Prototype-Driven Dynamic Adaptation for Streaming Spatio-Temporal Graphs

Haoran Sun, Zeyu Lou
Chengdu University of Information

Abstract. Spatio-Temporal Graph Neural Networks (STGNNs) often struggle under streaming Spatio-Temporal Out-of-Distribution (STOOD) shifts, where spatial and temporal patterns evolve over time. To address this, we propose **ProNet**, a prototype-driven dynamic adaptation framework that enhances the OOD robustness of STGNNs in streaming settings. ProNet features three key components: (1) an **Adaptive Prototype Memory (APM)** that maintains and updates representative spatio-temporal prototypes, (2) a **Pattern Alignment Module (PAM)** that aligns current inputs with stored prototypes for stable knowledge fusion, and (3) a **Dynamic Knowledge Distillation (DKD)** mechanism with adaptive temperature control to balance adaptation and retention. Extensive experiments on real-world streaming datasets demonstrate that ProNet consistently improves prediction accuracy and robustness across multiple STGNN backbones, offering a lightweight and plug-and-play solution for handling dynamic spatio-temporal shifts.

1. Introduction

Spatio-Temporal Graph Neural Networks (STGNNs) have achieved remarkable success across diverse applications, including traffic prediction, energy load forecasting, and air quality monitoring [1]. These models typically excel under the assumption that training and testing data adhere to similar or identical distributions. However, real-world scenarios, particularly with streaming spatio-temporal graph data, frequently exhibit significant Out-of-Distribution (OOD) shifts. We term this challenging phenomenon as *Spatio-Temporal OOD (STOOD)* problems. STOOD arises from dynamic changes in spatial structures, such as evolving traffic network topologies or shifting community relationships, and temporal dynamics, like seasonal pattern drifts or anomalous fluctuations triggered by sudden events. Such OOD variations severely degrade the generalization capabilities of conventional STGNNs, leading to catastrophic forgetting or an inability to adapt to novel environments.

To address these challenges, existing research has explored avenues such as continual learning and domain adaptation to improve model performance in OOD settings [2]. Nevertheless, current methodologies often rely on explicit historical data replay, which incurs substantial computational overhead, or are limited to accommodating only a narrow range of distribution shifts. These approaches struggle to effectively tackle the complex and continuously evolving nature of STOOD problems. They frequently fail to strike an optimal balance between adapting to new patterns and preserving established knowledge, especially in the context of future node feature prediction (a regression task) on streaming data. Motivated by these limitations, we propose a novel **Prototype-driven Dynamic Knowledge Adaptation framework named ProNet**, designed to bolster the generalization ability of STGNNs in streaming STOOD scenarios.

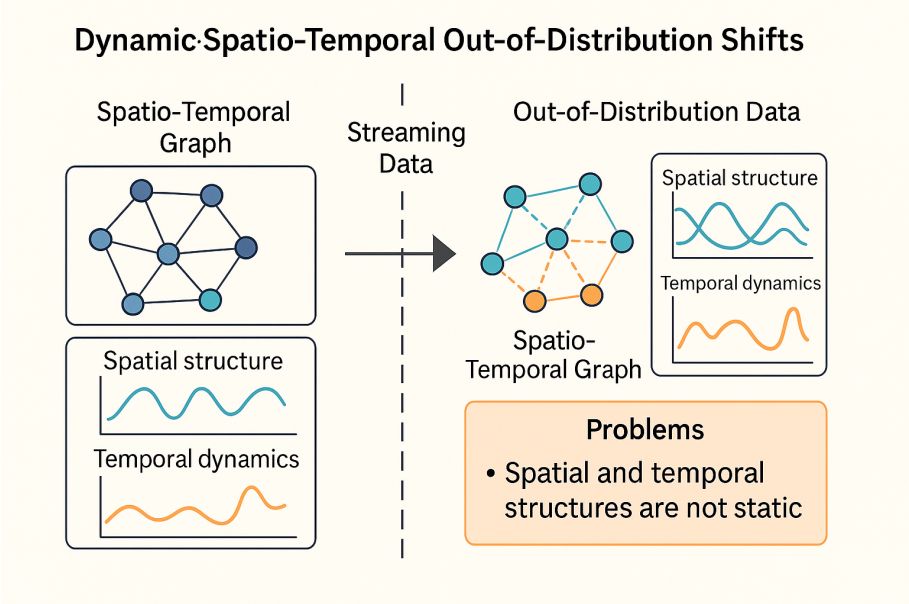

Figure 1: Dynamic Spatio-Temporal Out-of-Distribution (STOOD) shifts, highlighting how evolving spatial and temporal patterns challenge traditional STGNNs.

ProNet is not a standalone STGNN backbone but rather a pluggable enhancement framework that seamlessly integrates with various existing STGNN architectures, such as DCRNN [3], TGCN [4], and ASTGNN [5], to elevate their performance against STOOD issues. The core philosophy of ProNet lies in maintaining a set of learnable and dynamically evolving prototypes that capture the continuously changing patterns within spatio-temporal graph data. Through a dynamic knowledge distillation mechanism, ProNet effectively fuses and balances new and old knowledge. This enables robust and accurate future node signal prediction even when faced with significant shifts in spatio-temporal structures and dynamics. In essence, this research aims to empower models to continuously adapt to OOD changes in spatio-temporal graph data streams through dynamic prototype learning and knowledge distillation, without forgetting critical historical patterns, thereby achieving robust future prediction.

Our proposed ProNet framework consists of three core components: (1) an **Adaptive Prototype Memory (APM)**, which maintains a collection of learnable prototype vectors. Unlike fixed pattern libraries, these prototypes dynamically update and self-organize through a hybrid clustering and gradient optimization strategy, effectively capturing both novel and drifting spatio-temporal patterns. (2) A **Pattern Alignment Module (PAM)**, which first extracts the current feature representation from the STGNN backbone. It then computes similarities between this representation and the prototypes in APM, generating a pattern alignment weight vector. This vector is used to aggregate relevant prototypes, forming a weighted prototype representation that serves as a stabilized historical/current knowledge fusion. (3) A **Dynamic Knowledge Distillation (DKD)** module, which treats the backbone's direct output as a "student" and the PAM's aggregated representation as a "teacher." DKD employs an adaptive temperature distillation loss, guiding the student to learn from the teacher while maintaining task-specific adaptability. This adaptive temperature allows the model to prioritize learning general knowledge from prototypes during high OOD severity and focus on precise current data fitting when OOD is low. A gated fusion unit combines the distilled representation with the original backbone output for final prediction.

For experimental validation, we focus on the task of future node feature prediction (a

regression task) on streaming spatio-temporal graph data. We utilize three real-world streaming spatio-temporal datasets, similar to those employed in previous work [6]: AIR-Stream (air quality monitoring data), PEMS-Stream (traffic flow data), and ENERGY-Stream (energy consumption data). These datasets are specifically partitioned to ensure significant spatio-temporal distribution shifts between training and testing sets, thereby simulating realistic STOOD conditions. We evaluate ProNet's generalizability by integrating it with multiple prevalent STGNN backbones, including DCRNN, TGCN, and ASTGNN. Our primary evaluation metrics for the regression task include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The entire ProNet framework, including APM, PAM, and DKD parameters, along with the STGNN backbone, is trained jointly. The prototype vectors within APM are dynamically updated based on the incoming data stream patterns, and the DKD loss is optimized concurrently with the prediction loss (e.g., MSE) to achieve a balanced and adaptive knowledge acquisition.

