STKGNN: Scalable Spatio-Temporal Knowledge Graph Reasoning for Activity Recognition

Gözde Ayşe Tataroğlu Özbulak*
University of Lausanne
Lausanne, Switzerland
University of Applied Sciences and Arts Western Switzerland HES-SO
Sierre, Switzerland
gozdeayse.tatarogluozbulak@unil.ch

Yash Raj Shrestha
University of Lausanne
Lausanne, Switzerland
yashraj.shrestha@unil.ch

Jean-Paul Calbimonte
University of Applied Sciences and Arts Western Switzerland HES-SO
Sierre, Switzerland
The Sense Research Center
Lausanne, Switzerland
jean-paul.calbimonte@hevs.ch

Abstract

The emergence of dynamic, high-volume data streams demands advanced reasoning frameworks to capture complex spatio-temporal relationships that are essential for enabling contextual understanding. However, current approaches often lack scalable and adaptable semantic representations in dynamic and spatio-temporal scenarios. To answer this need, we introduce a novel Spatio-Temporal Knowledge approach based on Graph Neural Networks (STKGNN) for activity recognition. This framework performs graph-based reasoning over semantically enriched Spatio-Temporal Knowledge Graphs (STKGs) constructed from open-source video datasets. By leveraging these custom STKGs, we propose three advanced Graph Neural Network (GNN) based architectures to recognize various activities. Accordingly, we establish a comprehensive approach for spatio-temporal reasoning that adapts to diverse Knowledge Graph structures by addressing adaptability, scalability, and temporal complexities. This framework enhances activity recognition and provides a foundation for wider dynamic or real-time applications in different domains including healthcare, autonomous systems, video surveillance, and various other fields.

CCS Concepts

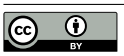
• **Computing methodologies** → **Knowledge representation and reasoning**; *Neural networks*; Activity recognition and understanding.

Keywords

Spatio-Temporal Knowledge Graph Reasoning, Graph Neural Networks, Activity Recognition

ACM Reference Format:
Gözde Ayşe Tataroğlu Özbulak, Yash Raj Shrestha, and Jean-Paul Calbimonte. 2025. STKGNN: Scalable Spatio-Temporal Knowledge Graph Reasoning for Activity Recognition. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761147

*Correspondence to:(gozdeayse.tatarogluozbulak@unil.ch.



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761147

1 Introduction

The rapid growth of dynamic, high-volume data streams across domains like healthcare or security systems demands scalable and interpretable reasoning frameworks which can capture contextual and temporal complexities. In this context, stream reasoning has shown promising results for real-time pattern inference and particularly in video-based systems, even though it poses unique challenges due to the spatio-temporal nature and rapidly changing interactions of streaming data.

In video-based applications such as activity recognition, it is essential to analyze dynamic interactions over time and space. Traditional methods, such as Convolutional Neural Networks (CNNs) and transformer-based architectures achieve impressive performance in extracting spatial and temporal features [3]. However, these methods often rely on pixel-level analysis, which is computationally demanding and lacks clear semantic insights. To alleviate this shortcoming, attention mechanisms, which are also widely integrated into Graph Neural Networks (GNNs), offer certain advantages due to their key feature weighting techniques. Despite that, such solutions struggle to generalize well for dynamic and multi-source datasets and fail to handle complex spatio-temporal relationships in reasoning tasks [26]. Knowledge Graphs (KGs) offer an alternative by creating an interpretable framework by their structured entities and relationships

KGs can encode domain knowledge for semantic representation, alleviating many reasoning problems with structured knowledge representations, and paving the way for higher-level semantic reasoning. In this context, KG embeddings have been integrated into a multi-graph CNN for efficient modeling of spatial-temporal dependencies, for example in traffic flow [15]. Also, in video-driven systems, Spatio-Temporal Knowledge Graphs (STKGs) extend this capability by encoding spatial relationships (e.g., object locations) and temporal dependencies (e.g., event sequences) [26]. Similarly, Temporal Knowledge Graph Embeddings [10] show how to model dynamic graph structures and time patterns by preserving semantic richness. Consequently, this brings more flexible reasoning and makes STKGs particularly well-suited for adaptive analysis of dynamic interactions.

While KGs offer numerous opportunities to represent the changes that occur in dynamic environments at the semantic level, they also present practical challenges in capturing the sequential dependencies and evolving structures in noisy video streams. Effective reasoning, especially on large-scale dynamic graphs, requires lightweight, scalable, and expressive models, but existing methods often

separate spatial and temporal reasoning. For example, the SGCN method in [9] focuses on static graph sparsification but neglects temporal complexities. These limitations emphasize the need to combine spatial and temporal reasoning while ensuring scalability.

In response to the needs described above, this work presents a novel context-aware framework that tackles the challenges of scalability and adaptability by integrating semantically rich spatial information with temporal information. This approach provides efficient STKG reasoning through our specialized GNN-based models for activity recognition in video streams. This modular design can pave the way for advances in video-derived reasoning systems in different domains by promising the adaptability of the approach beyond activity recognition. The main contributions of this work can be summarized as follows:

- *Construction of scalable STKGs:* We propose a novel algorithm to construct STKGs from open-source video datasets by encoding object-level spatial relationships and temporal dependencies across video frame sequences. This design supports structured and semantically enriched graph representations of complex video activities.
- *Specialized GNN models:* We introduce three progressively specialized GNN models which are designed to reasoning over custom STKGs with increasing structural and temporal complexity. These proposed architectures enable scalable and adaptive reasoning through static, temporal, and heterogeneous graphs (diverse activity types from multiple sources with varying complexity and contexts).
- *Comprehensive benchmarking protocol:* We provide a benchmarking setup that combines six diverse STKG datasets with a multi-metric, dual-setting (inductive and transductive reasoning) evaluation protocol. This enables rigorous, reproducible comparison of reasoning performance across different graph scales, class distributions, and temporal complexities. Also, this approach fills a critical gap in literature where such standardized STKG evaluation is absent.

