_{base} and selectively replace the *dynamic regions* with their perturbed versions, resulting in S_{dyna}. For static regions, we apply only weak spatial perturbations (e.g., random flipping) and substitute the corresponding *static regions* in S_{base}, producing S_{stat}.

Finally, S_{base}, S_{dyna}, and S_{stat} are fused to form the final sequence S_{pred}. This decomposition-augmentation-fusion strategy preserves temporal coherence, amplifies motion cues in dynamic regions, and stabilizes static postures, collectively improving fine-grained action recognition.

♥ **Physics-driven Acceleration Modeling.** In this module, we explicitly model acceleration dynamics using Lagrangian principles to enhance motion representation.

Recall the Lagrangian equation (Eq. 1), where we substitute the general coordinates q with the set of joint positions S:

$$M(S)\ddot{S} + C(S, \dot{S})\dot{S} + g(S) = \tau, \tag{4}$$

where M, C, g, and τ denote the inertia matrix, Coriolis matrix, gravity term, and driving force, respectively. Here S could be S_{base}, S_{dyna}, and S_{stat}, and we omit the subscripts for convenience. Our aim is to calculate the acceleration term $\ddot{S}$:

$$\ddot{S} = \{M(S)\}^{-1} \cdot \tau - \hat{C}(S, \dot{S})\dot{S} - \hat{g}(S). \tag{5}$$

For computational efficiency and to facilitate neural network estimation, we define: $\hat{C}(S, \dot{S}) \coloneqq M(S)^{-1}C(S, \dot{S})$, $\hat{g}(S) \coloneqq M(S)^{-1}g(S)$, and $\hat{M}(S) \coloneqq M(S)^{-1}$. To provide the necessary inputs for these estimators, we first extract both global and local features of the joint positions and velocities:

$$s^{(g)} = f_s^{(global)}(\{S^t\}_{t=0}^{T-1}),\ s_t = f_s^{(local)}(S_t); \tag{6}$$

$$\dot{s}^{(g)} = f_{\dot{s}}^{(global)}(\{\hat{v}_t\}_{t=0}^{T-1}),\ \dot{s}_t = f_{\dot{s}}^{(local)}(\hat{v}_t); \tag{7}$$

where $\hat{v}_\alpha^t$ is calculated using first-order finite differences as $\hat{v}_t = \frac{S_{t+1} - S_{t-1}}{2\Delta t}$. Each physical term in the dynamics equation is then estimated using neural networks $\mathbb{E}$. At a specific time t:

$$\hat{g}_t(S) = \mathbb{E}_g[s^{(g)}, s_t]. \tag{8}$$

Since τ is time-independent, we have:

$$\hat{\tau} = \mathbb{E}_\tau[s^{(g)}, \dot{s}^{(g)}] \tag{9}$$

The remaining two terms are matrices, which we assume to be symmetric. Therefore, we first estimate their upper triangular parts and obtain the final matrices by applying the symmetry operation $\mathcal{S}^\dagger$:

$$\hat{C}_t(S, \dot{S}) = \mathcal{S}^\dagger\{\mathbb{E}_C[s^{(g)}, s_t, \dot{s}^{(g)}, \dot{s}_t]\}, \tag{10}$$

$$\hat{M}_t(S) = \mathcal{S}^\dagger\{\mathbb{E}_M[s^{(g)}, s_t]\}. \tag{11}$$

The refined, physics-driven acceleration is then computed as:

$$\ddot{S}_t = \hat{M}_t(S) \cdot \hat{\tau} - \hat{C}_t(S, \dot{S}) \cdot \dot{s}_t - \hat{g}_t(S). \tag{12}$$

To further improve robustness, we combine this estimate with the pseudo-acceleration calculated via second-order finite differences: $\hat{a}_t = \frac{S_{t+1} - 2S_t + S_{t-1}}{(\Delta t)^2}$. The final fused temporal dynamics feature (joint acceleration) is:

$$\mathbf{a}_t = \text{Fusion}(\hat{a}_t, \ddot{S}_t) \in \mathbb{R}^{K \times 2}. \tag{13}$$

♠ **GCN-based Optimization Objectives.** The FineTec framework is trained in two stages: skeleton sequence completion and the action recognition task.