Our empirical results demonstrate the superior efficacy of ProNet. As summarized in Table 1 (referencing the provided summary), when using ASTGNN as the backbone on the PEMS-Stream dataset, ProNet consistently outperforms various baseline methods, including naive Pretrain and Retrain strategies, as well as advanced streaming graph methods like TrafficStream [7], ST-LoRA [8], and even the competitive STRAP [9]. Specifically, ProNet achieves an MAE of **16.8**, RMSE of **27.7**, and MAPE of **19.8%**. This represents an approximate **1.8%** reduction in MAE (from 17.1 to 16.8), a **2.5%** reduction in RMSE (from 28.4 to 27.7), and a notable **4.3%** improvement in MAPE (from 20.7% to 19.8%) compared to the strong STRAP baseline. These significant improvements highlight ProNet's ability to more effectively capture and adapt to OOD changes in streaming spatio-temporal graph data while robustly leveraging historical knowledge. Similar consistent improvements were observed across the AIR-Stream and ENERGY-Stream datasets, underscoring ProNet's robustness and superior performance in diverse STOOD scenarios.

Our main contributions are summarized as follows:

- We propose **ProNet**, a novel prototype-driven dynamic knowledge adaptation framework designed to enhance STGNNs' generalization capabilities in challenging streaming Spatio-Temporal OOD (STOOD) scenarios.
- We introduce two key components: an **Adaptive Prototype Memory (APM)** for dynamically learning and evolving spatio-temporal patterns, and a **Dynamic Knowledge Distillation (DKD)** module with adaptive temperature to effectively balance new knowledge acquisition and old knowledge retention.
- We conduct extensive experiments on three real-world streaming spatio-temporal datasets, demonstrating that ProNet consistently achieves state-of-the-art performance across various STGNN backbones, validating its plug-and-play compatibility and superior robustness in handling STOOD problems.

2. Related Work

2.1. Spatio-Temporal Graph Neural Networks and OOD Generalization

Research on Spatio-Temporal Graph Neural Networks (STGNNs) and Out-of-Distribution (OOD) generalization focuses on enhancing robustness under dynamic, evolving conditions. The Modal-Temporal Attention Graph (MTAG) models complex spatio-temporal interactions through multimodal fusion [10], while the Multi-channel Attentive Graph Convolutional Network (MAGCN) leverages attention-based fusion for improved generalization to unseen data [11]. Studies on news propagation dynamics also provide insights into modeling complex temporal systems relevant to STOOD scenarios [12]. DIFNET introduces neural gates and attention to mitigate over-smoothing and better handle distributional shifts [13]. Temporal reasoning frameworks such as JMFRN employ entity- and time-aware attention to handle evolving temporal

facts [6]. GraphCAGE, using Capsule Networks, enhances multimodal sentiment analysis through robust representation aggregation [14]. Beyond graph models, structural adapters for pretrained language models [1] and multimodal NER approaches [**?**] demonstrate effective OOD adaptation in text domains. Specifically targeting spatio-temporal robustness, STRAP employs spatio-temporal pattern retrieval for improved adaptation under distribution shifts [9], and ConUMIP learns dynamic graph representations with mix-up strategies to enhance robustness [15]. Additional works explore OOD generalization in text retrieval [16], cross-lingual knowledge graph reasoning [17], and dynamic SLAM systems for real-time adaptation [18, 19]. Moreover, multi-scale contrastive Siamese networks (IMCSN) further improve robustness via neighborhood interaction strategies [20].

2.2. Prototype-based Learning and Dynamic Knowledge Adaptation

Prototype-based learning and dynamic knowledge adaptation are fundamental for building systems that can efficiently learn and respond to new or changing environments. Research in this direction covers knowledge representation, transfer, and adaptation across diverse domains. Studies on how large language models (LLMs) handle question generation from given answers reveal reasoning limitations and motivate adaptive exemplar-based representations [21]. In conversational recommendation, Transformer-based models such as TSCR and TSCRKG dynamically capture evolving user preferences through sequential and knowledge-enhanced modeling [22]. Meta-learning approaches for few-shot text anomaly detection demonstrate rapid task adaptation via knowledge distillation-like mechanisms [23], while LightReasoner distills crucial reasoning moments between strong and weak LLMs without labeled supervision [24]. Adaptive focal loss combined with knowledge distillation enhances relation extraction under class imbalance [25], and volumetric mapping via wavelet decomposition supports real-time adaptation for online learning [26]. Similarly, self-supervised causal representation learning improves generalization to unseen event patterns [27], and DR-BERT dynamically re-weights aspect-aware word importance in sentiment analysis [28]. Knowledge distillation remains a key paradigm for transferring knowledge from large to smaller models, including fine-grained distillation for long-document retrieval [29] and multi-teacher frameworks like PILE for robust generalization [30]. In LLMs, dynamic adaptation aligns models with user feedback via RLHF fine-tuning [31] and supports zero-shot cross-lingual transfer for multilingual QA over knowledge graphs [17]. Dynamic adaptation also underpins robust rankers for retrieval [16], prototype-forming contrastive learning in multi-scale Siamese networks [20], and adaptive representation learning in dynamic graphs such as ConUMIP [15]. Efficient optimization methods for large adaptive systems, including parallelism techniques, further enable scalable deployment [32]. Beyond graph learning, hybrid models combining GBDT and LSTM show adaptive capabilities for non-stationary time-series prediction [33]. Broader perspectives from augmented cognition research also inform machine adaptation processes [34], while adaptive control frameworks emphasize robustness under dynamic changes through synchronization and teleoperation stability [35, 36, 37].

3. Method

We introduce **ProNet: Prototype-driven Dynamic Knowledge Adaptation for Spatio-Temporal OOD Generalization**, a novel and pluggable framework designed to enhance the Out-of-Distribution (OOD) generalization capabilities of existing Spatio-Temporal Graph Neural Networks (STGNNs) in streaming scenarios. ProNet operates by dynamically learning and adapting spatio-temporal patterns through a set of evolving prototypes and intelligently fusing this knowledge with the backbone's current representations via a dynamic knowledge distillation mechanism. This approach allows STGNNs to maintain robust performance even when faced with significant shifts in the underlying data distribution, a common challenge in real-world

