Decomposing Prompts, Composing Actions: A Multi-Granularity Prompting Approach for Incremental Action Learning

Xinyi Cheng[1], Chenghao Xu[2], Xi Wang[2], Jiexi Yan[1*], Yanhua Yang[1*]

[1]School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[2]School of Electronic Engineering, Xidian University, Xi'an 710071, China
{xinyicheng, chx}@stu.xidian.edu.cn, {wangxi6317, jxyan1995}@gmail.com, yanhyang@xidian.edu.cn

Abstract

Continual learning for action recognition is a critical capability for next-generation Extended Reality (XR) systems. Yet it faces a severe real-world challenge: strict user privacy that prohibits data rehearsal. While recent prompt-based continual learning methods show promise, we argue their core 'flat,' single-granularity design fundamentally misaligns with the complexity of human actions. This monolithic architecture fails to model the inherent hierarchical structure and overlooks standard action primitives shared across tasks, resulting in suboptimal performance and hindered knowledge transfer. To overcome this limitation, we propose DPCA, a novel spatio-temporal continual learning framework with multi-granularity adaptive prompting. DPCA learns three synergistic components to resolve this mismatch. First, the task-specific prompter employs a multi-granularity query system to capture the unique, compositional semantics of each action. Second, the task-agnostic prompter learns a globally shared vocabulary of "action primitives," providing a stable and generalizable knowledge base to mitigate catastrophic forgetting. Finally, we introduce a Dissimilarity Attention Rectification at each granularity level, leveraging a reverse attention mechanism to model class-agnostic background information and effectively alleviating overfitting. The synergy between these components enables robust model adaptation without requiring access to past data. Rigorous experiments on multiple large-scale benchmarks (including NTU RGB+D), under a strict rehearsal-free, few-shot protocol, confirm that DPCA establishes a new state-of-the-art. This advance paves the way for the realization of truly adaptive and privacy-respecting XR systems.

Introduction

Extended Reality (XR), as the next-generation computing platform, demands continuous, adaptive understanding of user actions. This necessitates models that sequentially learn new dynamic behaviors post-deployment, a challenge formally known as Class-Incremental Learning (CIL) (Rebuffi et al. 2017; Zhou et al. 2024; Mittal, Galesso, and Brox 2021). However, applying CIL to real-world XR systems introduces a more specific and formidable problem: Continual Action Recognition (CAR). CAR is uniquely defined by two stringent, non-negotiable constraints: (1) a strict, privacy-preserving paradigm (rehearsal-free) due to the sensitive, biometric nature of skeletal data, and (2) the necessity of a few-shot learning regime due to the impracticality of extensive user annotation. These dual constraints, far exceeding standard CIL benchmarks, render models exceptionally vulnerable to catastrophic forgetting (McCloskey and Cohen 1989). This performance decay on previously learned actions fundamentally undermines the application's reliability and the user's immersive experience.

In recent years, Prompt-based Continual Learning (PCL), a promising rehearsal-free paradigm from large-scale pre-trained models, has shown immense potential (Bowman et al. 2023; Razdaibiedina et al. 2023). Methods like L2P (Wang et al. 2022b), DualPrompt (Wang et al. 2022a), and CODA-Prompt (Smith et al. 2023) steer a frozen backbone using small, trainable prompt vectors, achieving strong performance in image CIL. This paradigm was extended to Graph Neural Networks (GNNs) (Fang et al. 2023; Wu et al. 2020) and subsequently to motion CIL by pioneers like POET (Garg et al. 2024). However, we argue this entire line of work shares a fundamental design limitation: its prompts are flat, single-granularity modules. This monolithic design is ill-suited for complex human actions, which are characterized by two fundamental properties. First, actions possess an inherent Hierarchical Structure (e.g., a "throw" comprises fine-grained joint rotations, limb swings, and holistic posture). Single-granularity prompts struggle to capture these multilevel semantics from only a few samples. Second, actions often share common Motion Primitives (e.g., "swinging an arm"). By learning independent prompts per task, current methods fail to build a compositional knowledge base of these motion priors, impeding knowledge transfer and exacerbating catastrophic forgetting.

To address these limitations, we propose DPCA, a multi-granularity prompting framework for class-incremental action recognition. DPCA decomposes learning into three synergistic components to model both action individuality and commonality. First, a task-specific prompter acts as a multi-granularity query system to capture the unique, hierarchical structure of each new action, enabling data-efficient learning of discriminative patterns from few samples. Second, a task-agnostic prompter learns a globally shared, compositional vocabulary of motion primitives (e.g., swinging,

*Corresponding author.

twisting). This component provides a stable, generalizable knowledge base, promoting positive knowledge transfer and anchoring against catastrophic forgetting. Third, to counter overfitting in the few-shot regime, we introduce a Dissimilarity Attention Rectification module. It uses a reverse attention mechanism to model and balance class-agnostic, non-discriminative information at each granularity level, enhancing stability. The synergy of these components enables DPCA to achieve a state-of-the-art stability-plasticity balance.

In summary, our main contributions are as follows.

- We are the first to identify and analyze a fundamental limitation of current prompt-based methods: their 'flat,' single-granularity design fails to model the hierarchical structure of actions and leverage shared motion priors.
- We design DPCA, a novel framework that enables fine-grained and robust GNN adaptation through the synergy of three components: a task-specific multi-granularity prompter, a task-agnostic dynamic basis prompter, and a novel Dissimilarity Attention Rectification.
- Extensive experiments on multiple large-scale benchmarks (including NTU RGB+D), under a stringent rehearsal-free, few-shot setting, conclusively demonstrate that our method establishes a new state-of-the-art, paving the way for intelligent and privacy-respecting XR interaction systems.