① For sequence completion, we use mean squared error (MSE) losses to measure the difference between the completed skeleton and the ground-truth sequence, for both the prompt and query pairs:

$$\mathcal{L}_{ICL} = \text{MSE}(S_{gt}, S_{base}) + \text{MSE}(S_{context}, S_{mask}) \tag{14}$$

② For action recognition, the model utilizes the fused skeleton sequence S_{pred} and fused acceleration sequence $\mathbf{a}_{pred}$, which are integrated through a cross-attention module to capture both positional and dynamic information. The resulting representations are processed by graph convolutional

Method	Input	G288-Min.		G288-Mod.		G288-Sev.		G99-Min.		G99-Mod.		G99-Sev.	
		Top-1	Mean	Top-1	Mean	Top-1	Mean	Top-1	Mean	Top-1	Mean	Top-1	Mean
ST-GCN *AAAI'18*	Skeleton	0.784	0.381	0.770	0.344	0.742	0.304	0.895	0.869	0.876	0.829	0.871	0.783
PYSKL-J *arXiv'22*	Skeleton	0.813	0.401	0.794	0.368	0.773	0.315	0.920	0.871	0.903	0.856	0.884	0.791
PYSKL-B *arXiv'22*	Skeleton	0.811	0.385	0.796	0.373	0.765	0.314	0.915	0.872	0.905	0.858	0.889	0.801
PoseC3D-J *CVPR'22*	Heatmap	0.793	0.297	0.771	0.284	0.747	0.253	0.916	0.871	0.904	0.854	0.873	0.770
PoseC3D-L *CVPR'22*	Heatmap	0.790	0.296	0.775	0.281	0.756	0.250	0.917	0.870	0.899	0.848	0.870	0.761
AAGCN *TIP'20*	Skeleton	0.755	0.281	0.765	0.279	0.744	0.263	0.907	0.856	0.902	0.846	0.874	0.795
CTRGCN *ICCV'21*	Skeleton	0.786	0.292	0.784	0.285	0.760	0.271	0.914	0.874	0.897	**0.859**	0.884	0.803
Sparse *CVPR'25*	Skeleton	0.765	0.282	0.740	0.268	0.683	0.237	0.898	0.860	0.876	0.827	0.808	0.725
FineTec (Ours)	Skeleton	**0.815**	**0.404**	**0.797**	**0.381**	**0.781**	**0.356**	**0.921**	**0.875**	**0.906**	0.851	**0.891**	**0.805**

Table 1: Fine-grained action recognition on Gym99-skeleton and Gym288-skeleton. Both Top-1 accuracy and mean class accuracy are reported under minor (Min.), moderate (Mod.), and severe (Sev.) temporal corruption.

Method	NTU-60			NTU-120		
	Min.	Mod.	Sev.	Min.	Mod.	Sev.
ST-GCN	0.894	0.890	0.879	0.810	0.803	0.781
PYSKL-J	0.885	0.883	0.875	0.809	0.808	0.790
PYSKL-B	0.893	0.887	0.885	0.815	0.810	0.790
PoseC3D-J	0.887	0.889	0.878	**0.823**	0.795	0.783
PoseC3D-L	0.899	0.897	0.877	0.816	0.812	0.785
AAGCN	0.891	0.886	0.873	0.813	0.807	0.796
CTRGCN	0.901	0.892	0.879	0.814	0.809	0.793
Sparse	0.895	0.896	0.864	0.813	0.793	0.767
FineTec (Ours)	**0.903**	**0.901**	**0.892**	0.819	**0.817**	**0.813**

Table 2: Coarse-grained action recognition on NTU-60-xsub and NTU-120-xsub. Top-1 accuracy is reported under minor, moderate, and severe temporal corruption.

networks (GCNs) (Duan et al. 2022a), followed by a recognition head that predicts class probabilities $\hat{y}$. The classification loss is defined as the cross-entropy between the predicted and ground-truth labels:

$$\mathcal{L}_{CE} = -\sum_i y_i \log \hat{y}_i, \tag{15}$$

where y_i is the ground-truth label. An additional loss is also calculated between $\hat{\mathbf{a}}$ and $\mathbf{a}$ over the three sequences:

$$\mathcal{L}_{Ac} = \frac{1}{3}\sum_{\alpha} \text{MSE}(\hat{\mathbf{a}}_\alpha, \mathbf{a}_\alpha), \quad \alpha \in \{base, dyna, stat\}. \tag{16}$$

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{Ac}, \tag{17}$$

where λ is a balancing hyperparameter.