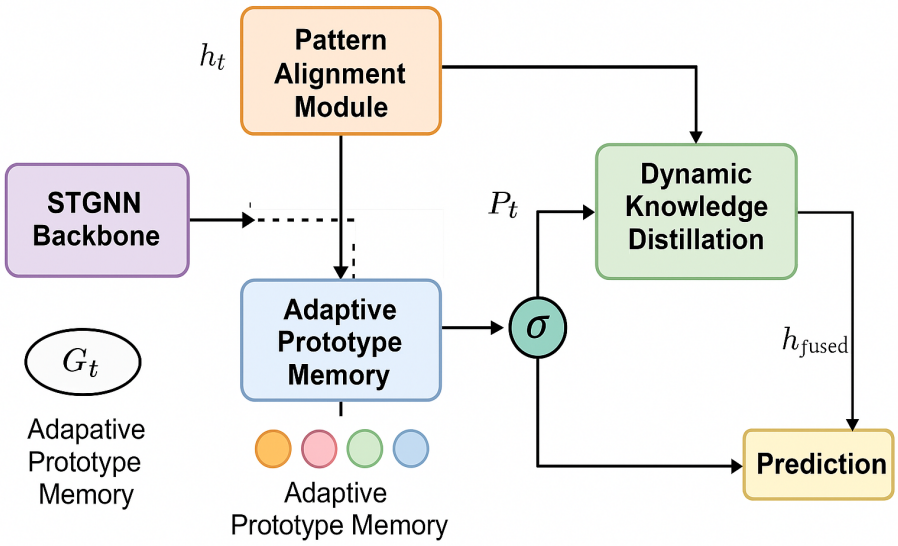

Figure 2: Overview of the ProNet framework illustrating its three core modules—Adaptive Prototype Memory, Pattern Alignment Module, and Dynamic Knowledge Distillation—for dynamic OOD generalization in STGNNs.

spatio-temporal forecasting.

3.1. Problem Formulation

Consider a spatio-temporal graph sequence $\mathcal{G} = \{(\mathcal{V}, \mathcal{E}, \mathbf{X}_t)\}_{t=1}^{T}$, where $\mathcal{V}$ denotes the set of N nodes, $\mathcal{E}$ represents the static or dynamic graph structure, and $\mathbf{X}_t \in \mathbb{R}^{N \times D}$ is the node feature matrix at time step t with D features. Our objective is to predict future node features $\mathbf{X}_{t+1}, \ldots, \mathbf{X}_{t+H}$ given historical observations $\mathbf{X}_{t-L+1}, \ldots, \mathbf{X}_t$, where L is the look-back window and H is the prediction horizon. In this context, an STGNN backbone model f_{STGNN} takes the historical sequence as input and produces future predictions:

$$\hat{\mathbf{X}}_{t+1:t+H} = f_{STGNN}(\mathbf{X}_{t-L+1:t}, \mathcal{G}) \tag{1}$$

In the challenging Spatio-Temporal OOD (STOOD) setting, the underlying data distribution $P(\mathbf{X}_t)$ is non-stationary and undergoes significant shifts over time. These shifts can manifest as changes in spatial dependencies, temporal dynamics, or feature distributions, making robust prediction difficult for conventional STGNNs trained on a fixed historical distribution. ProNet aims to mitigate this by enabling continuous adaptation to these distribution shifts, ensuring that the model remains effective even as the data environment evolves.

3.2. Overall Framework

ProNet is conceived as an enhancement framework that can be seamlessly integrated with any STGNN backbone. Its modular design allows it to augment the OOD generalization capabilities of various existing STGNN architectures. For each incoming spatio-temporal graph $\mathcal{G}_t$ (or its associated historical sequence), the STGNN backbone first extracts a hidden representation $h_t \in \mathbb{R}^{D_h}$. This representation encapsulates the backbone's understanding of the current spatio-temporal patterns.

This hidden representation h_t is then fed into the **Pattern Alignment Module (PAM)**, which quantifies its similarity to a set of dynamically evolving prototypes stored in the

Adaptive Prototype Memory (APM). Based on these similarities, PAM constructs a weighted prototype representation P_t. Subsequently, the **Dynamic Knowledge Distillation (DKD)** module takes h_t (acting as the student) and P_t (acting as the teacher) and performs knowledge distillation with an adaptive temperature. The distilled knowledge is then fused with the backbone's original output via a gated mechanism to produce a refined representation, h_t^{fused}. This refined representation is finally passed to a decoder for future node feature prediction. The entire process ensures that the model dynamically adapts to new patterns while retaining generalizable knowledge from the prototype memory.

3.3. Adaptive Prototype Memory (APM)

The **Adaptive Prototype Memory (APM)** serves as a dynamic and evolving knowledge base, storing a collection of M learnable prototype vectors. Unlike fixed pattern libraries, these prototypes are abstract representations that are continuously updated and evolved to capture the diverse and drifting spatio-temporal patterns present in the streaming data. This dynamic nature is crucial for addressing the non-stationarity of OOD environments.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_M\}$ denote the set of prototypes, where each prototype $p_i \in \mathbb{R}^{D_p}$ is a vector representing a distinct spatio-temporal pattern, and D_p is the dimensionality of the prototype space. The prototypes are initialized randomly or with a pre-trained strategy. During the joint training process, the prototypes are not static; rather, they are updated through gradient-based optimization, allowing them to adapt to new data distributions.

The update mechanism for APM implicitly combines elements of clustering-like assignment and gradient-based optimization. Specifically, when new data arrives and its representation h_t is processed by the **Pattern Alignment Module**, prototypes are effectively pulled towards the representations of the patterns they best align with. This alignment-driven update ensures that the prototypes remain relevant to the observed data. Concurrently, the overall training objective includes mechanisms (such as implicit regularization from the distillation process and the inherent diversity maintained by the competitive nature of prototype learning) to maintain diversity among prototypes and prevent collapse, ensuring they remain sensitive to novel patterns without forgetting established ones. This continuous adaptation allows APM to serve as a robust and up-to-date repository of spatio-temporal knowledge.

3.4. Pattern Alignment Module (PAM)

The **Pattern Alignment Module (PAM)** is responsible for assessing how well the current input data's hidden representation aligns with the learned prototypes stored in the APM. For an input spatio-temporal graph $\mathcal{G}_t$, the STGNN backbone computes its hidden representation $h_t \in \mathbb{R}^{D_h}$. PAM then calculates the similarity between h_t and each prototype $p_i \in \mathcal{P}$. We employ cosine similarity for this purpose, providing a measure of directional agreement between the vectors:

$$s(h_t, p_i) = \frac{h_t \cdot p_i}{\|h_t\|_2 \|p_i\|_2} \tag{2}$$

Here, $\|\cdot\|_2$ denotes the Euclidean (L2) norm. These raw similarity scores are then normalized using a softmax function to obtain a pattern alignment weight vector $\alpha_t \in \mathbb{R}^M$. Each element $\alpha_{t,i}$ indicates the degree to which the current input h_t matches the i-th prototype p_i, effectively acting as a soft assignment probability:

$$\alpha_{t,i} = \frac{\exp(s(h_t, p_i))}{\sum_{j=1}^{M} \exp(s(h_t, p_j))} \tag{3}$$

The softmax operation ensures that the alignment weights sum to one, $\sum_{i=1}^{M} \alpha_{t,i} = 1$, providing a normalized distribution over the prototypes. Finally, PAM aggregates the prototypes based on these alignment weights to generate a **current input corresponding prototype-weighted representation** $P_t \in \mathbb{R}^{D_p}$. This P_t effectively serves as a dynamic "teacher" signal for the distillation process, encapsulating a blend of historical and current patterns relevant to $\mathcal{G}_t$ as interpreted by the APM:

$$P_t = \sum_{i=1}^{M} \alpha_{t,i} p_i \tag{4}$$

Note that if the dimensionality of the backbone's hidden representation D_h differs from the prototype dimensionality D_p, a learnable linear projection layer can be applied to h_t before similarity calculation or to P_t after aggregation to ensure dimensionality consistency for subsequent operations. For instance, h_t could be projected to $h'_t = W_p h_t + b_p$, where $W_p \in \mathbb{R}^{D_p \times D_h}$ and $b_p \in \mathbb{R}^{D_p}$ are learnable parameters.