To sum up, this study introduces a reproducible framework for GNN-based semantic reasoning over STKGs generated from video frame by addressing key challenges such as adaptability, temporal modeling with semantically rich information and scalability. Beyond activity recognition, the modular design of our framework supports adaptation to broader domains for advancements in video-driven reasoning systems. For reproducibility and further research, we make our models and the algorithm developed to generate the STKG datasets publicly available[1]

2 Related Work

Traditional recognition methods generally relied on high-resource-dependent CNNs and transformer-based architectures. Although these structures achieve detailed spatio-temporal pattern analysis through pixel-level interpretation, they often encounter scalability and adaptability because of high computational needs, as in [17]. In this context, [23] introduced VideoMAE v2 which is a dual-masking approach to enhance spatial-temporal understanding in video frames. Nevertheless, its high computational cost needs prevent practical application in dynamic environments. Similarly, [25] proposes a multi-modal knowledge extraction based framework. This integrates pre-trained vision models for contextual understanding in video recognition tasks. However, these methods are only effective for spatial-temporal feature extraction. This kind of methods face limitations in adaptability and efficient source usage due to model's extensive memory and parameters needs.

Knowledge graph driven models for activity recognition provide structured representations to enhance efficiency in recognition activities. In this regard, [12] proposes a multi-modal approach, Visual Knowledge Graph (VKG), to capture contextual human actions by linking body parts and interacting objects within a video. This approach combines different type of analysis (e.g., text , image understanding) to recognize static relationships within frames without temporal dependency modeling. Addressing the need for temporal reasoning, [28] introduces the Temporal Reasoning Graph (TRG), which focuses on single-layer temporal interactions. This limits its flexibility in handling complexities of datasets with various activities. Moreover, [26] proposes the Simple Spatio-Temporal Knowledge Graph (SSTKG), which integrates both spatial and temporal embeddings as in our scenario. However the model in this approach lacks dynamic activity modeling capabilities. Conversely, [11] emphasizes the learning of dynamic spatio-temporal relations in human activity over videos without scaling dynamic interaction into diverse actions.

Transfer learning and interaction modeling have explored to extend activity recognition capabilities across datasets of different domains. For instance, [20] introduces Graph Interaction Network (GIN) which focuses on relation transfer between different human activities in videos. Yet, the method does not have explicit interaction dependencies. Similarly, [14] proposes an object-relation reasoning graph that transfers learned relations to recognize human interactions. However, the model only relies on human-object interactions, which restricts scalability for dynamic environment [19].

Dynamic knowledge graphs for spatio-temporal activity recognition is proposed by [24] with a Deep Reasoning Model (DRM) which serves as a foundation for dynamic graph-based reasoning in visual data. This model focuses on object relationships on still images for social relation understanding. DRM highlights benefits of KGs in visual reasoning by prioritizing context nodes. However, it has deficiencies capturing evolving interactions and dependencies in video settings.

In contrast to all these prior methods, our work introduces a scalable framework that constructs semantically enriched STKGs. Owing to STKGs' structure, our proposed models can capture evolving interactions and temporal dependencies by offering efficient reasoning for activity recognition. Also our framework can be adapted to any video-driven analysis. As highlighted in Table 1, our method uniquely combines all semantic-aware context, scalability, dynamic and reasoning capabilities unlike other compared approaches. It provides dynamic spatio-temporal recognition by ensuring adaptability to video stream. Unlike computationally expensive pixel-wise methods of multi-modal analysis, our framework benefits lightweight pre-trained pixel-wise models only for the knowledge graph creation process. This makes our approach a cost-effective solution for semantic reasoning over dynamic large-scale streaming data.

[1]The repository can be accessed at: https://github.com/aislab-hevs/STKGNN.

Table 1: Main differences between the existing approaches in literature and our method.

MAIN RECOGNITION METHODS	MAIN FEATURES				
	Semantic-aware Context	Scalability	Pixel-wise Recognition	Dynamic Capabilities	Reasoning Capabilities
Traditional Recognition Methods [17, 23, 25]	-	-	+	±	-
Knowledge Graph-driven models [11, 12, 26, 28]	+	-	+	±	±
Transfer Learning and Interaction Modeling [14, 19, 20]	-	-	+	-	±
Dynamic KGs for Spatio-Temporal Recognition [24]	+	-	±	±	+
Our Proposed Approach (STKGNN)	+	+	+	+	+

Notes: "+" indicates the feature is fully supported, "±" indicates partial support or limitations, and "-" indicates the feature is not supported

3 Spatio-temporal Knowledge Graph (STKG) Creation

In this section, we introduce a novel algorithm that generates scalable STKGs from video streams of varying complexity. Our aim is to create a spatio-temporal knowledge graph that contains the relationships between objects related to the activity over frames, which will fill the gap that we observed in the literature. In this proposed algorithm, while nodes represent the detected objects, edges encode both spatial and temporal relationships between objects across consecutive frames. Thus, we obtain STKGs that represent pixel-based information connected by semantic relationships. Thanks to these structures, contextual relationships and temporal dependencies specific to dynamic video activities can be analyzed in depth with GNN-based modeling. These STKGs effectively represent complex and evolving interactions through their spatial-temporal features. With this proposed algorithm, scalable STKG datasets which contain multiple video samples and activity classes can be generated. To increase the robustness of our algorithm, we used multiple open-source video-based datasets obtained from diverse sources.