Related Work

Skeleton-Based Action Recognition

While early deep learning approaches for action recognition focused on video data (Wu et al. 2020; Simonyan and Zisserman 2014; Park, Kang, and Han 2021), the field has increasingly shifted towards skeleton-based methods. The inherent robustness of skeleton data to appearance changes and its capacity to preserve user privacy drove this transition, leading to a surge in dedicated research (Yan, Xiong, and Lin 2018a; Duan et al. 2022; Du, Wang, and Wang 2015a; Xu et al. 2024a). Initially, this subfield achieved considerable success by adapting models like CNNs (Li et al. 2017; Du, Fu, and Wang 2015) and RNNs (Du, Wang, and Wang 2015b; Liao et al. 2021) to this new modality. However, the advent of Graph Convolutional Networks (GCNs) (Kipf and Welling 2016) marked a significant paradigm shift, as they are uniquely suited to model the inherent non-Euclidean graph structure of the human skeleton. This overcame the limitations of prior methods, which struggled to capture the complex, non-grid-like connections between joints. The pioneering ST-GCN (Yan, Xiong, and Lin 2018b) was the first to successfully apply this concept, learning spatio-temporal features directly on skeleton graphs. This foundational approach paved the way for subsequent advancements like CTR-GCN (Chen et al. 2021) and InfoGCN (Chi et al. 2022), which further refined performance by introducing more sophisticated, adaptive, and context-aware graph modeling techniques.

Continual Learning

Continual learning (CL) addresses the critical challenge of catastrophic forgetting, where a model's performance on previously learned tasks degrades upon learning new ones. Early approaches centered on memory replay, which either rehearses real data subsets (Wu et al. 2019, 2018; Kemker and Kanan 2017) or employs generative models to synthesize past data (Shin et al. 2017; Van de Ven, Siegelmann, and Tolias 2020; Gao and Liu 2023; Xu et al. 2024b; Xu, Yan, and Deng 2025). Regularization-based methods (Kang, Park, and Han 2022; Chen et al. 2021; Li and Hoiem 2017) were proposed as a data-free alternative, penalizing updates to model weights deemed critical for previous tasks. Recently, Prompt-based Continual Learning (PCL) (Bowman et al. 2023; Razdaibiedina et al. 2023) has emerged as a promising paradigm that circumvents this dilemma. By learning and storing only a small set of prompts instead of raw data, it is possible to overcome catastrophic forgetting.

Method

In this section, we will first introduce the task setting for Action Class-Continual Learning. We will then discuss our proposed DPCA method in detail, as illustrated in Fig.. 1. Next, we will outline the training and inference processes.

Problem Formulation

This work addresses the problem of Class-Incremental Learning (CIL) for skeleton-based action recognition. CIL is a challenging subfield of Continual Learning (CL) where a model must learn from a sequence of tasks arriving over time, with each task introducing new classes. Formally, we consider a sequence of $K+1$ tasks, $\{\mathbf{T}^0, \mathbf{T}^1, \ldots, \mathbf{T}^K\}$. Each task $\mathbf{T}^k$ is defined by a dataset $D^k = \{(\mathbf{X}_i^k, y_i^k)\}$, where $\mathbf{X}_i^k \in \mathbb{R}^{T\times J\times 3}$ represents a skeleton sequence of T frames with J joints in 3D coordinates, and $y_i^k \in C^k$ is the corresponding class label. A core stipulation of the CIL setting is that these class sets are mutually exclusive, i.e., $C^k \cap C^{k'} = \emptyset$ for any $k \neq k'$. We adhere to this standard definition as it aligns with the established protocol in the literature (e.g., POET), enabling a fair and rigorous comparison. While real-world applications might involve overlapping classes, this mutually exclusive setting is the necessary benchmark to precisely isolate and measure a model's vulnerability to catastrophic forgetting. Our method models the input as a spatio-temporal graph $G = (V, E)$, where the model is denoted as $f(\cdot) = f_c \circ f_g(\cdot)$. The vertices V represent skeletal joints, and the edges E define both intra-body structure within a frame and inter-frame temporal connections. After training in task $\mathbf{T}^k$, the performance of the model is evaluated on all classes seen so far $C^{all} = \bigcup_{i=0}^{k} C^i$. The primary challenge is to assimilate knowledge from new classes while mitigating catastrophic forgetting, which is the severe performance degradation of previously learned knowledge.

Prompt Pool Design

Our proposed DPCA framework, as illustrated in Fig.. 1, consists of three main synergistic components: a Task-