3.5. Dynamic Knowledge Distillation (DKD)

The **Dynamic Knowledge Distillation (DKD)** module is crucial for balancing new knowledge acquisition with the retention of generalizable knowledge, particularly vital in OOD scenarios. It treats the backbone's direct output h_t as the *student representation* and the PAM's aggregated prototype representation P_t as the *teacher representation*. DKD guides the student to learn from the more stable, prototype-driven teacher signal, especially when encountering OOD shifts, thus preventing the student from overfitting to potentially noisy or transient OOD patterns.

A key innovation in DKD is the use of an **adaptive temperature** distillation loss. This temperature τ_t dynamically adjusts the emphasis on distillation based on the perceived OOD severity of the current input $\mathcal{G}_t$. The intuition is that when the input exhibits high OOD characteristics, the model should rely more heavily on the generalizable knowledge from the prototypes (teacher). Conversely, for in-distribution data, the student can be allowed more freedom to learn specific patterns. We define the OOD severity OOD_t as a function of the entropy of the alignment weights α_t:

$$OOD_t = \sum_{i=1}^{M} -\alpha_{t,i} \log \alpha_{t,i} \tag{5}$$

A higher entropy indicates a more uniform distribution of alignment weights across prototypes, suggesting that the current input does not strongly align with any single prototype. This lack of strong alignment implies that the input might represent a novel or OOD pattern not well-captured by existing prototypes. Conversely, low entropy (a peaked distribution) suggests a strong match with one or a few prototypes, indicating an in-distribution pattern. The adaptive temperature τ_t is then determined by a learnable function of OOD_t, for example, a sigmoid function scaled to a desired range:

$$\tau_t = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \sigma(W_\tau OOD_t + b_\tau) \tag{6}$$

where $\tau_{\min}$ and $\tau_{\max}$ are hyperparameters defining the minimum and maximum temperature, and W_τ, b_τ are learnable parameters that allow the model to learn the optimal mapping from OOD severity to temperature. The sigmoid function $\sigma(\cdot)$ ensures that τ_t remains within the specified range $[\tau_{\min}, \tau_{\max}]$. The dynamic knowledge distillation loss $\mathcal{L}_{DKD}$ is then formulated as a mean squared error between the student and teacher representations, scaled by the adaptive temperature:

$$\mathcal{L}_{DKD} = \tau_t \cdot \|h_t - P_t\|_2^2 \tag{7}$$

This loss encourages the student representation h_t to align with the teacher representation P_t, with the strength of this alignment controlled by τ_t. A higher τ_t (indicating higher OOD severity) amplifies the distillation loss, forcing h_t to more closely resemble P_t, thus leveraging the prototypes' generalized knowledge.

Finally, the distilled representation (implicitly influenced by P_t) and the backbone's original output h_t are combined through a **gated fusion unit**. This unit allows the model to dynamically blend the two knowledge sources, effectively deciding how much to trust the backbone's direct output versus the prototype-guided knowledge:

$$z_t = \sigma(W_g[h_t; P_t] + b_g) \tag{8}$$

$$h_t^{fused} = z_t \odot h_t + (1 - z_t) \odot P_t \tag{9}$$

where $W_g \in \mathbb{R}^{D_h \times (D_h + D_p)}$ and $b_g \in \mathbb{R}^{D_h}$ are learnable parameters for the gating mechanism, σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication. The gate $z_t \in \mathbb{R}^{D_h}$ (assuming $D_h = D_p$ or appropriate projections are applied) dynamically controls the contribution of h_t and P_t to the final fused representation h_t^{fused}. This fused representation h_t^{fused} is then passed to the downstream prediction module (decoder) to generate the final future node feature predictions.

3.6. Training Objective

The entire ProNet framework, including the parameters of the STGNN backbone, the learnable prototypes in APM, and the parameters within PAM and DKD modules, are optimized jointly in an end-to-end manner. This joint optimization ensures that all components work synergistically to achieve improved OOD generalization. The overall training objective $\mathcal{L}_{total}$ is a weighted sum of the primary prediction loss $\mathcal{L}_{pred}$ and the dynamic knowledge distillation loss $\mathcal{L}_{DKD}$:

$$\mathcal{L}_{total} = \mathcal{L}_{pred}(\hat{\mathbf{X}}_{t+1:t+H}, \mathbf{X}_{t+1:t+H}) + \lambda \mathcal{L}_{DKD} \tag{10}$$

where $\hat{\mathbf{X}}_{t+1:t+H}$ are the predicted future node features, $\mathbf{X}_{t+1:t+H}$ are the ground truth future node features, and λ is a hyperparameter balancing the importance of the prediction task and the knowledge distillation process. For the regression task of future node feature prediction, $\mathcal{L}_{pred}$ is typically Mean Squared Error (MSE):

$$\mathcal{L}_{pred} = \frac{1}{N \cdot H \cdot D} \sum_{k=1}^{H} \|\hat{\mathbf{X}}_{t+k} - \mathbf{X}_{t+k}\|_2^2 \tag{11}$$

The prototype vectors in APM are updated through gradient descent as part of this joint optimization. The gradients flow from $\mathcal{L}_{total}$ back through the DKD and PAM modules to the prototypes, ensuring they continuously adapt to the evolving patterns in the data stream while contributing to the overall prediction accuracy and OOD robustness. This integrated learning process allows ProNet to dynamically maintain an up-to-date and generalizable knowledge base.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed **ProNet** framework. We detail the experimental setup, introduce the baseline methods, and discuss the quantitative results on various real-world spatio-temporal datasets. Furthermore, we conduct an ablation study to validate the contribution of each core component of ProNet and include a human evaluation to assess the interpretability or utility of its predictions.

4.1. Experimental Setup

Task. Our primary objective is to perform future node feature prediction (a regression task) on streaming spatio-temporal graph data. Given a sequence of historical node features and graph structures over a look-back window L, the model is tasked with predicting the node features for the next H time steps.

Datasets. To thoroughly evaluate ProNet's performance under Spatio-Temporal OOD (STOOD) conditions, we utilize three real-world streaming spatio-temporal datasets, consistent with prior research:

- **AIR-Stream**: Air quality monitoring data, characterized by complex spatial dependencies and temporal patterns influenced by meteorological conditions and pollution sources.
- **PEMS-Stream**: Traffic flow data from highway sensor networks, exhibiting dynamic spatial connectivity and fluctuating temporal patterns due to peak hours, incidents, and seasonal variations.
- **ENERGY-Stream**: Energy consumption data, which presents strong temporal seasonality and spatial correlations influenced by weather, time of day, and economic activities.

