SCIENCE CHINA
Information Sciences

• Supplementary File •

AS-CAR: Adaptive Topology Evolution with Semantic Alignment for Continual Action Recognition

Xingyu Zhu[1], Xiangbo Shu[1*], Binqian Xu[1], Liyan Zhang[2] & Jinhui Tang[3]

[1]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing* 210094*, China*
[2]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016*, China*
[3]*College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing* 210037*, China*

Appendix A Related Work

Appendix A.1 Video Class-Incremental Action Recognition

Early Video Class-Incremental Learning [7, 11] (VCIL) approaches primarily rely on replay and regularization mechanisms. TCD [11] encodes temporal information through sequential knowledge distillation, while vCLIMB [7] introduces a temporal-consistency loss to reduce stored frames and alleviate storage demands. FrameMaker [9] further compresses each video segment into key frames for efficient incremental updates. Else-Net [8] integrates human action semantics with skeletal structures, automatically identifying and updating modules for newly introduced actions. However, replay-based strategies incur additional storage overhead and raise privacy concerns. To address privacy and stability issues, STSP [3] introduces a spatiotemporal subspace projection strategy that reduces data leakage and catastrophic forgetting.

Recent advances have explored prompt learning [12] and multimodal pretraining [4, 5] to enhance temporal modeling. PIVOT [6] combines spatial prompts, memory replay, and a Transformer encoder to mitigate forgetting, whereas ST-Prompt [10] proposes a replay-free framework that leverages pretrained vision–language models and temporal prompts for temporal representation learning. Despite their effectiveness, these methods depend heavily on pretrained resources and are difficult to adapt to domains lacking such large-scale pretraining.

In contrast, Few-Shot Video Class-Incremental Learning (FSVCIL) remains underexplored. POET [2] extends the L2P [12] paradigm to this setting by constructing a spatiotemporal prompt pool, demonstrating the feasibility of prompt-based learning in non-Transformer architectures such as GCNs [16, 30], where "prompts" are reformulated as "offsets." However, directly transferring image-based methods to the video domain neglects the intrinsic motion regularities of actions. In particular, for skeletal action recognition, stage-specific motion topologies are crucial for maintaining consistent understanding of joint-level contextual relations.

To this end, we propose the Adaptive Topology Evolution (ATE) module, which enables GCN-based models to dynamically learn and preserve stage-specific skeletal graph structures. Unlike prompt-driven strategies, ATE explicitly models evolving motion topologies across sessions, mitigating the forgetting of structural patterns from previous tasks while remaining lightweight and compatible with most existing GCN frameworks.

Appendix A.2 Leveraging Pretrained Language Models in Visual Learning

Recent advances in pretrained models [37] have significantly transformed computer vision [15, 17, 34], offering rich semantic representations that enable effective cross-modal learning. In zero-shot and few-shot settings, pretrained language models excel through semantic embeddings and cross-modal feature alignment [32], inspiring the Class-Incremental Learning [33, 35, 36] (CIL) community to integrate vision–language optimization strategies, especially for improving base stage representations.

Building upon the structural adaptability introduced by our ATE module, which captures stage-specific motion topologies for robust representation learning, we further investigate how external semantic priors can complement such structural modeling in few-shot continual scenarios. Specifically, Pretrained Language Models [17] (PLMs) provide global category-level semantic knowledge that can be transferred into the visual domain, guiding the formation of more discriminative and semantically consistent class prototypes.

Representative methods such as LGCL [14] employ a language-guided prompt pool to project visual features into the linguistic domain and reduce modality gaps; LRT [4] introduces context-aware prompt learning to transfer textual knowledge into the visual modality; and PriViLege [5] applies semantic knowledge distillation using pretrained language priors for novel class learning. These approaches commonly rely on prompt learning or knowledge distillation modules to bridge modalities.

* Corresponding author (email: shuxb@njust.edu.cn)

In contrast, our method eliminates explicit prompt learning. Instead, we directly construct an anchor graph based on the intrinsic semantic topology encoded by a Pretrained Language Model [17], and transfer this class-relationship structure into the visual feature space through an optimal transport objective. This design effectively injects semantic topology into visual learning, refining decision boundaries without introducing additional training overhead and enabling a seamless fusion of semantic and structural information in Few-shot Class-Incremental Action Recognition.

Appendix A.3 Comparison with Existing Methods

Our work differs substantially from existing methods in both design motivation and technical implementation. Previous prompt-based continual learning methods, such as L2P [12], CODA-Prompt [13] and ASP [1], were originally developed for image domains and rely heavily on large pretrained visual models. L2P, designed for general continual learning, depends on fixed prompt pools, which often result in limited adaptability and weak generalization in few-shot scenarios. ASP extends this idea to few-shot continual learning by generating instance-level prompts, yet it suffers from instability across incremental stages. OrCo [28] improves feature alignment but remains constrained by reliance on image-based pretraining. In the action recognition field, POET [2] is an early attempt to apply prompt-based continual learning, but it largely inherits L2P's pool-based paradigm and thus faces similar adaptability issues. In contrast, our method is specifically developed for few-shot continual action recognition without using any large pretrained models. It introduces Adaptive Topology Evolution, which dynamically learns task-specific skeletal structures across sessions, and Semantic Structure Transfer, which transfers relational semantics from language models to guide prototype optimization. Together, these components enable our framework to jointly model temporal topology evolution and semantic consistency, achieving stronger adaptability and continual learning stability than existing prompt-based methods.