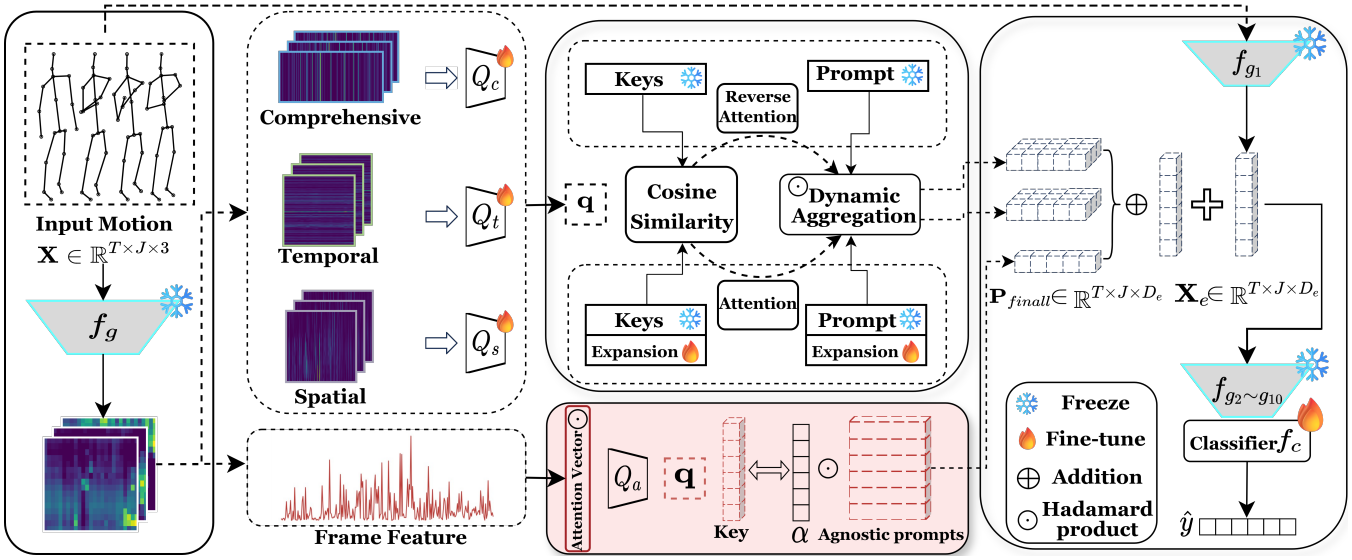

Figure 1: The DPCA framework synergizes three components for continual learning: (1) a multi-granularity Task-Specific Prompter captures unique action semantics; (2) a shared Task-Agnostic Prompter learns common "action primitives" to mitigate catastrophic forgetting; and (3) a Dissimilarity Attention Rectification filters non-discriminative features. The resulting prompts are injected into a frozen backbone, enabling a robust stability-plasticity balance for rehearsal-free learning.

Specific Prompt Pool designed to capture unique, hierarchical action semantics; a shared Task-Agnostic Prompt Pool that learns a global vocabulary of "action primitives"; and a Dissimilarity Attention Rectification module to filter non-discriminative features. These components work in concert, and we will detail each in the following subsections.

Task-Specific Prompting. To mitigate catastrophic forgetting while maintaining plasticity for new tasks, our core strategy is to adapt a frozen pre-trained backbone, f_g, using learnable task-specific prompts. For each new task $\mathbf{T}_k$ instead of a single prompt, we introduce a multi-granularity prompting mechanism designed to capture diverse features of skeleton-based actions. Specifically, the prompts for the task $\mathbf{T}_k$ denoted $\mathbf{P}_{specific}$ consist of three distinct prompt pools:

$$\mathbf{P}_{specific} = \{\mathbf{P}_c, \mathbf{P}_t, \mathbf{P}_s\}, \tag{1}$$

Each prompt pool, $\mathbf{P}_j$ (where $j \in \{s, t, c\}$), is composed of a set of N_j prompt embeddings $\{P_i^j\}_{i=1}^{N_j}$, where $P_i^j \in \mathbb{R}^{L_j\times D_j}$, and a corresponding set of queryable keys $\{K_i^j\}_{i=1}^{N_j}$, where $K^j \in \mathbb{R}^{N_j\times D_j}$. These pools correspond to spatial, temporal, and comprehensive granularities, respectively. The spatial and temporal prompts aim to capture fine-grained local joint dependencies and motion evolution, while the comprehensive prompt is designed to encode global, holistic action semantics. To achieve this functional specialization despite the inherent entanglement of spatio-temporal features, our architecture employs a targeted pooling strategy before generating the queries. Specifically, the query network Q_t generates its query by first pooling features across all spatial dimensions, compelling it to focus on the temporal signal. Conversely, Q_s generates its query by pooling across the temporal dimension (T frames), biasing it to focus on spatial configuration.

Rather than selecting a single prompt, we employ an attention-based dynamic aggregation mechanism to generate an instance-specific prompt for each input sample $\mathbf{X}_i$. For instance, at the comprehensive granularity ($j = c$), an input sample $\mathbf{X}_i$ is first passed through the frozen backbone f_g to obtain its deep feature representation. implemented as a lightweight MLP, with full architectural details in the Appendix) then maps this feature to a query vector:

$$q_{i,c} = Q_c(f_g(\mathbf{X}_i)),\ q_{i,c} \in \mathbb{R}^{D_c}. \tag{2}$$

The query $q_{i,c}$ is compared against all keys in the comprehensive key matrix K^c to compute attention scores. These scores, normalized via a softmax function, represent the relevance of each stored prompt to the current input.

$$\alpha_{i,c} = softmax(\frac{q_{i,c}(K^c)^T}{\sqrt{D_c}}),\ \alpha_{i,c} \in \mathbb{R}^{N_c}. \tag{3}$$

The final comprehensive prompt for the instance, $\mathbf{p}_c$, is computed as the weighted sum of all prompt embeddings in the pool $\mathbf{P}_c$:

$$\mathbf{p}_c = \sum_{m=1}^{N_c} \alpha_{i,c,m} \cdot P_m^c,\ \mathbf{p}_c \in \mathbb{R}^{L_c\times D_c} \tag{4}$$

Analogous operations are performed to generate the spatial prompt $\mathbf{p}_s$ and the temporal prompt $\mathbf{p}_t$. The three prompts are then fused via element-wise addition to form the final specific prompt.

$$\begin{aligned}\mathbf{p}_{specific} &= f(\mathbf{p}_c, \mathbf{p}_t, \mathbf{p}_s) \\ &= \mathbf{p}_c + \mathbf{p}_t + \mathbf{p}_s.\end{aligned} \tag{5}$$

Crucially, the prompt pools are kept isolated across tasks. When training on task $\mathbf{T}^k$, only the parameters of its corresponding pools $\{\mathbf{P}_c^k, \mathbf{P}_s^k, \mathbf{P}_t^k\}$ are available. All previously learned prompt pools remain frozen, thereby preventing catastrophic forgetting while ensuring the model is end-to-end differentiable for the current task.