Open-source Datasets to Create STKGs: To ensure the reliability of the KG structure, three well-known open source datasets in activity recognition were used as follows:

- The *HMDB-51 (Human Motion Database 51)* [8] dataset consists of 51 classes short video clips of general human movements with limited quality resolution, collected from online sources.
- The *UCF-101 (University of Central Florida 101 Action Recognition Dataset)* [18] video dataset consists of many videos collected from YouTube, including 101 classes of sports, human movements and daily activities.
- The *Kinetics 400* [6] dataset consists of 400 classes of YouTube videos, covering various activities and complexities. The long length of videos in the dataset enable capturing of more detailed temporal relationships.

These open-source datasets offer a wide range of node and edge types with various representations and contexts to create rich STKG structure. The details of creation procedures are defined in Algorithm 1, which includes the following major steps:

Frame Sampling: The selected videos from open-source datasets are sampled at a specific frame rate to avoid unnecessary repetition by skipping some consecutive frames. Only frames where meaningful changes exist are processed. In our scenario, we assumed that the frame rate of a classic video is 25 FPS and processed 5 frames per second.

Algorithm 1 Spatio-temporal Knowledge Graph Creation From Open-source Video-based Datasets

Input: D: Set of video datasets, C: Number of distinct activity classes per dataset, V: Number of videos per class, F: Frame rate for sampling, OD: Object detection model, ST: Spatial threshold for edge creation, TT: Time threshold for temporal edge creation
Output: $STKG$: Set of main knowledge graphs, one per dataset
Initialization: $KG \leftarrow \emptyset$
for each dataset $d \in D$ **do**
  Select C distinct classes from d
  Select V videos from each selected class
  for each video v **do**
    Sample frames from v using F
    $Nodes\_N \leftarrow$ Detect objects in frames using OD
    for each detected object o in $Nodes\_N$ **do**
      Create node n
      Assign bounding box of o as node feature for n
      Assign class label of the video v to n as node label
    end for
    for each pair of consecutive frames f_1, f_2 **do**
      Compute $time\_difference$ between f_1 and f_2
      for each object (node) $n_1 \in f_1$ and $n_2 \in f_2$ **do**
        Compute $spatial\_distance$ between n_1 and n_2
        if $time\_difference \leq TT$ **AND** $spatial\_distance \leq ST$ **then**
          Create edge (n_1, n_2)
          Assign $spatial\_distance$ and $time\_difference$ as edge features
        end if
      end for
    end for
    Add subgraph from video v to main knowledge graph G_d
  end for
  Add G_d to $STKG$
end for
Return $STKG$

Node Creation: For each sampled frame, a pre-trained object detection model is applied to detect objects relevant to the activity in video. We used Faster-RCNN [16] due to its lightweight structure and high accuracy in detecting objects with detailed bounding box knowledge as desired in our scenario. Each detected object is represented as a node in the graph. Also the node features are derived from the detected object's bounding box coordinates. Besides, each node is assigned a label based on the corresponding activity class in the video dataset.

Edge Creation: In order to create edges between two nodes, two criteria are examined simultaneously as temporal and spatial threshold. We considered temporal threshold as the time difference between two consecutive frames. If this time difference is smaller than this temporal threshold value, the probability of creating an edge between the nodes is validated. Based on our empirical analysis among 0.2, 0.5, and 1.0 seconds, we set the time threshold value as 0.5 seconds. For the second criteria, the spatial distances of the nodes in both frames (euclidean distance between bounding boxes)

are calculated. It is determined whether this distance is smaller than or equal to the spatial threshold–experimentally defined–as 20 pixels. After these two criteria are valid, an edge is created between these two nodes. The combination of these spatial and temporal knowledge are used as edge features. This ensures accurate capturing of spatio-temporal relationships between objects.

Threshold Sensitivity Analysis: We observed that overly small temporal thresholds (e.g., 0.2s) fragmented interactions and reduced recall, while large values (e.g., 1.0s) introduced spurious edges and harmed precision. Similarly, setting the spatial threshold below 20 pixel tended to miss short-range interactions, whereas values above 30 pixel admitted false correspondences between unrelated objects. The chosen thresholds (0.5s and 20 pixels) provided a balanced trade-off. Small perturbations around these values produced only marginal variations in performance. This indicates that our framework is robust to moderate parameter changes.

Subgraph Construction: For each video, we generate a subgraph where nodes represent detected objects in frames and edges capture spatial and temporal relationships. Then we merge these subgraphs to form a unified STKG that represents object interactions related with the activity throughout the video.

By following the steps that we proposed by Algorithm 1, we constructed six STKG datasets with varying class distributions and complexities.Three of these custom STKGs are derived from a single-source, namely: **HMDB–STKG, UCF–STKG, and Kinetics–STKG**. Each of them consist of 15 activity classes and 120 video samples. The other three STKGs are derived from multiple sources (MS–STKGs) constructed by combining samples from all three open-source datasets: **MS–STKG–Small** (15 classes, 60 videos); **MS–STKG–Medium** (15 classes, 120 videos) and **MS–STKG–Large** (30 classes, 150 videos). These representative datasets are randomly selected to support a comparative evaluation. These custom datasets enable both isolated evaluation on single-source STKGs and for generalization analysis on balanced multi-source STKGs. Thus, it shows the adaptability of STKGNN to various domains.