Appendix B Method

Appendix B.1 Progressive Momentum Adaptation

To ensure stable adaptation under few-shot incremental learning, we propose a Progressive Momentum Adaptation (PMA) strategy for the visual encoder $\mathcal{F}_\theta(\cdot)$, which is composed of L hierarchical layers $\{\mathcal{F}_\theta^1(\cdot), \mathcal{F}_\theta^2(\cdot), \ldots, \mathcal{F}_\theta^L(\cdot)\}$. After completing the base session training, the shallow layers $\{\mathcal{F}_\theta^1(\cdot), \ldots, \mathcal{F}_\theta^m(\cdot)\}$ are frozen to preserve the general motion semantics captured during base learning, as these layers primarily encode task-agnostic representations. During subsequent incremental sessions, only the deeper layers $\left\{\mathcal{F}_\theta^{m+1}(\cdot), \ldots, \mathcal{F}_\theta^L(\cdot)\right\}$ are updated to enable efficient adaptation to new action classes while preventing interference with previously learned knowledge. However, due to the limited number of samples per class in each incremental session, direct fine-tuning of the unfrozen layers can easily lead to overfitting. To mitigate this, PMA introduces an exponential moving average (EMA) update mechanism to gradually refine the trainable parameters, formulated as:

$$\theta_t = \beta\,\theta_{t-1} + (1-\beta)\,\hat{\theta}_t, \tag{B1}$$

where $\hat{\theta}_t$ represents the parameters that have been updated through gradient optimization for the current task t, and $\beta \in [0,1]$ is the momentum coefficient controlling temporal smoothness. This selective and progressive update stabilizes parameter evolution across sessions, effectively balancing model plasticity and stability. By freezing general layers and applying momentum-driven adaptation to deeper layers, PMA ensures efficient knowledge integration while maintaining robust generalization in FSCAR.

Appendix B.2 Adaptive Topology Evolution

In a continual skeleton-based action recognition framework, the evolution of graph topology in each session is essential for maintaining the continuity of human motion representations. To address the heterogeneous nature of skeletal data at incremental stage t, we introduce an Adaptive Topology Evolution (ATE) module that enables the model to capture session-specific inductive biases embedded in the skeletal topology, thereby reinforcing persistent understanding of human motion patterns. For each incremental session t, a learnable modulation matrix U^t, having the same shape as the base adjacency matrix A, is learned to represent the session-specific skeletal graph. The modulation is then integrated into the original graph convolution propagation as follows:

$$H^{(l+1)} = \sum_{s=1}^{S} \sigma\big(\big(A_s^{(l+1)} + \mu\, U_s^{t,(l+1)}\big) H^{(l)} M_s^{(l+1)}\big), \tag{B2}$$

where $U_s^{t,(l+1)}$ denotes the topology modulation matrix for channel s at layer $l{+}1$ in session t, and $\mu = 0.4$ controls the strength of topology adaptation. Throughout the continual learning process, the composite topologies $A + \mu\, U^t$ from all previous sessions $1, \ldots, t{-}1$ are kept frozen, while only the current session's U^t is updated. This mechanism allows the network to incrementally and adaptively evolve the skeletal graph structure across sessions.

Appendix B.3 Semantic Structure Transfer

In FSCAR, unclear boundaries between base and incremental classes often lead to overfitting. To address this, we propose a Semantic Structure Transfer (SST) strategy that introduces semantic relations from a pretrained language model into the visual feature space. For each action class, we prompt a large language model [21] to generate detailed motion descriptions

$\mathcal{A} = \{\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2, \ldots, \widetilde{\mathcal{C}}_C\}$, which are encoded by a text encoder $\mathcal{T}\varphi(\cdot)$ to obtain textual prototypes $P^{\text{text}} = \mathcal{T}\varphi(\mathcal{A}) \in \mathbb{R}^{C \times d_t}$. Corresponding visual prototypes $P^{\text{vis}} \in \mathbb{R}^{C \times d_v}$ are derived from the visual backbone.

We then align the pairwise relational graphs of P^{text} and P^{vis} using the entropy-regularized Gromov–Wasserstein (EGW) distance [31] to ensure structural consistency between the semantic and visual domains. Intuitively, EGW measures how the internal pairwise relations among class prototypes in the text space correspond to those in the visual space. Unlike conventional vector-based alignment that only matches features one by one, EGW compares the relational geometry of the two graphs, preserving their topological correspondence [29]. Specifically, let D_1 and D_2 denote the pairwise distance matrices of text and visual prototypes, respectively. The EGW loss is formulated as:

$$\begin{aligned} \mathcal{L}_{GW} &= \min_{T \in \Pi(p,q)} \sum_{i,j,k,l} \big(D_1[i,k] - D_2[j,l]\big)^2 \, T_{i,j} \, T_{k,l} \; - \; \epsilon \, H(T) \\ &= \min_{T \in \Pi(p,q)} \sum_{i,j} T_{i,j} \, M_{i,j} \; - \; \epsilon \, H(T), \end{aligned} \tag{B3}$$

where $\Pi(p,q)$ is the set of couplings with marginals $p = q = \frac{1}{C}\mathbf{1}C$, $H(T) = -\sum i, jT_{i,j} \log T_{i,j}$ denotes the entropy term, $M = (D_1 \odot D_1)\, p\mathbf{1}^\top + \mathbf{1}\big((D_2 \odot D_2)\, q\big)^\top - 2\, D_1 \, T \, D_2^\top$, and ϵ controls regularization strength. This formulation enables differentiable alignment between textual and visual structures, effectively transferring semantic topology across modalities.

Appendix B.4 Training Target

Our training procedure follows the FSCIL paradigm. In the base session, the visual prototype parameters are randomly initialized and optimized as classifier weights by jointly minimizing the following objectives:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{CE} + \alpha \, \mathcal{L}_{GW}. \tag{B4}$$

We conduct sufficient training on the base dataset to obtain a generalized and semantically aligned feature space, during which the visual encoder $\mathcal{F}_\theta(\cdot)$ is fully updated. In the subsequent incremental sessions, due to the scarcity of new samples, only the $(m+1)_{th} - L_{th}$ block of the visual encoder is updated while the first m layers are frozen. A momentum-based update is further applied following Eq.B1 to mitigate forgetting and overfitting. The training objective remains consistent with that of the base session. At the inference stage, we update the text prototypes, discard the text encoder, and perform action classification by computing the similarity between the input skeletal features and the learned class prototypes.