Task-Agnostic Prompting. While task-specific prompts excel at capturing the unique characteristics of each task, they do not explicitly model the shared knowledge across different tasks. To facilitate cross-task knowledge transfer and improve generalization, we introduce a complementary, shared Task-Agnostic Prompt Pool, $\mathbf{P}_a$. The objective of this shared pool is to distill a set of universal "action primitives", fundamental, reusable spatio-temporal patterns common to multiple actions (e.g., the 'arm lifting' component in both 'waving' and 'throwing'). This pool builds upon the same query-key-prompt architecture. Still, it introduces a key innovation: a Weighted Query Attention mechanism designed to isolate these universal patterns from noisy, instance-specific features. To achieve this, we augment each entry in the task-agnostic pool with a third learnable component: a gating attention vector, $A_i^a \in \mathbb{R}^{D_a}$. Instead of using the raw query q_a for similarity matching, we first use q_a to query the gating matrix A^a and compute an attention-based gate vector $\mathbf{g}_a$:

$$\alpha_{gate} = \text{softmax}(\frac{q_a(A^a)^T}{\sqrt{D_a}}) \quad \mathbf{g}_a = \sum_{m=1}^{N_a} \alpha_{gate,m} \cdot A_m^a \quad (6)$$

This gate vector $\mathbf{g}_a$ then modulates the raw query q_a via element-wise multiplication ($\odot$) to produce the final weighted query q_a':

$$q_a' = q_a \odot \mathbf{g}_a \quad (7)$$

This weighted query q_a' is then used for the subsequent similarity matching against the keys K^a.

Furthermore, to ensure that the learned "action primitives" are diverse and functionally distinct, we introduce an Action Primitive Codebook Regularization loss, $\mathcal{L}_{cb}$. This loss encourages the key and gating attention matrices (K^a, A^a) to form a near-orthonormal basis, which in turn compels each primitive to have both a unique location in the feature space and a distinct functional focus, thus creating a more expressive "codebook" of primitives.

$$\mathcal{L}_{cb} = \|K^a(K^a)^T - I\|_F^2 + \|A^a(A^a)^T - I\|_F^2 \quad (8)$$

Where $K^a, A^a \in \mathbb{R}^{N_a \times D_a}$, I is the identity matrix, and $||\cdot||_F^2$ is the squared Frobenius norm.

Dissimilarity Attention Rectification. While task-specific prompts are designed to capture the core discriminative patterns of an action, skeleton data contains significant intra-class variations that can harm generalization. These non-discriminative features stem from performer-specific attributes (e.g., body shape), variations in execution style (e.g., speed, amplitude), or minor movements in irrelevant body parts. To address this, we introduce Dissimilarity Attention Rectification (DAR), defined as $AR^j = (\{P_i^j\}_{i=1}^{N_j}, \{K_i^j\}_{i=1}^{N_j})$, which explicitly models and balances these non-essential features. Inspired by the use of reverse attention for modeling background information and non-content features, the AR operates in parallel with the task-specific prompts at each granularity $j \in \{c, s, t\}$. It employs a reverse attention mechanism to focus on features that are dissimilar to the query actively. Specifically, it assigns higher weights to query-key pairs with lower similarity. The attention scores are calculated as:

$$\alpha_j' = \max(0, -\text{sim}(q_j, K^j)), \quad j \in \{s, t, c\}. \quad (9)$$

By learning to represent these structured, non-discriminative variations, the DAR effectively reduces the intra-class feature variance. The final prompt passed to the model is a combination of the task-specific prompt and this new rectification prompt.

$$\mathbf{p}'_{specific} = \mathbf{p}_{specific} + \lambda \mathbf{p}_{DAR}. \quad (10)$$

where λ is a hyperparameter that balances the contribution of DAR. For brevity, $\mathbf{p}_{specific}$ will henceforth refer to $\mathbf{p}'_{specific}$ unless otherwise specified.

Optimizing with cross-entropy $\mathcal{L}_{cls}$ alone can compel the task-specific prompts to overfit to noise in pursuit of better class separation, thereby harming the model's generalization capability. To mitigate this, we proposed Prompt Semantic Alignment loss, $\mathcal{L}_{PSA}$, which encourages the prompts generated for all samples within the same class to be semantically similar. By minimizing the average cosine distance between all positive prompt pairs within a batch, we create a more structured and robust semantic space. Formally, given a batch of samples, we define the set of positive pairs as $\mathcal{P} = \{(i, j) | y_i = y_j, i \neq j\}$. The PSA loss is then:

$$\mathcal{L}_{PSA} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} (1 - \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{||\mathbf{p_i}||_2 ||\mathbf{p}_j||_2}), \quad (11)$$

where $\mathbf{p}_i$ represents the task-specific prompt for sample $\mathbf{x}_i$. This regularization term compels the model to ignore spurious, non-generalizable features and focus on the core, transferable spatio-temporal patterns that truly define the action. This not only enhances model generalization for seen classes but is also crucial for adapting to new classes in our continual learning setting.