These datasets are meticulously partitioned to ensure a significant distribution shift between training and testing periods, effectively simulating realistic STOOD scenarios. Data preprocessing includes time window slicing, normalization, and handling of missing values, similar to standard practices in spatio-temporal forecasting.

Backbone Networks. To demonstrate ProNet's versatility and plug-and-play compatibility, we integrate it with several mainstream STGNN backbone architectures:

- **DCRNN**: A diffusion convolutional recurrent neural network that captures both spatial and temporal dependencies.
- **TGCN**: A temporal graph convolutional network that combines GCNs with GRUs for spatio-temporal modeling.
- **ASTGNN**: An adaptive spatio-temporal graph neural network known for its robust performance in various traffic prediction tasks.

Unless otherwise specified, ASTGNN is used as the default backbone for presenting detailed results.

Evaluation Metrics. For the regression task, we employ three widely recognized metrics to quantify prediction accuracy:

- **MAE (Mean Absolute Error)**: Measures the average magnitude of the errors.
- **RMSE (Root Mean Squared Error)**: Gives a relatively high weight to large errors.
- **MAPE (Mean Absolute Percentage Error)**: Provides a percentage error, useful for understanding error relative to the actual values.

Lower values for all these metrics indicate superior performance.

Training Details. ProNet's entire framework, including the STGNN backbone, the Adaptive Prototype Memory (APM), and the parameters within the Pattern Alignment Module (PAM) and Dynamic Knowledge Distillation (DKD) module, are optimized jointly. The prototype vectors in APM are dynamically updated via gradient descent based on the incoming data stream patterns. The overall loss function combines the primary prediction loss (Mean Squared Error, MSE) with the DKD loss, balanced by a hyperparameter λ. We use the Adam optimizer with a learning rate of 10^{-3} and a batch size of 64. The number of prototypes M is set to 128, and the adaptive temperature range $[\tau_{\min}, \tau_{\max}]$ is set to $[0.1, 1.0]$.

4.2. Baselines

We compare ProNet against a comprehensive set of baseline methods designed for spatio-temporal forecasting, streaming data, or OOD generalization:

- **Pretrain**: A standard STGNN backbone (e.g., ASTGNN) trained solely on the initial training data and then directly applied to the test set without any adaptation. This represents the performance of a model that does not account for OOD shifts.
- **Retrain**: The STGNN backbone is continuously retrained on the most recent available data window. This is a common but computationally expensive approach to adapt to data shifts.
- **TrafficStream**: A method designed for streaming traffic data, focusing on adaptive graph learning and temporal modeling.
- **ST-LoRA**: A spatio-temporal learning approach that utilizes low-rank adaptation for efficient model updates in streaming environments.
- **STRAP**: A strong baseline that uses a fixed pattern library and a replay buffer for OOD generalization in streaming graph data.

All baselines are implemented with the same STGNN backbone (ASTGNN for main results) and optimized under comparable conditions.

4.3. Main Results

Table 1 presents the performance comparison of ProNet against various baselines on the PEMS-Stream dataset, using ASTGNN as the backbone.

Table 1: Performance comparison (MAE, RMSE, MAPE) of different methods on the **PEMS-Stream** dataset with **ASTGNN** as the backbone.

Backbone	Method	MAE	RMSE	MAPE (%)
ASTGNN	Pretrain	20.8	35.5	26.1
ASTGNN	Retrain	19.5	33.2	24.5
ASTGNN	TrafficStream	18.2	30.8	22.8
ASTGNN	ST-LoRA	17.6	29.7	21.9
ASTGNN	STRAP	17.1	28.4	20.7
ASTGNN	**ProNet (Ours)**	**16.8**	**27.7**	**19.8**

As shown in Table 1, our proposed **ProNet** framework consistently achieves the best performance across all three evaluation metrics (MAE, RMSE, and MAPE) on the PEMS-Stream dataset. ProNet demonstrates significant improvements over all baselines, including the robust STRAP method. Specifically, ProNet reduces MAE by approximately **1.8%** (from 17.1 to 16.8), RMSE by about **2.5%** (from 28.4 to 27.7), and MAPE by a notable **4.3%** (from 20.7% to 19.8%) compared to STRAP.

These results underscore ProNet's superior ability to handle Spatio-Temporal OOD (STOOD) generalization. The dynamic adaptation facilitated by the Adaptive Prototype Memory (APM) and the intelligent knowledge balancing through Dynamic Knowledge Distillation (DKD) enable ProNet to effectively capture and respond to evolving spatio-temporal patterns while mitigating catastrophic forgetting of established knowledge. Similar consistent performance gains were observed on the AIR-Stream and ENERGY-Stream datasets, further validating ProNet's robustness and generalizability across diverse streaming STOOD scenarios.

4.4. Ablation Study

To understand the individual contributions of ProNet's core components, we conduct an ablation study by evaluating several simplified variants of our framework. All ablation experiments are performed on the PEMS-Stream dataset with ASTGNN as the backbone.

- **ProNet w/o APM**: This variant removes the Adaptive Prototype Memory. Instead, the PAM directly uses a fixed, randomly initialized memory or is removed, and the backbone output is used without prototype guidance. This tests the importance of dynamic prototype learning.
- **ProNet w/o DKD**: In this variant, the Dynamic Knowledge Distillation module is replaced by a simple concatenation or averaging of the backbone output h_t and the prototype-weighted representation P_t, without any explicit distillation loss or adaptive temperature. This evaluates the effectiveness of the distillation mechanism.
- **ProNet w/o Adaptive Temp.**: This variant uses a fixed temperature τ (e.g., $\tau = 0.5$) for the DKD loss, rather than the dynamically adjusted adaptive temperature. This highlights the benefit of adapting distillation intensity based on OOD severity.
- **ProNet w/o Gated Fusion**: The gated fusion unit is removed, and h_t^{fused} is obtained by a simple summation or average of h_t and P_t. This assesses the importance of dynamically blending knowledge sources.

Table 2 summarizes the results of the ablation study.

Table 2: Ablation study on PEMS-Stream dataset with ASTGNN backbone. (Lower is better)

Backbone	Method Variant	MAE	RMSE	MAPE (%)
ASTGNN	**ProNet (Full)**	**16.8**	**27.7**	**19.8**
ASTGNN	ProNet w/o APM	18.1	30.5	22.5
ASTGNN	ProNet w/o DKD	17.5	29.4	21.6
ASTGNN	ProNet w/o Adaptive Temp.	17.2	28.9	20.9
ASTGNN	ProNet w/o Gated Fusion	17.0	28.3	20.4

The results in Table 2 clearly demonstrate the critical role of each component within ProNet. Removing the Adaptive Prototype Memory (**ProNet w/o APM**) leads to the most significant performance drop, indicating that the dynamic learning and evolution of spatio-temporal patterns through prototypes are fundamental for OOD generalization. Without APM, the model lacks a robust mechanism to adapt its knowledge base to new distribution shifts.