Appendix C Experiments

Appendix C.1 Datasets

Following POET [2], we evaluate our method on three standard benchmarks for skeletal action recognition: NTU RGB+D 60 (NTU-60), NTU RGB+D 120 (NTU-120), and PKU-MMD I.

NTU RGB+D 60 [18] (NTU-60) is a large-scale human action recognition dataset containing 56,880 skeleton sequences captured by depth sensors. Each action is performed by one or two of 40 distinct subjects and annotated with 25 body joints, resulting in 60 daily activity classes. The dataset provides two official evaluation protocols: (1) Cross-Subject (X-Sub), where 20 subjects are used for training and the remaining 20 for testing; and (2) Cross-View (X-View), where training samples are collected from camera views 2 and 3, and testing samples from view 1. In our few-shot class-incremental setting, we follow the X-Sub protocol and select 40 base classes and 20 incremental classes, introducing five new classes per session in a 5-way, 5-shot configuration.

NTU RGB+D 120 [19] (NTU-120) extends NTU-60 by incorporating 114,480 skeleton sequences spanning 120 action categories. Actions are performed by 106 subjects and captured from 32 different camera setups, making it significantly more diverse. Two official benchmarks are defined: (1) Cross-Subject (X-Sub), with 53 subjects for training and 53 for testing; and (2) Cross-Setup (X-Set), where training data come from 16 even setup IDs and testing data from 16 odd ones. For consistency, we adopt the X-Sub protocol and partition the dataset into 100 base and 20 incremental classes, using the same 5-way, 5-shot incremental protocol as NTU-60.

PKU-MMD I [20] is designed for multi-view skeleton-based action recognition and contains 6,952 sequence covering 51 action categories. The dataset is collected from various viewpoints, making it well-suited for testing generalization to cross-view variations. Following the standard cross-subject split (5,339 training and 1,613 test samples), we divide the dataset into 31 base and 20 incremental classes, also under the 5-way, 5-shot configuration.

Together, these datasets present complementary challenges, as NTU-60 and NTU-120 emphasize large-scale subject and viewpoint diversity, whereas PKU-MMD I focuses on viewpoint robustness, enabling a comprehensive evaluation of our model's continual learning capability across diverse skeletal domains.

Appendix C.2 Evaluation Metrics

To assess both adaptation and forgetting, we report three metrics:

- Average accuracy $A_{\text{avg}} = \frac{1}{T+1} \sum_{t=0}^{T} A_t$, where A_t is the Top-1 accuracy after session t.
- Performance degradation $\text{PD} = A_0 - A_t$.
- Forgetting measure $\text{FM} = \frac{1}{t-1} \sum_{k=1}^{t-1} \max_{k' \in \{1,\ldots,t-1\}} \{\, s_{k',k} - s_{t,k}\}$, where $s_{t,k}$ denotes accuracy on task k after session t.

Table C1 Default hyperparameters for AS-CAR

Configuration	Hyperparameter
α	0.35
β	0.9
μ	0.4
m	4
random rotation	True
window size	64
base lr	0.1
incremental lr	0.01
lr scheduler	cosine decay
base batch size	64
incremental batch size	25
base epoch	100
incremental epoch	10
optimizer	SGD

Appendix C.3 Implementation Details

Pretrained Language Model. To minimize resource consumption, we employ an offline language model to extract semantic topology knowledge. Specifically, we prompt GPT-4 [21] with "Describe the skeletal movement characteristics of action [CLASS] in detail." to obtain fine-grained action descriptions. These textual descriptions are then encoded using the CLIP ViT/B-16 [17] text encoder to derive text prototypes serving as prior semantic knowledge. The visual encoder employs the CTR-GCN [16] architecture and text encoder remains frozen throughout this process.

Base Training. During the base session stage, we train the model using the SGD optimizer with a batch size of 64 and an initial learning rate of 0.1. The learning rate decays step-wise at epochs 35 and 75 over a total of 100 epochs to achieve generalized feature representations.

Incremental Training. For each incremental session, training proceeds for 10 epochs with an initial learning rate of 0.01 and a batch size of 25. To mitigate forgetting, the first m layers of the backbone encoder are frozen, while the subsequent $(m+1)_{th} - L_{th}$ layers are updated with momentum. This strategy balances model plasticity and stability, reducing the risk of overfitting.

All experiments are conducted on a single NVIDIA RTX 4090 GPU using PyTorch. Each experiment is repeated with three random seeds $(7, 1024, 2025)$, and the reported results are averaged across these runs. In Table C1, we provide the default hyperparameter settings used for training our model.

Appendix C.4 Performance Comparison

Results on NTU-60. As shown in Table C2, our method achieves the highest overall performance among all baselines on the NTU-60 Cross-Subject benchmark. Traditional class-incremental learning methods such as LWF and LUCIR suffer from severe forgetting, with accuracy dropping to nearly zero on old tasks. More advanced prompt-based approaches like L2P [12], DualPrompt [25], and CODA-P [13] alleviate forgetting to some extent but still exhibit limited adaptability to dynamic skeletal structures.

Few-shot methods (CEC, FACT, and TEEN) demonstrate improved stability by decoupling representation learning and classification, yet they remain constrained by their shallow temporal modeling. POET and ASP further enhance performance through spatiotemporal prompting and adaptive subspace projection, respectively. However, both still face challenges in maintaining long-term knowledge retention.