4 Design of Specialized GNN Architectures for Reasoning Over Custom STKGs

Our semantic aware framework proposes three specialized GNN-based models designed to process various STKGs generated from open source video datasets. Instead of relying on a single model, we propose adaptable GNN architectures capable of processing a wide range of STKGs. This flexibility ensures robustness across different spatio-temporal reasoning scenarios in video-driven activity recognition tasks. For this purpose, we introduce a two-stage reasoning pipeline as illustrated by Figure 1, which represents the creation process of STKG over video frames and depicts our specialized GNN designs to model that STKG. In the first stage, STKGs are constructed by extracting contextual and temporal relationships from video frame sequences as defined in Algorithm 1. Since these STKGs serve as semantically rich representations, capturing the complex dynamics of human activity across heterogeneous datasets and proper modeling these relationships are major. Therefore, in the second stage, specialized GNN models integrated with CNN-based encoders are designed to perform scalable and generalizable reasoning on graphs of different sizes and complexity. Instead of developing a separate model for each dataset, we aim to combine reusable GNN components that can adapt to different contexts. All model structures are shown in Fig. 1, these introduce modular model architectures that support both structural consistency and computational efficiency in multi-source activity recognition scenarios.

4.1 StableGCN: Graph Convolutional Network (GCN) Based Model

Our GCN based [7] architecture brings together the components needed to deal with the challenges arising from structural complexity and relational dependencies in STKG reasoning. The StableGCN model depicted in Figure 1 uses layer normalization and skip connections to reduce oversmoothing and gradient distortion [5]. Cross-entropy loss and adaptive learning rate scheduling optimization are also incorporated into the architecture. This design approach provides a simpler and faster way to capture static details in STKG by modeling the contextual relations between node and edges.

GCN with layer normalization aggregate neighboring node information iteratively to adjust node features. In layer l, the feature update for node i is defined as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i) \cup \{i\}} \frac{1}{\sqrt{d_i d_j}} W^{(l)} h_j^{(l)}\right) \tag{1}$$

where d_i and d_j are degrees of nodes i and j, $h_i^{(l)}$ is the feature vector of node j at layer l, $W^{(l)}$ is the layer-specific weight matrix, and σ represents the ReLU activation.

Layer normalization is applied post GCN and standardizes feature distributions, thereby it stabilizes training and enables better generalization for diverse STKG.

Skip connections improves gradient flow and reduces information loss [5]. Besides, it helps retaining important node-specific information, temporal and relational differences by maintaining original input features at each layer, while also enhancing neighborhood aggregation. This process can be expressed as:

$$h_i^{(l+1)} = \text{GCNConv}(h_i^{(l)}, \text{edge\_index}) + h_i^{(0)} \tag{2}$$

$$L_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

where $h_i^{(0)}$ is the initial node feature, retained by all layers, $\hat{y}_i$ represents the predicted probability of the correct class.

Cross-entropy (CE) loss minimizes the difference between predicted and actual labels. It guides the model toward meaningful feature representation, which is critical for effective reasoning.

Learning rate scheduler enables more precise adjustments to the model parameters in later training stages, ensuring effective learning across diverse STKG scales. The learning rate η at step t is gradually reduced according to: $\eta_{t+1} = \gamma \cdot \eta_t$, with decay factor $\gamma = 0.75$ applied every 1000 steps, allowing finer adjustments in later epochs and robust performance across varied KG sizes.

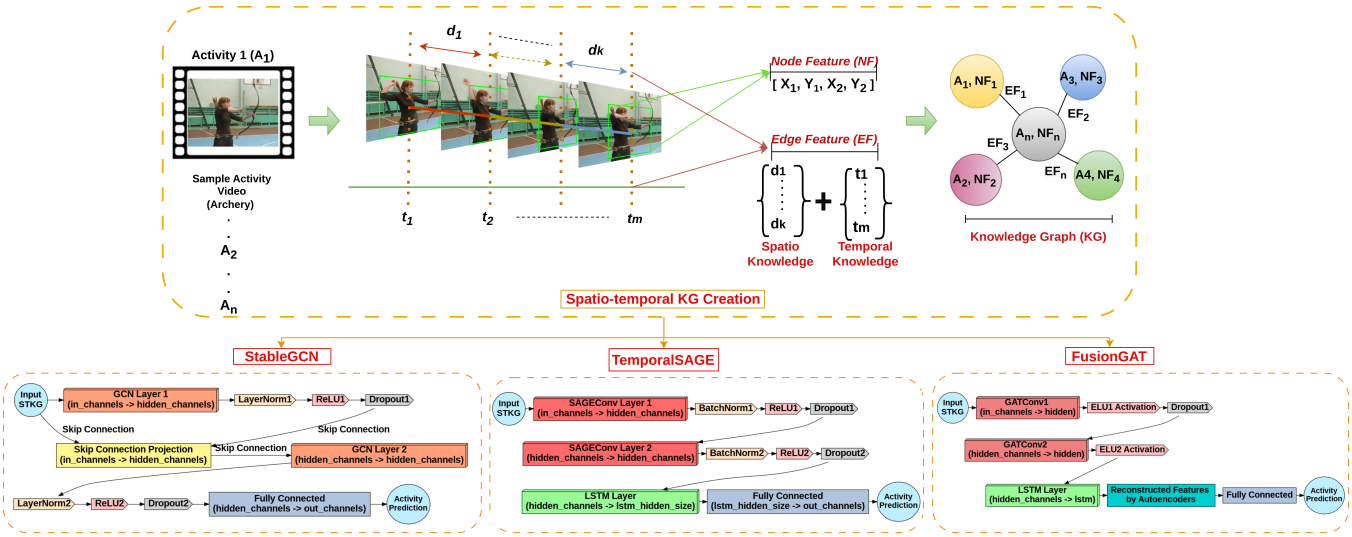


Figure 1: Overview of the proposed STKGNN framework that consists of STKG creation and specialized GNNs architectures. The top module shows the STKG construction, where each node A_i represents an activity element with node features (NF) derived from object coordinates, and edges (EF) represent spatio-temporal relationships based on spatial distance d_k and temporal intervals t_n. The bottom module represents the structure of three GNN-based models, namely StableGCN, TemporalSAGE and FusionGAT.