Learning Procedure and Inference Process

Learning Procedure. Our model's training follows a two-phase protocol. Initially, during the base training phase on task $\mathbf{T}^0$, we leverage the large-scale base dataset D_0 to establish a strong foundation by training a robust feature backbone, $f_g(\cdot)$, while concurrently instantiating the initial task-specific prompt pool, $\mathbf{P}^0_{specific}$, and the globally shared task-agnostic prompt pool $\mathbf{P}_{agnostic}$. For all subsequent few-shot tasks $\mathbf{T}^k$ (where $k > 0$), the model enters an incremental learning phase designed to mitigate catastrophic forgetting. In this phase, we adopt a parameter-efficient strategy where the feature backbone f_g, all previously learned task-specific pools, and the entire shared task-agnostic module ($\mathbf{P}_{agnostic}$ and its adapters) are entirely frozen.

The trainable parameters are thus strictly limited to the new, lightweight task-specific prompt pool $\mathbf{P}^k_{specific}$ and the

	Task 0	Task 1	Task 2	Task 3	Task 4			
Method	Avg (↑)	Avg (↑)	Avg (↑)	Avg (↑)	Old (↑)	New (↑)	Avg (↑)	A_{HM} (↑)
Continual Linear Probing								
FE	88.4	72.0 ± 1.1	60.4 ± 2.4	47.7 ± 2.1	40.0 ± 1.6	51.0 ± 2.3	40.9 ± 1.4	44.8 ± 1.1
FE, Frozen	88.4	76.1 ± 1.0	52.4 ± 4.1	38.3 ± 2.7	28.4 ± 1.6	22.4 ± 4.5	27.9 ± 1.4	24.8 ± 3.0
FE+Replay†	88.4	72.0 ± 1.5	59.5 ± 4.0	58.7 ± 2.8	56.7 ± 2.5	34.7 ± 5.6	54.9 ± 2.7	42.8 ± 4.4
FT	88.4	6.2 ± 1.4	4.3 ± 1.5	2.8 ± 1.0	0.2 ± 0.5	36.0 ± 10.1	3.2 ± 0.8	0.3 ± 1.0
Standard Continual Learning								
LWF (Li and Hoiem 2017)	88.4	6.2 ± 1.5	2.8 ± 0.7	3.7 ± 1.3	0.0 ± 0.0	38.9 ± 8.8	3.2 ± 0.7	0.0 ± 0.0
EWC (Kirkpatrick et al. 2017)	88.4	6.6 ± 1.5	4.1 ± 1.4	3.1 ± 0.9	0.0 ± 0.0	42.1 ± 9.5	3.5 ± 0.8	0.0 ± 0.0
Experience Replay	88.4	35.1 ± 8.3	50.6 ± 5.0	60.6 ± 5.4	54.6 ± 6.5	43.7 ± 14.6	53.7 ± 7.1	47.8 ± 11.2
Experience Replay†	88.4	6.2 ± 1.5	9.0 ± 2.6	11.2 ± 3.0	10.9 ± 2.6	34.6 ± 7.9	12.9 ± 3.0	16.3 ± 3.5
LUCIR (Hou et al. 2019)	87.9	4.3 ± 2.1	4.1 ± 1.3	2.7 ± 0.8	0.2 ± 0.4	26.0 ± 9.2	2.3 ± 0.9	0.4 ± 0.8
Continual Prompt Tuning								
CODA-P (Smith et al. 2023)	87.4	76.1 ± 1.0	66.7 ± 1.3	58.6 ± 2.7	56.5 ± 2.9	0.5 ± 0.4	51.8 ± 2.7	1.1 ± 0.7
L2P (Wang et al. 2022b)	88.6	78.9 ± 0.1	71.0 ± 1.0	64.2 ± 0.1	62.0 ± 0.7	0.0 ± 0.0	56.8 ± 0.6	0.0 ± 0.0
APT (Bowman et al. 2023)	86.6	27.3 ± 1.6	30.8 ± 3.4	37.6 ± 2.3	NA	33.4 ± 2.0	NA	NA
POET (Garg et al. 2024)	87.9	82.3 ± 0.6	76.8 ± 0.9	68.4 ± 0.7	57.2 ± 1.0	**55.8 ± 5.9**	57.1 ± 1.1	56.3 ± 3.2
DPCA (Ours)	**88.9**	**84.3 ± 0.8**	**77.9 ± 1.2**	**68.6 ± 1.0**	**60.4 ± 0.9**	55.3 ± 5.5	**60.0 ± 1.3**	**57.7 ± 3.3**

Table 1: **Activity Recognition Results (%, ↑), Comparison with SOTA:** NTU RGB+D dataset on CTR-GCN backbone. After training on each incremental task, we report Average of all classes seen so far ('Avg'). We also report (i) A_{HM}, (ii) old classes accuracy ('Old'), (iii) new classes accuracy ('New') in the last session.

new classifier head. The query adapters (e.g., Q_c, Q_s, Q_t) for all prompt pools, having been trained on the base task, remain frozen. This approach isolates new knowledge within compact task-specific modules, anchoring them to the stable, shared knowledge base to preserve past information effectively.

Inference Process. During inference, the model leverages its entire learned knowledge base to classify a given sample. For an input $\mathbf{X}$, the shared, frozen query network first generates multi-granularity task-specific and task-agnostic queries. Critically, the model then performs a global search, matching these queries against the task-specific prompt pools from all previously learned tasks (from $\mathbf{T}^0$ to $\mathbf{T}^K$) and the task-agnostic pool. The resulting prompts, $\mathbf{p}_{specific}$ and $\mathbf{p}_{agnostic}$, are injected into the first layer of the backbone via additive feature perturbation as defined in Equation (12).

$$\begin{aligned} \mathbf{x}_e &= f_{g_1}(\mathbf{X}) \\ \mathbf{x}'_e &= \mathbf{x}_e + \mathbf{p}_{specific} + \mathbf{p}_{agnostic}. \end{aligned} \tag{12}$$

Finally, these perturbed features are processed by the remainder of the backbone and the classifier to yield the final prediction.