In contrast, AS-CAR achieves an average accuracy of 78.9%, outperforming the strongest baseline (ASP) by 0.5%, while also attaining the lowest performance drop (17.7%) and forgetting measure (3.5). These results highlight AS-CAR's superior ability to preserve base-class knowledge and adapt to new classes effectively, owing to its combination of structure-aware graph modeling and semantic-guided transfer.

Results on NTU-120. As illustrated in Table C3, the NTU-120 dataset presents a more challenging few-shot incremental learning scenario due to its larger class diversity and higher motion complexity. Nevertheless, our method maintains consistently strong performance across all sessions. Compared with the strongest baseline ASP [1], which achieves an average accuracy of 78.9%, our approach further improves this to 79.3%, demonstrating superior adaptability to large-scale incremental learning.

In particular, our method achieves stable accuracy gains across all sessions, reaching 81.2%, 78.4%, 77.5%, and 76.2% from S1 to S4. This steady trend contrasts with the notable performance degradation observed in most baselines, especially those dependent on static prompting (e.g., POET and DualPrompt). Moreover, the performance dropping rate is reduced to 7.2%, the lowest among all compared methods, indicating robust knowledge retention and minimal forgetting.

These results confirm that the combination of adaptive topology learning and semantic structure transfer allows our framework to generalize effectively across sessions, preserving motion semantics and sustaining discriminative power even under the more complex NTU-120 protocol.

Table C2 Top-1 accuracy A_t (%) in each incremental task, average accuracy A_{avg} and performance dropping rate (PD) on NTU-60 Cross-Subject dataset.

Methods	S0	S1	S2	S3	S4	$A_{\text{avg}}(\uparrow)$	PD (↓)	FM (↓)
	A_0	A_1	A_2	A_3	A_4			
Standard Class-Incremental Learning								
LWF [23]	88.4	6.2	2.8	3.7	3.2	20.9	85.2	-
LUCIR [24]	87.9	4.3	4.1	2.5	2.4	20.3	85.6	-
L2P [12]	88.6	78.9	71.0	64.2	56.8	71.9	31.8	-
DualP [25]	88.2	76.2	71.3	65.1	59.2	72.3	29.0	-
CODA-P [13]	87.4	76.1	66.7	58.6	51.8	68.1	35.6	-
Few-Shot Class-Incremental Learning								
CEC [22]	87.2	80.3	72.4	66.8	61.8	73.7	25.4	-
FACT [27]	87.0	79.5	71.8	63.5	58.6	72.1	28.4	-
TEEN [26]	88.2	80.9	75.6	67.1	63.6	75.1	24.7	-
POET [2]	87.9	82.3	76.8	68.4	57.1	74.5	30.8	29.9
ASP [1]	88.6	82.7	**77.4**	73.2	70.2	78.4	18.4	3.6
Ours	**88.9**	**83.4**	77.0	**74.0**	**71.2**	**78.9**	**17.7**	**3.5**

Table C3 Top-1 accuracy A_t (%) in each incremental task, average accuracy A_{avg} and performance dropping rate (PD) on NTU-120 Cross-Subject dataset.

Methods	S0	S1	S2	S3	S4	$A_{\text{avg}}(\uparrow)$	PD (↓)	FM (↓)
	A_0	A_1	A_2	A_3	A_4			
Standard Class-Incremental Learning								
LWF [23]	83.1	13.4	12.7	14.9	4.2	25.7	78.9	-
LUCIR [24]	82.8	12.6	13.2	11.8	5.3	25.1	77.5	-
L2P [12]	82.2	73.5	68.0	65.2	60.4	69.9	21.8	-
DualP [25]	82.4	74.2	69.2	66.5	62.4	70.9	20.0	-
CODA-P [13]	82.4	71.7	64.6	61.3	55.2	67.0	27.2	-
Few-Shot Class-Incremental Learning								
CEC [22]	82.4	76.8	71.9	67.9	66.2	72.9	16.2	-
FACT [27]	82.1	76.5	70.4	64.2	63.7	71.4	18.4	-
TEEN [26]	83.2	77.2	74.3	68.8	66.8	74.1	16.4	-
POET [2]	82.8	79.5	77.1	71.5	62.9	74.8	19.9	-
ASP [1]	**83.9**	80.4	77.7	76.8	75.8	78.9	8.1	1.5
Ours	83.4	**81.2**	**78.4**	**77.5**	**76.2**	**79.3**	**7.2**	1.5

Table C4 Top-1 accuracy A_t (%) in each incremental task, average accuracy A_{avg} and performance dropping rate (PD) on PKU-MMD I Cross-Subject dataset.

Methods	S0	S1	S2	S3	S4	$A_{\text{avg}}(\uparrow)$	PD (↓)	FM (↓)
	A_0	A_1	A_2	A_3	A_4			
Standard Class-Incremental Learning								
LWF [23]	94.6	15.6	16.3	13.4	8.3	29.6	86.3	-
LUCIR [24]	94.2	17.4	13.1	11.6	11.3	29.5	82.9	-
L2P [12]	94.2	82.1	73.6	65.3	58.3	74.7	35.9	-
DualP [25]	94.6	83.6	72.9	66.7	58.6	75.3	36.0	-
CODA-P [13]	94.6	80.7	70.3	61.4	55.9	72.6	38.6	-
Few-Shot Class-Incremental Learning								
CEC [22]	94.5	86.8	76.9	70.1	66.5	79.0	27.9	-
FACT [27]	94.2	84.9	75.4	67.5	64.9	77.4	29.3	-
TEEN [26]	95.2	87.5	78.7	71.8	68.3	80.3	26.9	-
POET [2]	94.8	89.3	81.5	75.6	67.8	81.8	27.0	-
ASP [1]	**96.1**	**90.5**	80.0	74.4	71.7	82.5	24.4	**3.6**
Ours	96.0	90.4	**81.9**	**75.6**	**72.1**	**83.2**	**23.9**	3.7

Table C5 Comparison of model complexity across incremental sessions for different methods.