Experiment

Gym288-Skeleton Dataset

In this work, to enable comprehensive evaluation of fine-grained and temporally corrupted action recognition, we construct the Gym288-skeleton dataset by extending the open-source Gym99 dataset. We manually annotate $\sim$ 1.1w initial-frame bounding boxes, apply OSTrack (Ye et al. 2022) for athlete tracking, and perform pose estimation within the tracked boxes for each video. This provides a large-scale dataset with skeleton annotations for 288 fine-grained action classes, advancing the scope and difficulty of existing benchmarks. Statistics are shown in Figure 3, and the detailed constructing process and analysis are shown in the supplementary material.

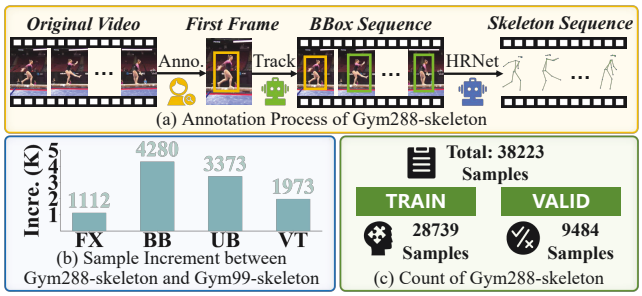

Figure 3: The Construction Process and Statistics of the constructed Gym288-skeleton Dataset.

Evaluation Settings

➳ **Datasets:** We conduct fine-grained action recognition experiments on two benchmark datasets: Gym99-skeleton and the constructed Gym288-skeleton (Shao et al. 2020a). For coarse settings, we use the NTU datasets, including NTU60 (Shahroudy et al. 2016) and NTU120 (Liu et al. 2020). ➳ **Baselines:** We compare our method with several representative skeleton-based approaches: ST-GCN (Yan, Xiong, and Lin 2018), PYSKL (Duan et al. 2022a), PoseC3D (Duan et al. 2022b), AAGCN (Shi et al. 2020), CTRGCN (Chen et al. 2021), and Sparse (Xie et al. 2025). ➳ **Evaluation Metrics:** For action recognition, we report Top-1 and Top-5 accuracy. Due to the significant class imbalance in Gym288 (Shao et al. 2020a), Mean class accuracy (Mean) is also included as a more informative metric. For skeleton restoration, we employ standard metrics: Mean Per

Method	Gym99-Min.			Gym99-Mod.			Gym99-Sev.		
	MPJPE↓	N-MPJPE↓	MPJVE↓	MPJPE↓	N-MPJPE↓	MPJVE↓	MPJPE↓	N-MPJPE↓	MPJVE↓
L-R Copy	0.136	0.133	0.246	0.332	0.318	0.445	0.713	0.665	0.575
R-L Copy	0.136	0.132	0.246	0.327	0.313	0.441	0.699	0.650	0.571
siMLPe	0.175	0.139	0.119	0.208	0.168	0.129	0.245	0.199	0.139
SiC-Stat	0.210	0.102	0.351	0.397	0.181	0.508	0.584	0.252	0.544
SiC-Dyna	0.188	0.100	0.196	0.164	0.144	0.205	0.192	0.174	0.321
Ours	**0.106**	**0.098**	**0.047**	**0.119**	**0.109**	**0.085**	**0.147**	**0.132**	**0.113**

Table 3: Skeleton restoration results on Gym99-Skeleton. Experiments are conducted under three levels of temporal corruption: minor, moderate, and severe. The evaluation metrics include MPJPE, N-MPJPE, and MPJVE. And the "L-R" and "R-L" denote left-to-right and right-to-left respectively.

Index	Cont.	Skel.	Phys.	Minor	Moderate	Severe
❶	✗	✓	✓	0.812	0.785	0.751
❷	✓	✗	✓	0.787	0.780	0.770
❸	✓	✓	✗	0.789	0.776	0.775
❹	✓	✓	✓	**0.815**	**0.797**	**0.781**

Table 4: Ablations of different modules, including the Context-aware Sequence Completion (Cont.), Skeleton-based Spatial Decomposition (Skel.), and Physics-driven Acceleration Modeling (Phys.).