Table 2: Functional comparison of the proposed GNN architectures based on their design goals and spatio-temporal capabilities.

Core Design Aspects	StableGCN	TemporalSAGE	FusionGAT
Temporal Dependency Handling	Low (neighborhood-based)	Captured via LSTM	Fine-grained via Attention + LSTM
Generalization Strategy	Skip connections	Augmentation (edge manipulation)	Autoencoder-based latent reconstruction
Target KG Types	Simple, low-noise graphs	Mid-complex dynamic graphs	Complex, heterogeneous graphs
Design Purpose	Fast, lightweight reasoning	Temporal-aware intermediate reasoning	Deep semantic reasoning for complex graphs

4.2 TemporalSAGE: SAGEConv and LSTM Based Model

This model was developed to capture deeper relational dynamics and sequential dependencies within spatio-temporal KGs derived from video data and specialized activity recognition tasks. It is designed to handle the spatial-temporal interactions in activity data. It also combines SAGEConv [4] layers for efficient neighborhood feature collection, Batch Normalization for stable training, and a Long Short-Term Memory (LSTM) [2] layer to handle temporal dependencies. It also provides a composite loss function by integrating Cross Entropy Loss and Focal Loss to handle class imbalances in our specific STKGs. With this structure, the model improves reasoning capabilities in dynamic environments where contextual and temporal information is critical for activity understanding. The architecture of TemporalSAGE model is represented in Figure 1.

SAGEConv layers are used mainly to collect node neighborhood's features for static structure modeling. These layers are essential for handling heterogeneous STKGs where node and edge types vary. For each node i at layer l the following update function is applied:

$$h_i^{(l+1)} = \sigma\left(W_1^{(l)} h_i^{(l)} + W_2^{(l)} \sum_{j \in N(i)} \frac{h_j^{(l)}}{|N(i)|}\right) \tag{4}$$

where $h_i^{(l)}$ represents the feature vector of node i, $N(i)$ denotes the set of neighboring nodes, $W_1^{(l)}$ and $W_2^{(l)}$ are weight matrices, and σ is the ReLU activation function. This approach enables a more nuanced capture in spatial relationships by enhancing the semantic depth of node interactions.

LSTM layers process sequential dependencies, allowing the model to retain the temporal context, which is essential for understanding unfolding activities. Also, it allows to converge faster and achieve better generalization, particularly across STKGs with diverse feature distributions and scale variations.

Batch normalization is applied after each convolution layer to standardize feature distributions. It also improves training stability and convergence rates. By normalizing diverse features, the model is equipped to handle STKGs with diverse feature distributions and scale variations.

Composite loss optimizes classification performance by combining Cross-Entropy Loss and Focal Loss. This enhances robustness in case class imbalance. Focal Loss is defined as:

$$L_{\text{Focal}} = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{5}$$

where α is the weighting factor for large number of classes, and γ adjusts the focus on hard-to-classify instances. This composite

setup is calculated as:

$$L_{\text{composite}} = 0.1 \cdot L_{\text{CE}} + 0.9 \cdot L_{\text{Focal}} \tag{6}$$

where: L_{CE} is the Cross-Entropy Loss for classification accuracy, L_{Focal} aides in class balance and increasing robustness. The 0.9 ratio of this function in composite loss was determined empirically. TemporalSAGE also uses a *Learning Rate Scheduler* to improve convergency like StableGCN, yet $\gamma = 0.5$ is the decay factor which is applied in every 500 steps.

4.3 FusionGAT: Graph Attention Network (GAT) Based Model

This architecture distinguishes from previous models by providing deeper structure. It integrates GAT [21], LSTM for sequence processing, and a Graph Autoencoder (GAE) [22] for node feature reconstruction. It also defines Reconstruction Loss to improve the latent representation of STKG structures and benefits from *Batch Normalization* and *Learning Rate Scheduler* with the same parameters as in TemporalSAGE as visualized in Figure 1. This complex structure is useful for improving the generalization ability of the model in dynamic and multi-source graph structures.

GAT layers are used to enhance KG reasoning by selectively weighting the neighbors of nodes by using attention coefficients which allows the model to focus on the most relevant and important relationships. This is effective for modeling our STKGs where nodes and edge types have varying semantic importance. The flexibility of GAT layers to adapt to changing graph structures also ensures their effectiveness in dynamic and even in real-time applications. This is because they continuously recalibrate attention weights based on the most recent neighborhood context. For each node i at layer l, the GAT update is represented as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W^{(l)} h_j^{(l)}\right) \tag{7}$$

where $h_i^{(l)}$ is the feature vector of node i at layer l, $W^{(l)}$ represents the learnable weight matrix, and α_{ij} is the attention coefficient calculated based on the relevance of neighbor j to node i. Also, σ denotes the Exponential Linear Unit (ELU) activation function which provides faster and more stable convergence in sophisticated structures. *LSTM for sequential dependency modeling* supports spatio-temporal reasoning on our KGs by capturing essential temporal dependencies to recognize activity patterns over time.

Batch Normalization is applied after each GAT layer to standardize feature distributions, improving training stability and convergence rates. By normalizing features, the model is better equipped to handle the diverse input scales inherent to video-derived KGs.