Methods	Params (M)	FLOPs (G)	IT (ms)
POET [2]	1.63	1.88	1.80
ASP [1]	29.42	3.60	2.62
Ours	1.41	1.79	1.38

Table C6 Memory cost of ATE across 5 Tasks.

	Memory Cost (MB)
U^t	0.5

Results on PKU-MMD I. Table C4 presents the results on the PKU-MMD I dataset, which poses additional challenges due to its diverse recording viewpoints and larger inter-subject variations. Our method consistently outperforms all baselines across every incremental session. Specifically, it achieves 96.0% accuracy in the base session (S0) and maintains strong performance through the final session (S4) with 72.1%, leading to the highest overall average accuracy of 83.2%. This improvement over ASP demonstrates the superior adaptability of our framework under complex, cross-domain skeletal scenarios.

Conventional CIL methods such as LWF and LUCIR experience severe performance degradation, with accuracy falling below 20% after the initial few sessions, whereas our model remains stable and effectively mitigates catastrophic forgetting. Compared to few-shot methods such as POET and TEEN, AS-CAR exhibits smoother accuracy transitions between sessions, showing its capability to balance old-class retention and new-class acquisition. Moreover, the lowest performance dropping rate (23.9%) further indicates that our method effectively stabilizes feature representations during continual updates.

Overall, these results highlight that integrating semantic structure alignment with adaptive topology modeling allows AS-CAR to robustly generalize across varied motion domains while maintaining incremental learning stability.

Appendix C.5 Efficiency Comparison

Table C5 compares model complexity among representative methods in terms of parameter size (Params), floating point operations (FLOPs), and inference time (IT) on a single RTX 4090 GPU. POET and ASP both incur higher computational costs due to their additional prompt mechanisms. Specifically, POET requires maintaining a learnable prompt pool, while ASP integrates prompts with multiple auxiliary modules, leading to larger model sizes and slower inference.

In contrast, our method achieves comparable or better performance with significantly fewer parameters and lower FLOPs. The lightweight architecture ensures faster inference and reduced memory overhead, making it more scalable for long-term continual learning scenarios. These results demonstrate that our design effectively balances efficiency and performance, offering a more practical solution for few-shot continual action recognition.

Table C6 reports the memory cost of the ATE module across 5 tasks. The results show that the additional memory required to store the learned topology matrices U^t is about 0.5 MB, which is negligible compared to the overall model size. This indicates that ATE introduces minimal storage overhead, even when new topology matrices are generated at each incremental stage.

Such lightweight design demonstrates that ATE is highly scalable for long-term continual learning. By representing task-specific topologies in a compact form and avoiding redundant parameterization, ATE effectively maintains structural adaptability without imposing significant memory or computational burdens.

Appendix D Ablation Studies

Appendix D.1 Effect of Different Components

Table D1 summarizes the ablation study of AS-CAR by selectively removing each major component. Both the ATE and the SST contribute notably to the final performance. When ATE is removed, the average accuracy drops from 78.9% $\rightarrow$ 75.5%, 79.3% $\rightarrow$ 74.2%, and 83.2% $\rightarrow$ 80.9% on NTU-60, NTU-120, and PKU-MMD I, respectively. This decline shows that ATE plays a key role in refining class prototypes through dynamic relational encoding.

Similarly, removing SST results in an even larger performance decrease across all datasets (e.g., –2.9% on PKU-MMD I), suggesting that structural guidance is crucial for maintaining discriminative representation alignment during class expansion. Notably, both modules exhibit consistent benefits across different scales and domains, indicating their complementary roles: ATE stabilizes topology adaptivity, while SST enhances semantic boundary.

Overall, the ablation results demonstrate that the superior performance of AS-CAR arises from the combined effect of ATE and SST, since each component alone is insufficient to fully capture the cross-task structural consistency required for few-shot class-incremental skeleton recognition.

Appendix D.2 Necessity of EGW

To validate the necessity of entropy-regularized Gromov–Wasserstein (EGW) alignment in the SST module, we replaced it with the graph module from LRT [4] and conducted experiments on NTU-60. As shown in Table D2, the average accuracy dropped from 78.9% to 75.3% after removing EGW, demonstrating its essential role in maintaining semantic consistency. Unlike vector-based similarity metrics that only consider node-level correspondence, EGW alignment captures

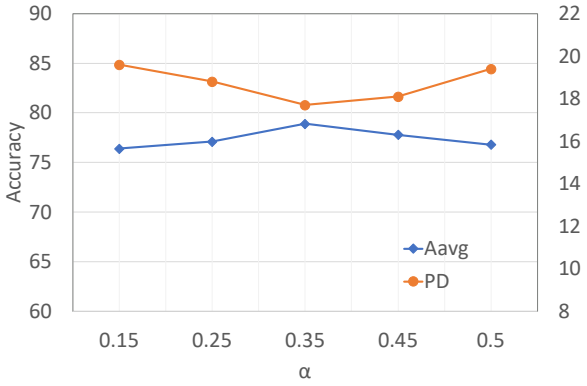

Figure D1 Ablation Results of Different α on NTU-60.

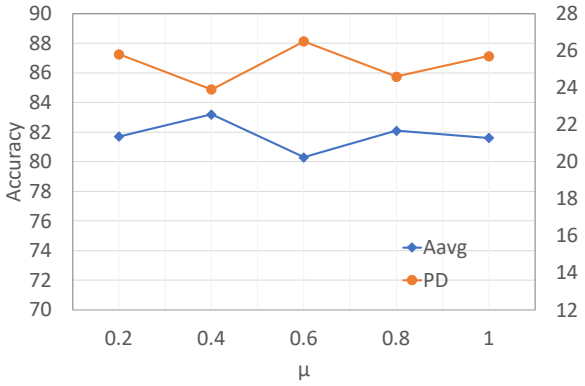

Figure D2 Ablation Results of Different μ on PKU-MMD I.

the relational dependencies between intra-graph edges, allowing a more precise transfer of semantic topology from textual to visual domains. This structural preservation ensures that the semantic relationships learned in the language space are faithfully reflected in the visual representation space, which is crucial for stable and discriminative adaptation in FSCAR.