Index	S_{dyna}	S_{stat}	Moderate	Severe
❶	✗	✓	0.790	0.774
❷	✓	✗	0.786	0.764
❸	✓	✓	**0.797**	**0.781**

Table 5: Analysis on augmented skeleton sequences.

Joint Position Error (MPJPE) and Mean Per Vertex Position Error (MPVPE). Main results are presented in this section, with further details provided in the Appendix.

Main Results

Results on Fine-grained Action Recognition. The main quantitative results on the two fine-grained skeleton datasets, Gym99 and Gym288-skeleton, are presented in Table 1. These results are reported across three difficulty levels: minor (25% frame missing), moderate (50% frame missing), and severe (75% frame missing). It can be observed that the proposed FineTec framework consistently achieves the best performance under all conditions. Notably, in the most challenging scenario—Gym288-skeleton with severe frame missing—FineTec attains a Top-1 accuracy of 78.1%, surpassing all previous skeleton-based methods. In terms of mean class accuracy, FineTec improves upon the best baseline by 13%, and outperforms the latest work (Xie et al. 2025) by 50%. Overall, these results demonstrate that FineTec achieves outstanding effectiveness across fine-grained datasets and under all levels of difficulty.

Results on Coarse-grained Action Recognition. To further evaluate the generalization and robustness of FineTec, we conduct experiments on the coarse-grained skeleton action recognition benchmarks, including NTU-60-xsub and NTU-120-xsub, and UCF101 (Soomro, Zamir, and Shah 2012). The results are shown in Table 2, showing that FineTec outperforms all competitive baselines, particularly under the most challenging (severe) condition, achieving Top-1 accuracy improvements of 1.3% on NTU-60-xsub and 1.7% on NTU-120-xsub. Results on UCF101 are provided in the Appendix due to space limitations.

Main Results on Skeleton Restoration. We quantitatively evaluate the ability of FineTec to restore temporally corrupted skeleton sequences, as presented in Table 3. Notably, our method achieves the lowest MPJPE across all corruption levels, substantially outperforming all baselines. Compared to the strongest competing method (SiC-Dyna), MaskICL achieves MPJPE reductions of 43.6% under minor corruption, 27.4% under moderate corruption, and 23.4% under severe corruption. Consistent improvements are also observed for N-MPJPE and MPJVE, where our method achieves the best performance across all settings.

Ablations and Analysis

Ablation Studies of Modules. We validate FineTec's design through ablation studies of its three main modules and their variants on the Gym288-skeleton dataset, evaluated in terms of top-1 accuracy. ❶ Module Ablation: Table 4 demonstrates that removing any FineTec module results in a clear performance drop, confirming the necessity of each component and the effectiveness of the overall design. ❷ Analysis of Skeleton Decomposition: Table 5 shows that combining both S_{dyna} and S_{stat} achieves higher accuracy than using either alone, demonstrating that spatial decomposition and differentiated processing enrich skeleton features and enhance fine-grained action recognition. ❸ Fusion Strategy in Physics-driven module: We compare cross-attention (CA) fusion and MLP-based integration for integrating acceleration cues. CA consistently outperforms MLP in both moderate (0.797 vs. 0.779) and severe (0.781 vs. 0.771) settings (Top-1 Acc.), validating the effectiveness of the selected fusion strategy.

Impact of Sequence Completion on Action Recognition. We investigate how different sequence completion methods affect fine-grained action recognition accuracy, as shown in

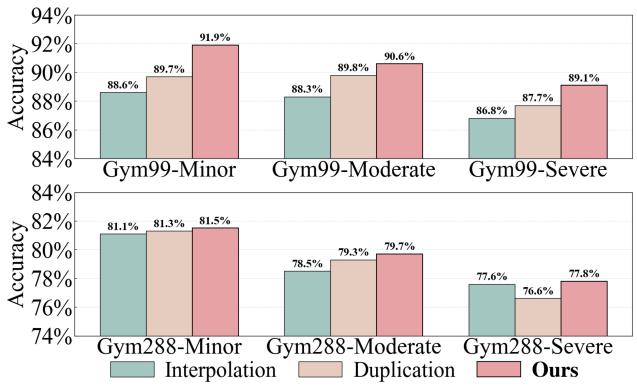

Figure 4: Comparison of Skeleton Restoration Methods on Gym99-skeleton and Gym288-skeleton. Top-1 accuracy of FineTec (Ours), Interpolation, and Duplication is reported under varying levels of temporal corruption.