Reconstruction loss is used to ensure the latent representation of node features. These are reconstructed accurately to enhance the model's understanding of the underlying graph structure. To reconstruct the original features from the latent embeddings, the GAE in FusionGAT model employs GAT layers as the encoder by creating an embedding for input features, subsequently applying a linear transformation in the decoder. We benefit from this approach to define our reconstruction loss function by Mean Squared Error (MSE) which measures similarity between the input and reconstructed features:

$$L_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 \tag{8}$$

where x_i and $\hat{x}_i$ represent the original and reconstructed node features, respectively, and N is the total number of nodes. This method enhances the hidden representation of node features by capturing fundamental patterns in the STKG structure.

Total loss is formulated by combining the classification and reconstruction losses to optimize both accuracy to improve complex spatio-temporal relationships in KGs as follows:

$$L_{\text{total}} = L_{\text{CE}} + 0.2 \cdot L_{\text{recon}} \tag{9}$$

where L_{CE} is the Cross-Entropy Loss, supporting activity classification; and L_{recon} is the MSE Loss for reconstructing node features. The 0.2 rate of this loss in our total objective function was defined by empirical analysis process.

The Learning Rate Scheduler is used by this model too with the same configuration as in the previous model. Since this strategy enhances overall model performance on complex, spatio-temporal based KG structures, it is preferred to be used in all custom models.

Table 2 summarizes the complementary motivations behind developing these three custom GNN architectures. Rather than designing a single large-scale model, we introduce a functional progression through the models from lightweight to complex architectures which are optimized for different reasoning demands. *StableGCN* offers fast inference over static graph structure; *TemporalSAGE* supplies temporal dynamics and robustness through LSTM and graph augmentation. *FusionGAT* enables deeper semantic understanding and generalization on heterogeneous (diverse activity types from multiple sources with varying complexity and contexts) by integrating attention mechanisms with autoencoder-based reconstruction. Together, these models form a scalable reasoning framework, where each model addresses a specific challenge and collectively validate the framework's adaptability to various STKG complexities.

5 Implementation Details

In this study, multiple evaluation metrics were used to address the different difficulty levels presented by multi-class and imbalanced classification tasks. Since reporting performance results only with accuracy metrics may be insufficient to capture performance variations across different class distributions, we present performance evaluations using precision, recall, F1, and Mean Reciprocal Rank (MRR) metrics to provide a more detailed characterization of model behavior. On the contrary to traditional metrics, we used MRR to measure the ranking quality of the model in our multi-class scenario. The MRR measure the quality of predictions' average reciprocal rank of the true label across all test instances. Thus, we aimed to reflect the capability of models to prioritize correct predictions.

In addition, we report performance metrics by using both inductive and transductive reasoning to prove that our proposed approach is compatible with the STKG data structure. Unlike traditional classification approaches that focus on labeling independent data points, our method emphasizes reasoning using structured KGs to detect semantic relationships and dependencies in spatial-temporal data. In this regard, our experimental studies use two main reasoning paradigms: **Inductive Reasoning** to evaluate the

model's ability to generalize nodes, edges, and features that are not previously seen during training, which are crucial for the analysis its adaptability to new scenarios; and **Transductive Reasoning** to test the model's performance using the full graph structure of the test set, except for its labels, similar to traditional classification.

On the other hand, to manage outliers, each model employs a targeted strategy that adjusts the representation of classes. For this purpose, excessive samples are aligned with the average sample count across all classes. This reduces class imbalance without affecting the distribution of other classes. Moreover, a five-fold cross-validation process is applied to verify generalizability and robustness of model performances.

The structure of our three models was designed separately considering their performances in order to address the difficulties of our proposed STKG structures. FusionGAT focused on complexity, StableGCN on faster reasoning, while TemporalSAGE, constitues a medium-scaled model. We apply a unique augmentation strategy involving edge addition and removal by optimizing the representation of complex, evolving relationships of TemporalSAGE. Data Augmentation strategy consists of the following two steps:

Edge addition via cosine similarity as first step consists in adding edges between nodes with high similarity in their feature representations, as demonstrated effectively in [13]. The similarity between nodes is calculated using the cosine similarity metric:

$$\text{cosine\_similarity}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \tag{10}$$

where v_i and v_j are the feature vectors of nodes i and j, respectively. An edge (v_i, v_j) is added to the graph only if the similarity score between nodes is greater or equal to a threshold value, $\tau_{\text{sim}} = 0.8$ which was determined empirically. Added edges are limited to 10% to ensure meaningful connections and avoid graph over-density.

Edge removal via degree centrality as second step is a kind of edge pruning process, which is applied based on degree centrality to manage graph density while retaining essential structural elements. By using Freeman's definition in [1], the degree centrality $c(v_i)$ for each node v_i is calculated as the ratio the degree of v_i to the maximum degree within $Graph$:

$$c(v_i) = \frac{\text{degree of } v_i}{\text{max degree in Graph}} \tag{11}$$

For an edge $e_{ij} = (v_i, v_j)$, we define a centrality score as: $(e_{ij}) = c(v_i) + c(v_j)$. Then edges are sorted by these scores, removing the the lowest 10% of edges. This ratio indirectly supports filtering out less significant edges by maintaining structural balance in the augmented graph.

6 Experimental Results

We conducted extensive experiments and reported our results across various configurations and analyses to address the absence of comparable benchmarks with similar data characteristics in the literature. To assess how effectively each model captures complex spatio-temporal relationships within KGs, we developed multiple models based on state-of-art GNN structures. Although there are similar methods in the literature, there is no direct model has graph based reasoning over STKGs with the same structural design and domain as we propose. Therefore, we designed our unique GNN-based models to be suitable for our custom STKG structure. To the best of our knowledge, there is currently no public STKG dataset that provides reasoning over video frames. To fill this gap, we propose a reproducible and adaptable data construction algorithm that allows researchers to create their own STKGs from open-source video datasets according to their domains.