Table D1 Ablation study of removing each component from AS-CAR respectively.

Schemes	A_{avg}		
	NTU-60	NTU-120	PKU-MMD I
Ours w/o ATE	75.5	74.2	80.9
Ours w/o SST	76.9	77.4	80.3
Ours	78.9	79.3	83.2

Table D2 Validation of GW alignment effectiveness in SST on NTU-60.

Stragtey	A_{avg}
Graph [4]	75.3
EGW	78.9

Appendix D.3 Parameters Sensitivity

As shown in Figure D1, the balance coefficient α in Eq. B4 determines the relative contribution of the $\mathcal{L}_{GW}$, which transfers semantic topology from the language space to the visual domain. When α is too small, the SST constraint is weak and the alignment between semantic and visual prototypes becomes insufficient, limiting the structural consistency learned by the ATE module.

With larger α, the transferred topology enhances representation coherence and improves the generalization to new classes. The best performance is achieved at $\alpha = 0.35$, where the average accuracy reaches its peak and the performance drop is minimized. Further increasing α reduces feature adaptability, indicating that a moderate balance between ATE and SST yields the most stable and discriminative representations.

Figure D2 illustrates the impact of parameter μ on ATE. It is observed that as μ increases from 0.2 to 0.4, both A_{avg} and PD improve, indicating that a moderate integration of task-specific topology enhances the adaptability of ATE. This balance allows the model to capture task-dependent graph variations while maintaining stable structural representations, leading to more consistent motion understanding across sessions.

However, when μ continues to grow beyond 0.4, the performance gradually decreases. This suggests that excessive topological injection disturbs the shared structural information learned from previous tasks, resulting in suboptimal generalization. Therefore, setting $\mu = 0.4$ achieves the best trade-off between adaptability and stability, ensuring that ATE effectively models evolving skeletal structures without disrupting previously learned representations.

Appendix D.4 Incremental Learning with Fewer Shots

As shown in Figure D3, model accuracy consistently decreases as the number of shot decreases, reflecting the inherent challenge of few-shot incremental learning. When the sample size is extremely limited, the model struggles to form stable prototypes, leading to weaker feature generalization and higher sensitivity to inter-class variations.

Despite this trend, the performance drop remains relatively smooth across sessions, demonstrating the robustness of the proposed framework. The combination of ATE and SST helps stabilize structural learning and semantic transfer, enabling the model to maintain reasonable accuracy even under 1-shot conditions. This indicates that our method effectively alleviates overfitting and knowledge forgetting in data-scarce scenarios.

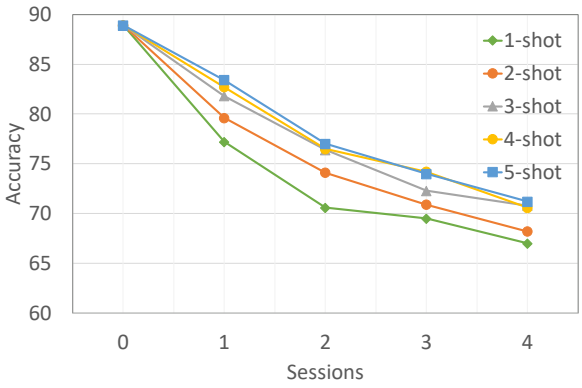

Figure D3 Accuracy with different number of shots k during incremental sessions on NTU-60.

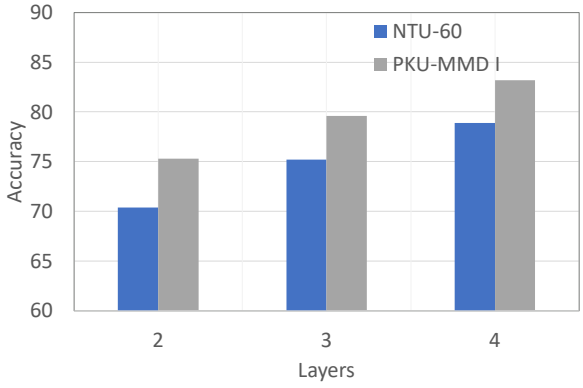

Figure D4 Impact of Different Layer Freezing Positions m on Performance.