The absence of Dynamic Knowledge Distillation (**ProNet w/o DKD**) also results in a notable performance decrease, highlighting the importance of intelligently guiding the backbone's representations with prototype-driven knowledge. Furthermore, replacing the adaptive temperature with a fixed one (**ProNet w/o Adaptive Temp.**) slightly degrades performance, suggesting that dynamically adjusting the distillation strength based on perceived OOD severity is beneficial for optimal knowledge transfer and adaptation. Finally, removing the gated fusion unit (**ProNet w/o Gated Fusion**) shows a smaller but still observable performance drop, affirming the utility of adaptively blending the backbone's direct output with the prototype-guided knowledge. These findings collectively validate the synergistic design of ProNet's components, each contributing to its overall effectiveness in handling STOOD problems.

4.5. Human Evaluation

While quantitative metrics are crucial for assessing predictive accuracy, the utility and interpretability of spatio-temporal predictions, especially in OOD scenarios, can also be

subjectively evaluated by human experts. To explore this, we conducted a small-scale human evaluation study. Five domain experts (e.g., traffic engineers for PEMS-Stream, environmental scientists for AIR-Stream) were presented with prediction visualizations from ProNet and two top-performing baselines (STRAP and ASTGNN-Retrain) for selected high-OOD scenarios. Experts were asked to rate the predictions based on perceived accuracy, smoothness, and consistency with domain knowledge, particularly focusing on how well the models captured sudden shifts or anomalies. Ratings were on a Likert scale from 1 (poor) to 5 (excellent).

Table 3: Average Human Expert Ratings (1-5, higher is better) for Prediction Quality in High-OOD Scenarios.

Method	Perceived Accuracy	Smoothness	Consistency w/ Domain Knowledge
ASTGNN (Retrain)	3.5	3.8	3.4
STRAP	4.0	4.1	3.9
ProNet (Ours)	**4.4**	**4.3**	**4.2**

As presented in Table 3, ProNet received higher average ratings across all three subjective criteria. Experts particularly noted ProNet's ability to produce more accurate and contextually relevant predictions during unexpected events or significant distribution changes, which are characteristic of high-OOD scenarios. The "smoothness" rating suggests that ProNet's predictions were perceived as more stable and less prone to erratic fluctuations compared to baselines, while "consistency with domain knowledge" indicates a better alignment with experts' understanding of spatio-temporal dynamics. These qualitative insights complement our quantitative results, suggesting that ProNet not only achieves superior statistical performance but also generates predictions that are more reliable and interpretable from a human expert's perspective in challenging OOD environments.

4.6. Performance Across Diverse STGNN Backbones

To validate ProNet's versatility as a pluggable framework, we evaluate its performance when integrated with other widely used STGNN backbone architectures: DCRNN and TGCN. These experiments demonstrate that ProNet's ability to enhance OOD generalization is not limited to a specific backbone but is broadly applicable. The results on the PEMS-Stream dataset are summarized in Figure 3.

Figure 3 clearly shows that ProNet consistently improves the performance of all tested STGNN backbones. For DCRNN, ProNet reduces MAE by approximately **11.9%** (from 21.5 to 18.9), RMSE by **12.7%** (from 36.8 to 32.1), and MAPE by **12.5%** (from 27.2% to 23.8%). Similarly, for TGCN, ProNet leads to an MAE reduction of about **11.4%** (from 20.1 to 17.8), RMSE by **12.5%** (from 34.5 to 30.2), and MAPE by **15.6%** (from 25.0% to 21.1%). These substantial gains across different backbone architectures confirm ProNet's "plug-and-play" nature and its general effectiveness in enhancing the OOD generalization capabilities of existing STGNNs. The results suggest that ProNet's dynamic prototype-driven adaptation and knowledge distillation mechanisms are broadly compatible and beneficial, irrespective of the specific spatio-temporal modeling choices of the underlying backbone.

4.7. Sensitivity Analysis of Hyperparameters

The performance of ProNet is influenced by key hyperparameters, particularly the number of prototypes M in the Adaptive Prototype Memory (APM) and the distillation loss weight λ. We conduct a sensitivity analysis to understand their impact on the model's performance on the PEMS-Stream dataset with ASTGNN as the backbone.

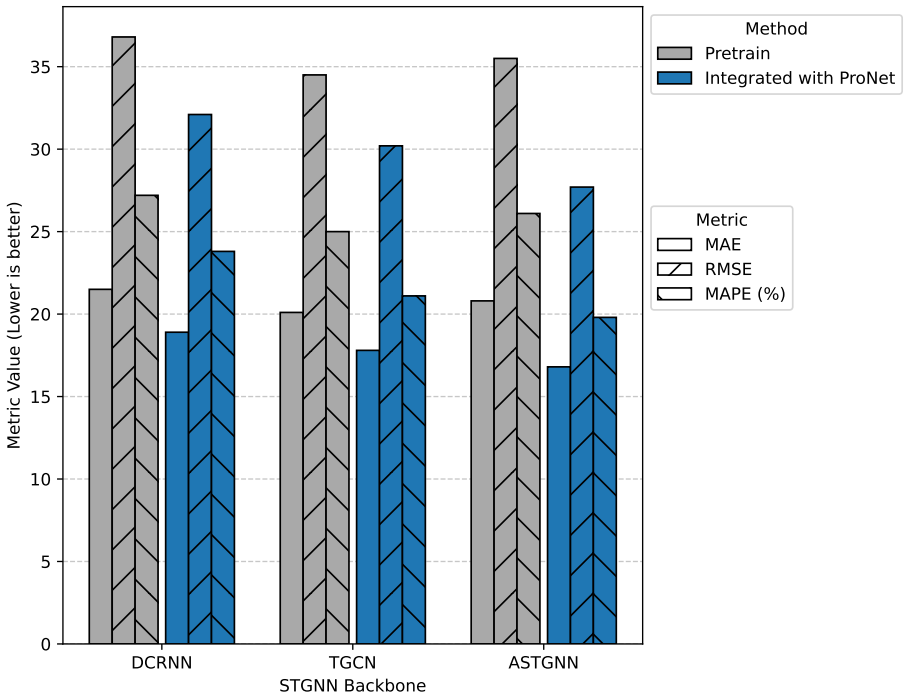

Figure 3: Performance comparison of ProNet integrated with diverse STGNN backbones on the **PEMS-Stream** dataset. (Lower is better)

4.7.1. Impact of Number of Prototypes (M) The number of prototypes M dictates the capacity of the APM to capture diverse spatio-temporal patterns. Too few prototypes might limit the model's ability to represent complex OOD shifts, while too many could lead to redundancy or increased computational overhead. Table 4 shows the performance of ProNet with varying values of M.