Perturb.	G99-Min.	G99-Sev.	G288-Min.	G288-Sev.
S-low	0.91/0.985	0.842/0.971	0.790/0.911	0.764/0.905
S-high	0.900/0.983	0.825/0.967	0.785/0.908	0.736/0.884
T-low	0.898/0.988	0.869/0.981	0.791/0.918	0.777/0.915
T-high	0.896/0.987	0.862/0.979	0.789/0.917	0.774/0.914
w/o	**0.926/0.995**	**0.896/0.988**	**0.819/0.928**	**0.804/0.924**

Table 6: Robustness analysis under spatial and temporal perturbations. Yellow rows indicate spatial (Gaussian noise) perturbations, and blue rows indicate temporal (frame dropping) perturbations. Top-1 / Top-5 accuracy are reported.

Figure 4. Compared to standard approaches such as Interpolation and Duplication, FineTec consistently yields the highest Top-1 accuracy on both fine-grained action recognition datasets, across varying levels of temporal degradation. Specifically, on the Gym99 dataset, FineTec achieves Top-1 accuracies of 0.919 (minor), 0.906 (moderate), and 0.885 (severe), demonstrating notable robustness even under significant data loss. The improvements are even more pronounced on the more challenging Gym288 dataset, where FineTec attains Top-1 accuracies of 0.815 (minor), 0.797 (moderate), and 0.778 (severe), outperforming all baselines. In particular, under severe degradation, FineTec significantly surpasses both Duplication and Interpolation. These results show that effective sequence completion significantly benefits action recognition under temporal corruption.

Qualitative Visualization of Skeleton Restoration. Figure 5 presents a qualitative comparison of skeleton sequence restoration for a sample from the Gym288 dataset (Label 122: *"Salto backward stretched with 1.5 twist"*) under severe corruption. We compare our context-aware completion module with an ablation variant that excludes in-context learning. Both methods use identical training settings, yet our approach better reconstructs the missing frames and preserves the fine-grained motion details.

Robustness to Noisy Inputs. We assess the robustness of FineTec under both spatial and temporal perturbations. Spatial noise is introduced by adding Gaussian noise at two severity levels ("S-low" and "S-high"). Temporal robustness is evaluated by randomly dropping half of the input frames, also at two severity levels ("T-low" and "T-high"). As summarized in Table 6, FineTec maintains high recognition accuracy under all perturbations. These results demonstrate FineTec's strong resilience to both spatial and temporal input corruptions, ensuring reliable performance even in challenging real-world scenarios.

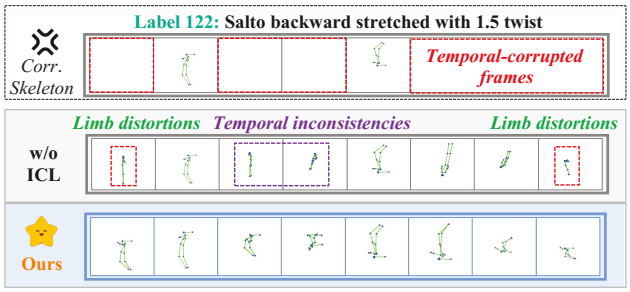

Figure 5: Qualitative Results of Skeleton Restoration. Our context-aware completion method more accurately reconstructs missing frames and preserves fine-grained motion details compared to the ablation without in-context learning.

Discussion and Future Works. FineTec establishes a strong foundation for robust fine-grained action recognition under temporal corruption, demonstrating the potential of detailed spatio-temporal and physics-driven modeling. Building on this foundation, there remain many promising directions for future research. For example, the current use of a fixed skeleton bank in sequence completion and manually defined subgroup partitioning can inspire future work on more adaptive, data-driven approaches. Additionally, extending our joint-level acceleration modeling to subgroup or limb-level dynamics may further enrich the model's understanding of human motion. We believe these directions, building on the foundation established by FineTec, will drive continued progress in robust action recognition.