Loss Function

To effectively address the inherent trade-off between plasticity and stability in few-shot continual learning, we design a composite loss function, $\mathcal{L}_{total}$. This function is composed of three key components: (1) a classification loss, $\mathcal{L}_{cls}$, to ensure **plasticity** by learning new categories; (2) the Action Primitive Codebook Regularization loss $\mathcal{L}_{cb}$ (8); and (3) the Prompt Semantic Alignment loss $\mathcal{L}_{PSA}$ (11). **The latter two losses act as regularizers** to enhance prompt diversity and structure, providing crucial **stability** and generalization against few-shot overfitting. The total loss function balances these objectives:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{cb}\mathcal{L}_{cb} + \lambda_{PSA}\mathcal{L}_{PSA} \tag{13}$$

where λ_{cb} and λ_{PSA} are the hyperparameters that balance the contribution of each loss term.

Experiments and Results

Experiment Setup

Datasets. We conduct our primary analysis on the NTU RGB+D dataset (Shahroudy et al. 2016), following the official Cross-Subject (CS) benchmark. To validate the generality of our method, we also report performance on the PKU-MMD and SHREC-2017 benchmarks.

Backbone and Protocol. Our main feature backbone is CTR-GCN (Chen et al. 2021). We also include results with ST-GCN (Yan, Xiong, and Lin 2018b) to demonstrate architectural robustness. We adopt the stringent, privacy-aware, few-shot protocol from POET (Garg et al. 2024). For the NTU dataset, this splits the 60 action classes into a 40-class base set and a 20-class incremental set. The incremental set is learned sequentially over 4 tasks in a 5-way, 5-shot setting.

Input Representation. Unless otherwise specified, each input sample is a 3D skeleton sequence of 64 temporal frames with 25 joints, represented as a tensor of shape $\mathbb{R}^{64\times25\times3}$.

Evaluation Metrics. Following standard CIL protocols, we evaluate performance using several key metrics. We report the average accuracy across all seen classes (A_{avg}), though this metric is known to be biased towards the large base set in our few-shot setting. Therefore, to precisely measure the critical stability-plasticity trade-off, we focus on the

accuracy on previously learned ("old") classes (A_{old}), which measures model **stability**, and the accuracy on new classes (A_{new}), which measures **plasticity**. Our primary metric for quantifying this balance is the Harmonic Mean (A_{hm}). Unlike a simple arithmetic mean (which can be misleadingly high if one metric is perfect and the other is zero), the harmonic mean heavily penalizes imbalance. It is therefore the most rigorous metric for this trade-off, as a high A_{hm} value indicates that the model is simultaneously strong in both retaining old knowledge (A_{old}) and adapting to new classes (A_{new}). It is defined as:

$$A_{hm} = \frac{2A_{old} \cdot A_{new}}{A_{old} + A_{new}}. \tag{14}$$

A higher A_{hm} value indicates a better balance between retaining old knowledge and adapting to new classes.

Implementation Details. To ensure a fair comparison, our experiments follow the protocol and baseline settings established by POET (Garg et al. 2024). We first pre-train the CTR-GCN backbone on the base dataset D^0 for 50 epochs. In the subsequent base training phase, the prompt components and their corresponding query networks (implemented as lightweight MLPs) are optimized. We use a batch size of 64, with an initial learning rate of 0.1 for the prompts and 0.01 for the query networks. During the few-shot incremental stage, we use a lower learning rate of 0.05 and train for only 5 epochs. We observed that standard classifiers are susceptible to catastrophic forgetting; consequently, we utilize a cosine classifier for all experiments. All models are trained using Stochastic Gradient Descent (SGD). Our method is highly parameter-efficient. Each new task adds only ∼50K trainable parameters (less than 1.5% of the backbone) and introduces less than 3% inference overhead.

Main Results

Baselines. To evaluate our method, we conduct a comprehensive comparison against a wide range of baselines on the NTU RGB+D dataset, following the rigorous protocol established by POET. The baselines include several categories: Continual Linear Probing methods, such as full Fine-Tuning (FT) and Feature Extraction (FE), which represent the extremes of either updating the entire model or freezing the backbone completely. Standard Continual Learning strategies like LwF (Li and Hoiem 2017), EWC (Kirkpatrick et al. 2017), and LUCIR (Hou et al. 2019), which apply regularization to backbone updates. Continual Prompt Tuning approaches, including the previous state-of-the-art POET (Garg et al. 2024), and adaptations of methods like L2P (Wang et al. 2022b), CODA-P (Smith et al. 2023), and APT (Bowman et al. 2023), which were originally designed for image-based tasks. As shown in Table.1, DPCA establishes a new state-of-the-art in the stringent rehearsal-free, few-shot setting. It achieves a superior stability-plasticity balance, with a state-of-the-art harmonic mean accuracy (A_{HM}) of 57.7%.