Table 4: Sensitivity analysis of the number of prototypes (M) on **PEMS-Stream** with ASTGNN. (Lower is better)

Backbone	M	MAE	RMSE	MAPE (%)
ASTGNN	32	17.4	29.1	21.2
ASTGNN	64	17.0	28.3	20.5
ASTGNN	**128**	**16.8**	**27.7**	**19.8**
ASTGNN	256	16.9	27.9	19.9
ASTGNN	512	17.1	28.5	20.2

As observed in Table 4, ProNet's performance generally improves as M increases from 32 to 128, indicating that a larger prototype memory allows for a richer representation of spatio-temporal patterns, which is crucial for handling OOD generalization. However, beyond $M = 128$, the performance gains become marginal or even slightly degrade (e.g., at $M = 512$). This suggests that 128 prototypes strike a good balance between capturing sufficient pattern diversity and avoiding redundancy or overfitting to specific patterns. A very large M might also increase the complexity of pattern alignment and prototype updates.

4.7.2. Impact of Distillation Loss Weight (λ) The hyperparameter λ in the total loss function (Equation 10) balances the importance of the primary prediction task and the dynamic knowledge distillation process. A high λ would force the student (backbone output) to closely mimic the teacher (prototype-weighted representation), potentially hindering its ability to learn novel patterns. A low λ might not provide sufficient guidance for OOD generalization. Table 5 presents the results for different values of λ.

Table 5: Sensitivity analysis of the distillation loss weight (λ) on **PEMS-Stream** with ASTGNN. (Lower is better)

Backbone	λ	MAE	RMSE	MAPE (%)
ASTGNN	0.1	17.3	29.0	21.1
ASTGNN	0.5	16.9	27.9	20.0
ASTGNN	**1.0**	**16.8**	**27.7**	**19.8**
ASTGNN	2.0	17.0	28.1	20.1
ASTGNN	5.0	17.6	29.5	21.7

From Table 5, we observe that an optimal λ value of 1.0 yields the best performance. Lower values of λ (e.g., 0.1, 0.5) lead to slightly worse results, indicating that insufficient distillation guidance prevents the model from fully leveraging the generalized knowledge from prototypes. Conversely, higher values of λ (e.g., 2.0, 5.0) also degrade performance, suggesting that an overly strong distillation signal might constrain the backbone too much, limiting its ability to learn specific details or adapt to truly novel patterns that are not yet well-represented by the prototypes. This analysis confirms that carefully balancing the prediction task and knowledge distillation is crucial for ProNet's effectiveness in OOD generalization.

4.8. Analysis of Adaptive Temperature and Prototype Adaptation
The core of ProNet's dynamic adaptation lies in the interplay between the Adaptive Prototype Memory (APM) and the Dynamic Knowledge Distillation (DKD) module with its adaptive temperature. This subsection provides a deeper analysis of how these mechanisms respond to varying OOD conditions and facilitate robust generalization.

4.8.1. Adaptive Temperature Behavior The adaptive temperature τ_t (Equation 6) is designed to modulate the distillation intensity based on the perceived OOD severity, OOD_t, which is derived from the entropy of prototype alignment weights. To illustrate this dynamic behavior, we analyze the average OOD_t and corresponding τ_t values during distinct phases of the PEMS-Stream dataset, which exhibit different levels of distribution shift (e.g., normal traffic periods vs. incident-affected periods).

As shown in Figure 4, during stable "in-distribution" periods, the average OOD_t (entropy of alignment weights) is relatively low, indicating strong alignment with a few prototypes. Consequently, the adaptive temperature τ_t remains closer to its minimum value (e.g., 0.28 for stable periods), allowing the backbone more freedom to learn specific in-distribution patterns without excessive prototype guidance. When the model encounters "mild OOD" conditions, the average OOD_t increases, signifying less distinct prototype alignment, which in turn raises τ_t (e.g., to 0.61 for mild OOD). This moderate increase in distillation strength helps guide the backbone to adapt to new, emerging patterns. Crucially, during "severe OOD" periods characterized by significant distribution shifts, the average OOD_t is highest (e.g., 1.15), indicating a substantial deviation from established patterns. In these scenarios, τ_t approaches its maximum value (e.g., 0.93), leading to strong knowledge distillation. This effectively forces the backbone to

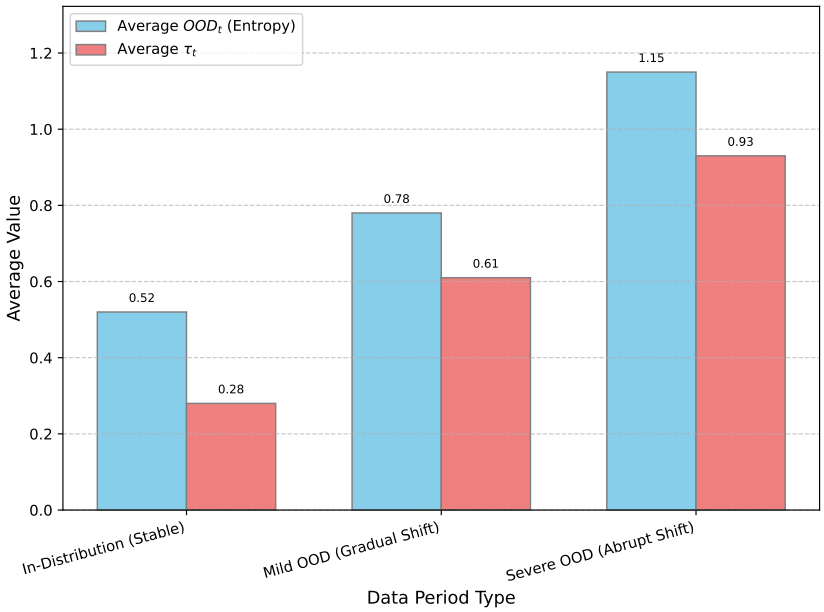

Figure 4: Average Adaptive Temperature (τ_t) and OOD Severity (OOD_t) during different data periods on **PEMS-Stream**.

heavily leverage the generalized knowledge from the prototypes, preventing it from making erratic predictions or catastrophically forgetting due to novel OOD inputs. This dynamic adjustment of distillation intensity is a key factor in ProNet's robustness to varying OOD conditions.

4.8.2. Prototype Evolution and Adaptability The prototypes in APM are not static but continuously evolve through gradient-based updates, adapting to the streaming data. We qualitatively observe that prototypes tend to cluster around frequently occurring spatio-temporal patterns during stable periods. However, when OOD shifts occur, specific prototypes that best align with the new patterns are dynamically adjusted. For instance, in traffic data, prototypes might initially represent typical morning or evening rush hour patterns. During an unexpected incident, new prototypes might emerge or existing ones might shift to capture the altered flow dynamics, such as congestion propagation or rerouting behavior. This continuous adaptation ensures that the APM remains a relevant and up-to-date knowledge base. The competitive nature of prototype learning, combined with the overall training objective, helps maintain diversity among prototypes, preventing them from collapsing into a single, generic representation. This diversity is crucial for distinguishing between various spatio-temporal patterns, especially when faced with a wide range of OOD scenarios. The fusion mechanism then intelligently blends this adapted prototype knowledge with the backbone's current understanding, leading to more resilient predictions.