Conclusion

In this work, we tackle the challenging problem of fine-grained action recognition under temporal corruption. We introduce FineTec, a unified framework specifically designed to address this issue. FineTec first restores temporal continuity from corrupted inputs via a Context-aware Sequence Completion module. It then employs Skeleton-based Spatial Decomposition, guided by biological priors, to partition the skeleton and amplify subtle motion distinctions. Finally, a Physics-driven Acceleration Modeling module leverages Lagrangian dynamics to capture discriminative motion cues beyond simple displacement. Extensive experiments on both coarse-grained and fine-grained benchmarks demonstrate that FineTec consistently outperforms existing methods under varying degrees of temporal corruption. Future directions involve expanding our framework to multi-modal contexts and enhancing biomechanical modeling.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC) under Grant 62306239, and supported by the Sanqin Talents Introduction Plan of Shaanxi Province, China.

References

Andriluka, M.; Tabanpour, B.; Freeman, C. D.; and Sminchisescu, C. 2025. Learned Neural Physics Simulation for Articulated 3D Human Pose Reconstruction. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 320–336. Cham: Springer Nature Switzerland. ISBN 978-3-031-72907-2.

Chen, B.; Jiang, H.; Liu, S.; Gupta, S.; Li, Y.; Zhao, H.; and Wang, S. 2025a. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6178–6189.

Chen, H.; Huang, H.; Yin, X.; and Shao, D. 2025b. FineQuest: Adaptive Knowledge-Assisted Sports Video Understanding via Agent-of-Thoughts Reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 2909–2918. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13359–13368.

Chi, H.-g.; Ha, M. H.; Chi, S.; Lee, S. W.; Huang, Q.; and Ramani, K. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20186–20196.

Chi, S.; Chi, H.-G.; Huang, Q.; and Ramani, K. 2025. InfoGCN++: Learning Representation by Predicting the Future for Online Skeleton-Based Action Recognition . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(01): 514–528.

Duan, H.; Wang, J.; Chen, K.; and Lin, D. 2022a. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7351–7354.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022b. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2969–2978.

Gärtner, E.; Andriluka, M.; Coumans, E.; and Sminchisescu, C. 2022a. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13190–13200.

Gärtner, E.; Andriluka, M.; Xu, H.; and Sminchisescu, C. 2022b. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13106–13115.

Huang, Y.; Chen, H.; Xu, Z.; Jia, Z.; Sun, H.; and Shao, D. 2025. SeFAR: Semi-supervised Fine-grained Action Recognition with Temporal Perturbation and Learning Stabilization. *arXiv preprint arXiv:2501.01245*.

Jain, A.; and Rodriguez, G. 1995. Diagonalized Lagrangian robot dynamics. *IEEE Transactions on Robotics and Automation*, 11(4): 571–584.

Jiang, Y.; and Deng, H. 2024. Lighter and faster: A multi-scale adaptive graph convolutional network for skeleton-based action recognition. *Engineering Applications of Artificial Intelligence*, 132: 107957.

Kim, K.; Park, G.; Lee, Y.; Yeo, W.; and Hwang, S. J. 2025. VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3295–3305.

Leong, M. C.; Zhang, H.; Tan, H. L.; Li, L.; and Lim, J. H. 2022. Combined CNN transformer encoder for enhanced fine-grained human action recognition. *arXiv preprint arXiv:2208.01897*.

Li, T.; Foo, L. G.; Ke, Q.; Rahmani, H.; Wang, A.; Wang, J.; and Liu, J. 2022. Dynamic spatio-temporal specialization learning for fine-grained action recognition. In *European Conference on Computer Vision*, 386–403. Springer.

Lin, L.; Zhang, J.; and Liu, J. 2023. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2363–2372.

Lin, M.; Wang, X.; Wang, Y.; Wang, S.; Dai, F.; Ding, P.; Wang, C.; Zuo, Z.; Sang, N.; Huang, S.; et al. 2025. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*.

Liu, H.; Liu, Y.; Ren, M.; Wang, H.; Wang, Y.; and Sun, Z. 2025. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29248–29257.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2020. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10): 2684–2701.