Most notably, DPCA surpasses the previous state-of-the-art, POET. This superiority stems from our compositional, multi-granularity design, which moves beyond POET's flat

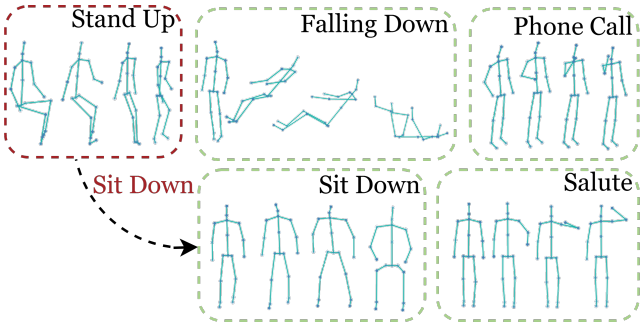

Figure 2: Visualization of the "sit down" action (red box) with other kinematically similar motions, illustrating a key challenge for traditional, single-granularity models.

temporal prompting. By learning a hierarchical vocabulary of spatio-temporal patterns and shared "action primitives" (via our task-agnostic prompter), DPCA achieves a more robust understanding. This allows it to distinguish kinematically similar actions (e.g., "sit down" vs. "stand up" as shown in Fig. 2). The robustness of this compositional design is further validated on two additional benchmarks (PKU-MMD and SHREC-2017) and with a different backbone (ST-GCN), where DPCA consistently outperforms POET. DPCA also significantly outperforms other prompt-based methods (L2P, CODA-P, APT). Their static query mechanisms fail in the few-shot regime, leading to poor plasticity (A_{new}). In contrast, our dynamic, multi-granularity query system(task-specific prompter) excels at generating instance-specific prompts, enabling efficient learning from scarce data. Furthermore, DPCA proves more robust than standard CL strategies (LwF, EWC, LUCIR). These backbone-tuning methods suffer severe catastrophic forgetting by overfitting to new few-shot tasks. Our backbone-frozen approach, which tunes only lightweight prompts, maintains high fidelity on previous tasks.

Ablations and Analysis

To validate the effectiveness of our proposed framework, we conduct a series of ablation studies on the NTU RGB+D dataset. These experiments (summarized in Table.2) are designed to isolate and analyze the specific contributions of DPCA's key components: the multi-granularity design (P_c, P_t, P_s), the task-agnostic prompter (P_a), and the Dissimilarity Attention Rectification (DAR).

Task-Specific Prompting

We conducted ablations to verify the necessity of our multi-granularity design (P_c, P_t, P_s). As shown in Table.2, the full model, which combines all three granularities, achieves the top A_{hm} of 57.7%. This confirms the design's effectiveness, as ablating any component results in a significant performance collapse. For instance, using only a single comprehensive prompt (w/o P_s&P_t) drops the A_{hm} to 40.6%, demonstrating that a single-granularity prompt is insufficient. Furthermore, the severe performance drops from removing only the temporal pool (w/o P_t, 46.8% A_{hm}) or

	Task 0	Task 1				Task 2				Task 3				Task 4			
Method	Avg	Old	New	Avg	A_{HM}	Old	New	Avg	A_{HM}	Old	New	Avg	A_{HM}	Old	New	Avg	A_{HM}
w/o $\mathbf{P}_a$	88.7	86.0	62.1	83.4	72.1	77.8	57.5	75.4	66.1	67.2	52.8	65.7	59.1	57.9	50.7	57.1	54.1
w/o $\mathbf{P}_t$	88.1	84.2	53.5	80.4	65.4	75.1	48.2	71.7	58.7	64.5	44.1	61.8	52.4	54.9	40.8	53.4	46.8
w/o $\mathbf{P}_s$	88.5	85.5	60.1	82.5	70.6	77.4	53.8	74.6	63.5	66.9	49.5	65.0	56.9	57.2	47.6	56.3	52.0
w/o $\mathbf{P}_{DAR}$	**89.0**	86.8	**62.1**	84.0	**73.3**	78.5	59.1	76.2	67.4	68.0	57.0	66.9	62.0	58.8	54.8	58.4	56.7
w/o $\mathbf{P}_s$&$\mathbf{P}_t$	87.2	83.0	45.1	77.9	58.4	73.5	40.2	69.8	52.0	62.1	35.6	59.9	45.3	52.5	33.1	50.8	40.6
DPCA (Full)	88.9	**87.1**	61.4	**84.3**	72.0	**80.0**	**58.3**	**77.9**	**67.4**	**69.8**	**56.8**	**68.6**	**62.6**	**60.4**	**55.3**	**60.0**	**57.7**

Table 2: **Ablations and Analysis on NTU RGB+D dataset (%, ↑):** 'w/o' denotes removing that component from DPCA. This revised table presents more realistic performance trade-offs, where the cumulative effect of each component's absence becomes more pronounced over the continual learning sequence. A_{HM} remains the key indicator for the stability-plasticity balance.

Method	PKU-MMD	SHREC-2017	NTU (ST-GCN)
POET	55.2	55.8	54.3
DPCA (Ours)	**56.5** (+1.3)	**57.3** (+1.5)	**55.7** (+1.4)

Table 3: Generality analysis on additional datasets and backbones (A_{HM}, %).

only the spatial pool (w/o P_s, 52.0% A_{hm}) reveal that the pools are not redundant. This confirms that each granularity captures unique and complementary spatio-temporal information, which is crucial for achieving a robust stability-plasticity balance.

Task-Agnostic Prompting

We conducted ablations to verify the necessity of our multi-granularity design (P_c, P_t, P_s). As shown in Table.2, the full model, which combines all three granularities, achieves the top A_{hm} of 57.7%. This confirms the design's effectiveness, as ablating any component results in a significant performance collapse. For instance, using only a single comprehensive prompt (w/o P_s&P_t) drops the A_{hm} to 40.6%, demonstrating that a single-granularity prompt is insufficient. Furthermore, the severe performance drops from removing only the temporal pool (w/o P_t, 46.8% A_{hm}) or only the spatial pool (w/o P_s, 52.0% A_{hm}) reveal that the pools are not redundant. This confirms that each granularity captures unique and complementary spatio-temporal information, which is crucial for achieving a robust stability-plasticity balance.