In light of the above discussion, we present various experimental analysis as in Table 3 compares the performance of our three GNN models on structured diverse STKGs. All of the described experiments were completed on a machine with an Intel(R) Core(TM) i7 14700KF 14 Generation, 20 Cores, 61 MB Cache, NVIDIA(R) GeForce RTX(TM) 4080 SUPER, 16 GB GDDR6X. Both the training and evaluation were executed on the GPU using CUDA version 12.9. All results are obtained by appliying 5-fold cross-validation to ensure average performance metrics.

FusionGAT excels in both transductive reasoning (98% accuracy) and inductive reasoning (88.45%), demonstrating strong generalization capability owing to its complex structure. Despite this, our lightweight model, StableGCN performs well on simpler data but struggles with heterogeneous (data with diverse type and multi-source) datasets. Lastly, the moderate-complexity model, TemporalSAGE brings together temporal dependencies by achieving moderate results but falling short on high-variability data. While all models show worst performance on the noisy Kinetics-STKG dataset through both inductive and transductive reasoning, best performance is observed by MS-STKG-Small owing to its small variant and balanced class representation. Furthermore, the steady increase in performance across models on MS-STKG-Large suggests a scalable compromise between model capabilities and graphical complexity. It is also worth noting that, across all datasets, the highest performance was achieved by FusionGAT, which excels at processing multi-source and multi-variant datasets.

As previously we highlighted, a direct one-to-one **comparative analysis** is due to the lack of publicly available STKG datasets and spatio-temporal reasoning targeted in our proposed framework. Consequently, we focus inductive and transductive reasoning settings through our structured STKGs in the activity recognition domain. While no prior work supports our reasoning structures, we extend our analysis to include GNN-based traditional activity recognition baselines that are relevant in terms of domain but not reasoning-oriented approaches. In the light of these common objectives, we re-implemented three closely related, representative models—STIP-GCN [27], TRG [28], and AKU [12]—adapting them to operate on our MS-STKG-Medium dataset as shown in Table 4. To ensure compatibility, only the action recognition module of the multi-modal AKU method was retained. The other implementations were took over by considering the main details of the proposed model architectures. All these implementations were evaluated by using Top-1 and Top-5 accuracy metrics and served as a complementary benchmark to highlight the performance gap between traditional pixel-wise recognition pipelines and our reasoning enabled models. Experimental results shows that our proposed GNN based models consistently outperforms these baselines, by achieving highest Top-1 accuracy (85.07%) with our FusionGAT model which is significantly exceeding STIP-GCN (72.50%), TRG (70.51%), and AKU (67.42%). The remarkable performance of FusionGAT is due to its

Table 3: Performance comparison of our models on different sized STKGs derived from open-source video datasets.

DATASET	StableGCN (%)		TemporalSAGE (%)		FusionGAT (%)	
	Inductive ± Std	Transductive ± Std	Inductive ± Std	Transductive ± Std	Inductive ± Std	Transductive ± Std
HMDB-STKG	83.18 ± 0.63	94.09 ± 0.18	86.50 ± 0.27	95.17 ± 0.15	87.34 ± 0.24	98.10 ± 0.12
UCF-STKG	80.87 ± 0.24	90.73 ± 0.12	82.66 ± 0.21	93.09 ± 0.10	84.60 ± 0.44	96.73 ± 0.16
Kinetics-STKG	70.77 ± 0.98	82.49 ± 0.27	70.98 ± 0.44	88.67 ± 0.11	74.37 ± 0.84	93.28 ± 0.22
MS-STKG-Medium	80.71 ± 0.34	90.91 ± 0.15	81.88 ± 0.39	93.10 ± 0.09	84.15 ± 0.68	96.70 ± 0.15
MS-STKG-Large	75.90 ± 0.17	85.55 ± 0.21	78.48 ± 0.39	89.43 ± 0.13	81.67 ± 0.20	95.07 ± 0.21
MS-STKG-Small	86.79 ± 0.49	95.47 ± 0.21	87.83 ± 0.46	96.48 ± 0.19	88.45 ± 0.72	97.79 ± 0.23

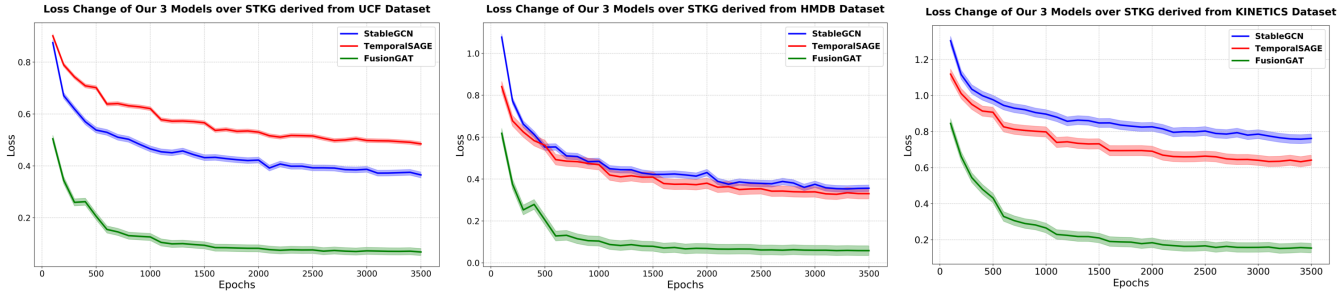


Figure 2: Loss characteristic of models with our STKGs derived from three open-source datasets.