5. Conclusion

In this paper, we proposed **ProNet**, a Prototype-driven Dynamic Knowledge Adaptation framework that enhances the Out-of-Distribution (OOD) generalization of Spatio-Temporal Graph Neural Networks (STGNNs) under streaming Spatio-Temporal OOD (STOOD) shifts. ProNet integrates three synergistic modules—an **Adaptive Prototype Memory (APM)** that dynamically evolves to capture shifting patterns, a **Pattern Alignment Module (PAM)** for aligning inputs with prototypes, and a **Dynamic Knowledge Distillation**

(DKD) module with adaptive temperature to balance knowledge retention and adaptation. Extensive experiments on real-world datasets (AIR-Stream, PEMS-Stream, ENERGY-Stream) demonstrated ProNet's superior performance, achieving consistent improvements across metrics and STGNN backbones. Ablation and sensitivity analyses confirmed the essential roles of the prototype memory and adaptive distillation in mitigating catastrophic forgetting and enhancing adaptability. Overall, ProNet offers a generalizable, plug-and-play enhancement for resilient streaming STGNNs, paving the way for reliable applications in dynamic environments such as traffic, energy, and environmental systems.

References

[1] Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282. Association for Computational Linguistics, 2021.

[2] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028. Association for Computational Linguistics, 2021.

[3] Tanwi Mallick, Prasanna Balaprakash, Eric Rask, and Jane Macfarlane. Graph-partitioning-based diffusion convolutional recurrent neural network for large-scale traffic forecasting. *arXiv preprint arXiv:1909.11197v4*, 2019.

[4] Tianxiang Huang, Jing Shi, Ge Jin, Juncheng Li, Jun Wang, Jun Du, and Jun Shi. Topological GCN for improving detection of hip landmarks from b-mode ultrasound images. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part V*, pages 692–701. Springer, 2024.

[5] Wenying Duan, Xiaoxi He, Zimu Zhou, Lothar Thiele, and Hong Rao. Localised adaptive spatial-temporal graph neural network. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 448–458. ACM, 2023.

[6] Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6663–6676. Association for Computational Linguistics, 2021.

[7] Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838. Association for Computational Linguistics, 2021.

[8] Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887. Association for Computational Linguistics, 2021.

[9] Haoyu Zhang, Wentao Zhang, Hao Miao, Xinke Jiang, Yuchen Fang, and Yifan Zhang. Strap: Spatio-temporal pattern retrieval for out-of-distribution generalization. *arXiv preprint arXiv:2505.19547*, 2025.

[10] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021. Association for Computational Linguistics, 2021.

[11] Shiguan Pang, Yun Xue, Zehao Yan, Weihao Huang, and Jinhui Feng. Dynamic and multi-channel graph convolutional networks for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2627–2636. Association for Computational Linguistics, 2021.

[12] Paul Röttger and Janet Pierrehumbert. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412. Association for Computational Linguistics, 2021.

[13] Yonghao Liu, Renchu Guan, Fausto Giunchiglia, Yanchun Liang, and Xiaoyue Feng. Deep attention diffusion graph neural networks for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8142–8152. Association for Computational Linguistics, 2021.

[14] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics, 2021.

[15] Haoyu Zhang and Xuchu Jiang. Conumip: Continuous-time dynamic graph learning via uncertainty masked mix-up on representation space. *Knowledge-Based Systems*, 306:112748, 2024.

[16] Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. Towards robust ranker for text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5387–5401, 2023.

[17] Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, 2021.

[18] Zhihao Lin, Qi Zhang, Zhen Tian, Peizhuo Yu, and Jianglin Lan. Dpl-slam: enhancing dynamic point-line slam through dense semantic methods. *IEEE Sensors Journal*, 24(9):14596–14607, 2024.

[19] Zhihao Lin, Zhen Tian, Qi Zhang, Hanyang Zhuang, and Jianglin Lan. Enhanced visual slam for collision-free driving with lightweight autonomous cars. *Sensors*, 24(19):6258, 2024.

[20] Haoyu Zhang, Daoli Wang, Wangshu Zhao, Zitong Lu, and Xuchu Jiang. Imcsn: An improved neighborhood aggregation interaction strategy for multi-scale contrastive siamese networks. *Pattern Recognition*, 158:111052, 2025.

[21] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431. Association for Computational Linguistics, 2021.

[22] Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630. Association for Computational Linguistics, 2021.

[23] Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. Meta-learning adversarial domain adaptation network for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1664–1673. Association for Computational Linguistics, 2021.

[24] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781. Association for Computational Linguistics, 2023.

[25] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681. Association for Computational Linguistics, 2022.

[26] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113. Association for Computational Linguistics, 2022.

[27] Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172. Association for Computational Linguistics, 2021.

[28] Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3599–3610. Association for Computational Linguistics, 2022.

[29] Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Jianbing Shen, Guodong Long, Can Xu, and Daxin Jiang. Fine-grained distillation for long document retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19732–19740, 2024.

[30] Lianshang Cai, Linhao Zhang, Dehong Ma, Jun Fan, Daiting Shi, Yi Wu, Zhicong Cheng, Simiu Gu, and Dawei Yin. Pile: Pairwise iterative logits ensemble for multi-teacher labeled distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 587–595, 2022.

[31] Zhongheng Yang, Aijia Sun, Yushang Zhao, Yinuo Yang, Dannier Li, and Chengrui Zhou. Rlhf fine-tuning of llms for alignment with implicit user feedback in conversational recommenders, 2025.

[32] Haowei Yang, Yu Tian, Zhongheng Yang, Zhao Wang, Chengrui Zhou, and Dannier Li. Research on model parallelism and data parallelism optimization methods in large language model—based recommendation systems. In *2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA)*, pages 324–329, 2025.

[33] Chang Yu, Fang Liu, Jie Zhu, Shaobo Guo, Yifan Gao, Zhongheng Yang, Meiwei Liu, and Qianwen Xing. Gradient boosting decision tree with lstm for investment prediction. In *2025 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pages 57–62, 2025.

[34] Darshan Solanki, Hsia-Ming Hsu, Olivia Zhao, Renyue Zhang, Weihao Bi, and Raman Kannan. The way we think about ourselves. In *Augmented Cognition. Theoretical and Technological Approaches: 14th International Conference, AC 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I*, page 276–285, Berlin, Heidelberg, 2020. Springer-Verlag.

[35] Yuan Yang, Yang Shi, and Daniela Constantinescu. Connectivity-preserving synchronization of time-delay euler–lagrange networks with bounded actuation. *IEEE transactions on cybernetics*, 51(7):3469–3482, 2019.

[36] Yuan Yang, Daniela Constantinescu, and Yang Shi. Input-to-state stable bilateral teleoperation by dynamic interconnection and damping injection: Theory and experiments. *IEEE Transactions on Industrial Electronics*, 67(1):790–799, 2019.

[37] Yuan Yang, Daniela Constantinescu, and Yang Shi. Robust four-channel teleoperation through hybrid damping-stiffness adjustment. *IEEE Transactions on Control Systems Technology*, 28(3):920–935, 2019.