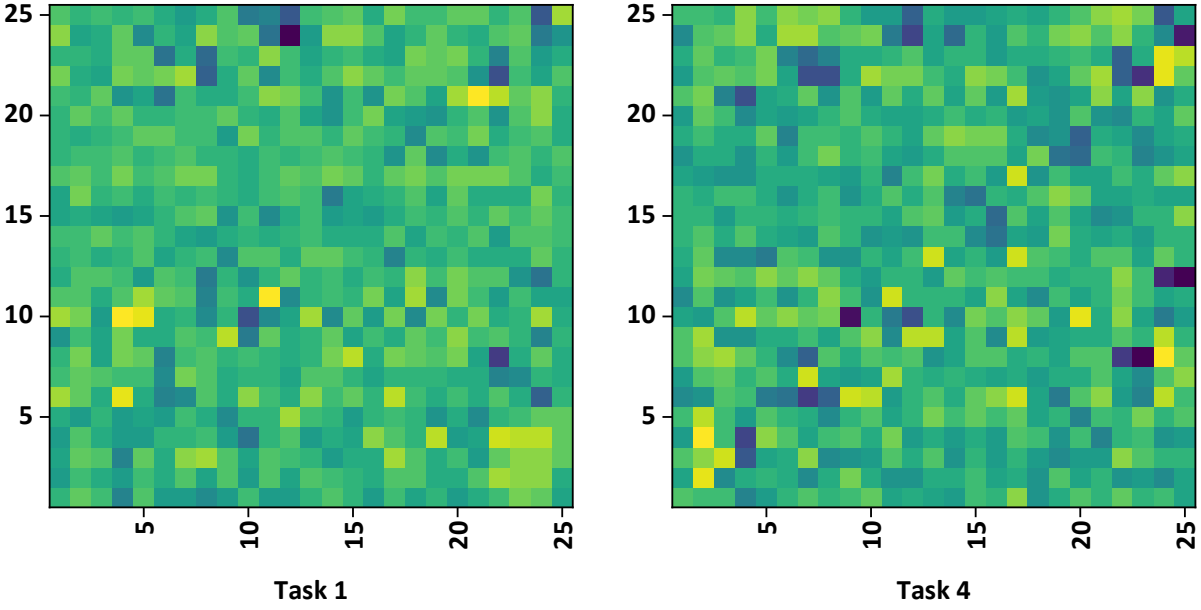

Figure D5 Visualization of Adjacent Matrices U^t in the ATE Module on NTU-120.

Appendix D.5 Analyze the Effect of Layer Freezing at Different Positions

Figure D4 shows that the A_{avg} increases as deeper layers are unfrozen, reaching the best performance when the first 4 layers are fixed. This indicates that the shallow layers indeed capture general and transferable motion semantics, which should remain stable during incremental learning, while the deeper layers are more suitable for adapting to new tasks.

When fewer layers are frozen, excessive parameter updates disturb previously learned representations, leading to knowledge forgetting. Conversely, freezing too many layers limits model plasticity and reduces adaptation to new actions. The optimal configuration at layer 4 effectively balances stability and adaptability, confirming the effectiveness of our PMA strategy in preserving base knowledge while maintaining strong incremental learning capability.

Appendix E Qualitative Analysis

Appendix E.1 Visualization of ATE

The visualization (Figure D5) illustrates the topology matrices learned by the ATE for Task 1 and Task 4. Each matrix corresponds to the spatiotemporal correlations among 25 skeletal joints, revealing how ATE dynamically adapts the graph structure to capture distinct motion paradigms as new tasks emerge. Compared with Task 1, the matrix of Task 4 shows more localized and differentiated connectivity patterns, reflecting the evolution of inter-joint relationships when encountering new action categories.

These results confirm that ATE effectively models stage-specific topological dependencies while preserving structural consistency across sessions. By progressively refining joint interactions instead of reusing a fixed topology, ATE enhances the model's capacity to encode evolving motion semantics, thereby mitigating catastrophic forgetting and improving adaptability

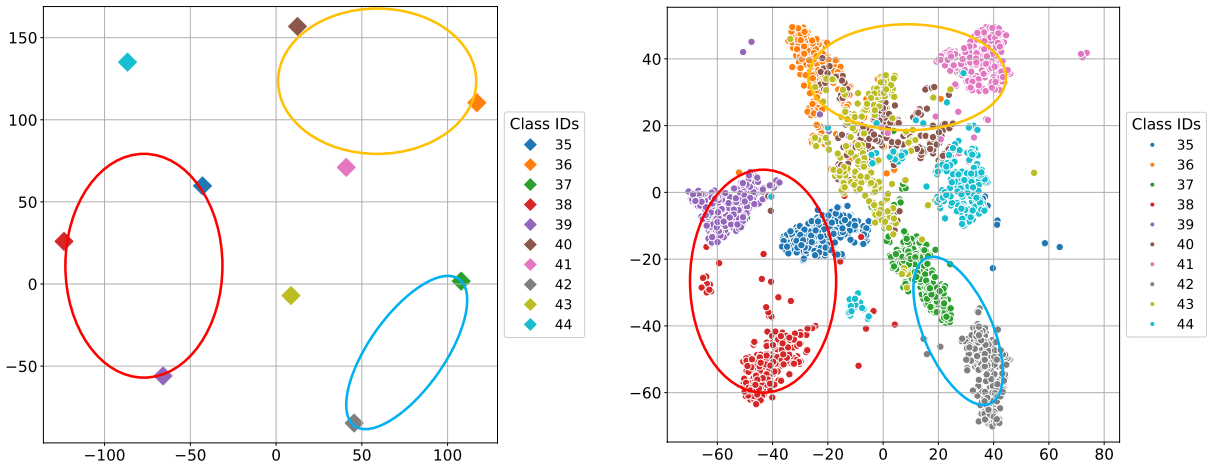

Figure E1 t-SNE Visualization of Transferring Textual Relations from PLMs (Left) to Visual Space (Right) by SST on NTU-60.

in long-term continual skeleton action recognition.

Appendix E.2 Visualization of SST

Figure E1 illustrates the t-SNE visualization of SST transferring class relationships derived from PLMs (left) to the visual space (right). After transfer, visually encoded prototypes become more clustered and preserve the relative semantic affinities that existed in the text space, as evidenced by class IDs 35, 38, and 39 remaining close to one another in the visual t-SNE. This indicates that SST successfully projects language-model topology into the visual domain rather than merely changing local distances arbitrarily.

By injecting semantic topology into the visual prototypes, SST increases intra-class compactness and maintains meaningful inter-class relations, which improves discrimination under few-shot conditions. These visualizations confirm that the semantic anchor graph guides the formation of a semantically consistent visual feature space, supporting more robust decision boundaries during incremental learning.

Appendix F Discussion on Applicability

The proposed framework demonstrates broad applicability to various continual learning scenarios beyond skeleton-based action recognition. Its modular design, combining transferable semantic structures with adaptive topology modeling, enables effective knowledge retention and fast adaptation across heterogeneous tasks. Since the approach relies on general feature representations and task-agnostic prototype alignment, it can be readily extended to other modalities such as gesture, human-object interaction, or even video event understanding. Moreover, its efficiency and low memory footprint make it suitable for deployment in real-world applications that require online adaptation, such as surveillance, healthcare monitoring, and human-computer interaction.