Dissimilarity Attention Rectification

The Dissimilarity Attention Rectification (DAR) module is designed to explicitly model and balance non-essential, class-agnostic variations, allowing the task-specific prompts to focus on true discriminative features. To verify its contribution, we ablated this component (w/o P_{DAR}). The results in Table.2 show that this component provides a significant contribution. Removing the DAR module causes a 1.0% A_{hm} drop, from 57.7% (full model) down to 56.7%. This confirms its essential role in refining the feature space to enhance both stability and plasticity.

Method	Old (%)	New (%)	A_{HM} (%)
DPCA (Inc=2)	**61.1**	53.5	57.0
DPCA (Inc=3) (Ours)	60.4	**55.3**	**57.7**
DPCA (Inc=4)	59.8	54.1	56.8
DPCA (Inc=5)	59.5	53.2	56.2

Table 4: **Ablation on the Size of incremental prompts (Inc).** Inc=N denotes adding N prompts per new task. The Inc=3 setting achieves the best trade-off between representational capacity and overfitting, validating our approach.

Size of the Incremental Prompt Pool

To investigate the optimal capacity for learning new tasks, we conducted an ablation study on the number of prompts added per granularity for each task (Inc=N). The results in Table.3 shows a clear performance trade-off. Our default setting, Inc=3, achieves the optimal A_{HM} of 57.7%. A smaller pool (Inc=2) results in a lower A_{HM} (57.0%), suggesting insufficient representational capacity. Conversely, larger pools (Inc=4 and Inc=5) also degrade performance (56.8% and 56.2% A_{HM} respectively), indicating overfitting in the stringent 5-shot setting.

Conclusions

This paper challenges prevailing 'flat' prompt-based methods, arguing that their single-granularity design is structurally mismatched with the hierarchical nature of human actions in continual learning. To resolve this, we introduce DPCA, a framework that synergizes a multi-granularity task-specific prompter for unique action details with a task-agnostic prompter that learns a shared vocabulary of "action primitives" to mitigate catastrophic forgetting without data rehearsal. Rigorous experiments validate our approach, with DPCA establishing a new state-of-the-art. It achieves a superior stability-plasticity balance, evidenced by the highest harmonic mean accuracy (A_{HM}) that consistently outperforms prior methods across all incremental stages. By shifting from monolithic to compositional prompts, our work charts a new course for creating truly intelligent, private, and continuously adapting interactive systems, highlighting that future model design should prioritize alignment with the inherent, structural properties of the problem domain.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), the Joint Fund of the Ministry of Education of China (8091B022149, 8091B02072404), National Natural Science Foundation of China (62132016, 62571412, 62302372).

References

Bowman, B.; Achille, A.; Zancato, L.; Trager, M.; Perera, P.; Paolini, G.; and Soatto, S. 2023. a-la-carte prompt tuning (apt): Combining distinct data via composable prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14984–14993.

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.

Du, Y.; Fu, Y.; and Wang, L. 2015. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, 579–583. IEEE.

Du, Y.; Wang, W.; and Wang, L. 2015a. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Du, Y.; Wang, W.; and Wang, L. 2015b. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

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

Fang, T.; Zhang, Y.; Yang, Y.; Wang, C.; and Chen, L. 2023. Universal prompt tuning for graph neural networks. *Advances in Neural Information Processing Systems*, 36: 52464–52489.

Gao, R.; and Liu, W. 2023. Ddgr: Continual learning with deep diffusion-based generative replay. In *International Conference on Machine Learning*, 10744–10763. PMLR.

Garg, P.; Joseph, K.; Balasubramanian, V. N.; Camgoz, N. C.; Wan, C.; Kin, K.; Si, W.; Ma, S.; and De La Torre, F. 2024. Poet: Prompt offset tuning for continual human action adaptation. In *European Conference on Computer Vision*, 436–455. Springer.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 831–839.

Kang, M.; Park, J.; and Han, B. 2022. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16071–16080.

Kemker, R.; and Kanan, C. 2017. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2017. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE international conference on multimedia & expo workshops (ICMEW)*, 597–600. IEEE.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liao, S.; Lyons, T.; Yang, W.; Schlegel, K.; and Ni, H. 2021. Logsig-RNN: A novel network for robust and efficient skeleton-based action recognition. *arXiv preprint arXiv:2110.13008*.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Mittal, S.; Galesso, S.; and Brox, T. 2021. Essentials for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3513–3522.

Park, J.; Kang, M.; and Han, B. 2021. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13698–13707.

Razdaibiedina, A.; Mao, Y.; Hou, R.; Khabsa, M.; Lewis, M.; and Almahairi, A. 2023. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

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.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11909–11919.

Van de Ven, G. M.; Siegelmann, H. T.; and Tolias, A. S. 2020. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1): 4069.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022a. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, 631–648. Springer.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 139–149.

Wu, C.; Herranz, L.; Liu, X.; Van De Weijer, J.; Raducanu, B.; et al. 2018. Memory replay gans: Learning to generate new categories without forgetting. *Advances in neural information processing systems*, 31.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 374–382.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Yu, P. S. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xu, C.; Lyu, G.; Yan, J.; Yang, M.; and Deng, C. 2024a. LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures. In *ACL*, 5004–5013.

Xu, C.; Yan, J.; and Deng, C. 2025. Keep and Extent: Unified Knowledge Embedding for Few-shot Image Generation. *IEEE TIP*.

Xu, C.; Yan, J.; Yang, M.; and Deng, C. 2024b. Rethinking Noise Sampling in Class-Imbalanced Diffusion Models. *IEEE TIP*.

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

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

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.