Liu, S.; Ren, Z.; Gupta, S.; and Wang, S. 2024. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, 360–378. Springer.

Liu, S.-L.; Ding, Y.-N.; Zhang, J.-R.; Liu, K.-Y.; Zhang, S.-F.; Wang, F.-L.; and Huang, G. 2023. Multi-Dimensional Refinement Graph Convolutional Network with Robust Decouple Loss for Fine-Grained Skeleton-Based Action Recognition. arXiv:2306.15321.

Myung, W.; Su, N.; Xue, J.-H.; and Wang, G. 2024. Degcn: Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33: 2477–2490.

Rajendran, M.; Tan, C. T.; Atmosukarto, I.; Ng, A. B.; and See, S. 2024. Review on synergizing the Metaverse and AI-driven synthetic data: enhancing virtual realms and activity recognition in computer vision. *Visual Intelligence*, 2(1): 27.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019.

Shao, D.; Shi, M.; Xu, S.; Chen, H.; Huang, Y.; and Wang, B. 2025. FinePhys: Fine-grained Human Action Generation by Explicitly Incorporating Physical Laws for Effective Skeletal Guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1905–1916.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020a. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2616–2625.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020b. Intra-and inter-action understanding via temporal action parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 730–739.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29: 9532–9545.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *Computer Science*.

Thai Duong, J. S., Abdullah Altawaitan; and Atanasov, N. 2023. Port-Hamiltonian Neural ODE Networks on Lie Groups For Robot Dynamics Learning and Control. *arXiv preprint arXiv:2401.09520*.

Ugrinovic, N.; Pan, B.; Pavlakos, G.; Paschalidou, D.; Shen, B.; Sanchez-Riera, J.; Moreno-Noguer, F.; and Guibas, L. 2024. Multiphys: Multi-person physics-aware 3d motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2331–2340.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2740–2755.

Wang, M.; Huang, Z.; Kong, X.; Shen, G.; Dai, G.; Wang, J.; and Liu, Y. 2025. Action Detail Matters: Refining Video Recognition with Local Action Queries. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19132–19142.

Wang, X.; Fang, Z.; Li, X.; Li, X.; Chen, C.; and Liu, M. 2024. Skeleton-in-context: Unified skeleton sequence modeling with in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2436–2446.

Xie, J.; Meng, Y.; Zhao, Y.; Nguyen, A.; Yang, X.; and Zheng, Y. 2024. Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 6225–6233.

Xie, J.; Zhao, Y.; Meng, Y.; Zhao, H.; Nguyen, A.; and Zheng, Y. 2025. Are Spatial-Temporal Graph Convolution Networks for Human Action Recognition Over-Parameterized? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24309–24319.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yang, C.; Xu, Y.; Shi, J.; Dai, B.; and Zhou, B. 2020. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 591–600.

Zhang, C.; Gupta, A.; and Zisserman, A. 2021. Temporal query networks for fine-grained video understanding. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 4486–4496.

Zhang, H.; Leong, M. C.; Li, L.; and Lin, W. 2024a. PGVT: Pose-Guided Video Transformer for Fine-Grained Action Recognition. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6631–6642.

Zhang, Y.; Kephart, J. O.; Cui, Z.; and Ji, Q. 2024b. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2305–2317.

Zhang, Y.; Kephart, J. O.; and Ji, Q. 2024. Incorporating Physics Principles for Precise Human Motion Prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6164–6174.

Zhang, Z.; Zhu, Y.; Rai, R.; and Doermann, D. 2022. PIM-Net: Physics-Infused Neural Network for Human Motion Prediction. *IEEE Robotics and Automation Letters*, 7(4): 8949–8955.

Zheng, G.; Lin, S.; Zuo, H.; Fu, C.; and Pan, J. 2024. Nettrack: Tracking highly dynamic objects with a net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19145–19155.

Zhou, Y.; Yan, X.; Cheng, Z.-Q.; Yan, Y.; Dai, Q.; and Hua, X.-S. 2024. BlockGCN: Redefining Topology Awareness for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, A.; Zhu, J.; Bailey, J.; Gong, M.; and Ke, Q. 2025. Semantic-guided Cross-Modal Prompt Learning for Skeleton-based Zero-shot Action Recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13876–13885.