References

1 Liu C, Wang Z, Xiong T, et al. Few-shot class incremental learning with attention-aware self-adaptive prompt. In: Proceedings of European Conference on Computer Vision (ECCV), 2024. 1–18
2 Garg P, Joseph KJ, Balasubramanian V N, et al. POET: Prompt Offset Tuning for Continual Human Action Adaptation. In: Proceedings of European Conference on Computer Vision (ECCV), 2024. 436–455
3 Cheng H, Yang S, Wang C, et al. Stsp: Spatial-temporal subspace projection for video class-incremental learning. In: Proceedings of European Conference on Computer Vision (ECCV), 2024. 374–391
4 Zhao Y, Li J, Song Z, Tian Y. Language-Inspired Relation Transfer for Few-Shot Class-Incremental Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024
5 Park K-H, Song K, Park G-M. Pre-trained vision and language transformers are few-shot incremental learners. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 23881–23890
6 Villa A, Alcázar J L, Alfarra M, et al. Pivot: Prompting for video continual learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 24214–24223
7 Villa A, Alhamoud K, Escorcia V, et al. vclimb: A novel video class incremental learning benchmark. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 19035–19044
8 Li T, Ke Q, Rahmani H, et al. Else-net: Elastic semantic network for continual action recognition from skeleton data. In: Proceedings of International Conference on Computer Vision (ICCV), 2021. 13434–13443
9 Pei Y, Qing Z, Cen J, et al. Learning a condensed frame for memory-efficient video class-incremental learning. Advances in Neural Information Processing Systems (NeurIPS), 2022, 35: 31002–31016
10 Pei Y, Qing Z, Zhang S, et al. Space-time prompting for video class-incremental learning. In: Proceedings of International Conference on Computer Vision (ICCV), 2023. 11932–11942
11 Park J, Kang M, Han B. Class-incremental learning for action recognition in videos. In: Proceedings of International Conference on Computer Vision (ICCV), 2021. 13698–13707
12 Wang Z, Zhang Z, Lee C-Y, et al. Learning to prompt for continual learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 139–149
13 Smith J S, Karlinsky L, Gutta V, et al. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 11909–11919

14 Khan M G Z A, Naeem M F, Van Gool L, et al. Introducing language guidance in prompt-based continual learning. In: Proceedings of International Conference on Computer Vision (ICCV), 2023. 11463–11473
15 Zhu X, Shu X, Tang J. Motion-Aware Mask Feature Reconstruction for Skeleton-Based Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10718–10731
16 Chen Y, Zhang Z, Yuan C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of International Conference on Computer Vision (ICCV), 2021. 13359–13368
17 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning (ICML), 2021. 8748–8763
18 Shahroudy A, Liu J, Ng T-T, Wang G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 1010–1019
19 Liu J, Shahroudy A, Perez M, et al. Ntu RGB+D 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2684–2701
20 Liu C, Hu Y, Li Y, Song S, Liu J. PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding. arXiv preprint arXiv:1703.07475, 2017
21 Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023
22 Zhang C, Song N, Lin G, et al. Few-shot incremental learning with continually evolved classifiers. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 12455–12464
23 Li Z, Hoiem D. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935–2947
24 Hou S, Pan X, Loy C C, Wang Z, Lin D. Learning a unified classifier incrementally via rebalancing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 831–839
25 Wang Z, Zhang Z, Ebrahimi S, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In: Proceedings of European Conference on Computer Vision (ECCV), 2022. 631–648
26 Wang Q-W, Zhou D-W, Zhang Y-K, Zhan D-C, Ye H-J. Few-shot class-incremental learning via training-free prototype calibration. Advances in Neural Information Processing Systems (NeurIPS), 2023, 36: 15060–15076
27 Zhou D-W, Wang F-Y, Ye H-J, Ma L, Pu S, Zhan D-C. Forward compatible few-shot class-incremental learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 9046–9056
28 Ahmed N, Kukleva A, Schiele B. Orco: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 28762–28771
29 Han Y, Hui L, Jiang H, Qian J, Xie J. Generative subgraph contrast for self-supervised graph representation learning. In: Proceedings of European Conference on Computer Vision (ECCV), 2022. 91–107
30 Zhu X, Shu X, Tang J. Client-Unbiased Skeletal Action Recognizer in Federated Learning. IEEE Transactions on Image Processing, 2025
31 Peyré G, Cuturi M, Solomon J. Gromov-wasserstein averaging of kernel and distance matrices. In: Proceedings of International Conference on Machine Learning (ICML), 2016. 2664–2672
32 Zhu X, Shu X, Huang P, Tang J. Prompt-Guided Prototype-Aware Commonality and Discrimination Learning for Zero-Shot Skeleton-Based Action Recognition. IEEE Transactions on Multimedia, 2025
33 Zhang J, Liu L, Silven O, Pietikäinen M, Hu D. Few-shot class-incremental learning for classification and object detection: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025
34 Jiang Z, Yuan Y, Ma D, Wang Q, Yuan Y. Implicit CLIP Prior Decoupling for Few-Shot Remote Sensing Image Segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2025
35 Liu J, Ji Z, Pang Y, Yu Y. Ntk-guided few-shot class incremental learning. IEEE Transactions on Image Processing, 2024
36 Wang X, Ji Z, Yu Y, Pang Y, Han J. Model attention expansion for few-shot class-incremental learning. IEEE Transactions on Image Processing, 2024
37 Zhang Y, Ji Z, Pang Y, Han J, Li X. Modality-experts coordinated adaptation for large multimodal models. Science China Information Sciences, 2024, 67(12): 220107