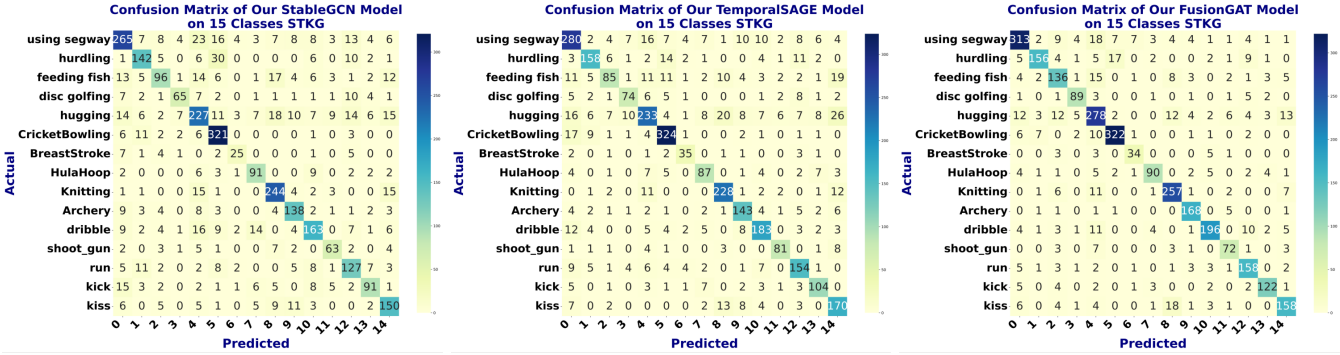


Figure 3: Confusion matrices for our 3 models over 15 classes MS-STKG-Medium derived from 'HMDB+UCF+Kinetics'.

Table 4: Comparison of Top-1 and Top-5 test accuracies (%) achieved by our proposed models and closely related GNN-based baselines on the MS-STKG-Medium dataset.

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)
FusionGAT (Ours)	85.07	98.39
TemporalSAGE (Ours)	80.30	97.75
StableGCN (Ours)	79.38	97.74
STIP-GCN [27]	72.07	97.65
TRG [28]	70.51	95.50
AKU [12]	67.42	95.32

well-designed architecture that incorporates attention mechanisms, temporal sequence modeling, and graph autocoder-based semantic reconstruction. Unlike STIP-GCN, which lacks temporal modeling, TRG, which omits spatial semantics, and AKU, which relies on early-stage visual fusion without reasoning. However our models effectively capture both spatial and temporal dependencies. Moreover, the performance gap between Top-1 and Top-5 accuracy serves as a robust indicator of model certainty and semantic precision. Our FusionGAT model narrows this Top-1/Top-5 gap (13.3%) further by indicating not only strong semantic recognition capabilities but also precise ranking of prediction confidence. This analysis supports broader comparative analysis discussed in Section 2 and reinforces scalability and uniqueness of our framework. Although these baselines do not incorporate explicit STKG reasoning, they represent the closest comparable activity recognition approaches in the literature, thereby contextualizing the performance gains of our framework.

Figure 2 presents the training dynamics of our three specialized models on our custom STKGs derived from open source datasets. It

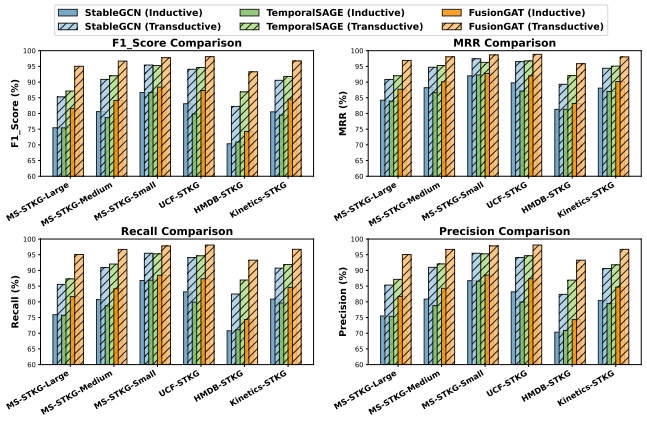


Figure 4: Performance metrics across our 3 unique models over various STKGs. Solid bars represent inductive reasoning and striped bars represents transductive reasoning.

provides the opportunity to observe how learning trends change on which dataset. For UCF-STKG, which has consistent videos, all models show a steady decline in loss. Moreover, owing to the balanced character of HMDB-STKG, all models presents faster loss optimization. For the Kinetics-STKG, all models excepting FusionGAT perform slightly worse learning tendency because of the noisy characteristics of the dataset. On the other hand, FusionGAT achieves the lowest final loss, demonstrating strong adaptability through all the three datasets. However, in general it can be said that all models can adapt to different dataset variations owing to their reasonable model structures.

Figure 3 compares the class-level performance of our proposed GNN models on the MS-STKG-Medium dataset. It is observed that all models correctly classify structured clear actions such as "archery" and "CricketBowling" which have distinct motion patterns. It can be said that misclassifications are more frequent in activities such as "haulRope" and "hurdling" where overlapping dynamic visuals and subtle motion differences cause confusion. On the other hand, it is observed that FusionGAT shows the most consistent class accuracy by using its attention and temporal modules to better process overlapping spatio-temporal signals. TemporalSAGE shows a slight improvement over StableGCN by including temporal context.

Figure 4 illustrates comprehensive model performance by different evaluation metrics over our various STKG datasets through inductive and transductive reasoning. The aim of this figure is to demonstrate the adaptability, robustness and generalizability of the framework we propose in a sophisticated manner. In this context, FusionGAT, shown in orange, demonstrates high success in both inductive and transductive reasoning in complex and multi-source datasets (MS-STKG-Large, MS-STKG-Medium), demonstrating that it can generalize the learned relationships and make effective inferences over spatio-temporal relations. Also, TemporalSAGE is better than StableGCN, especially in terms of F1 Score and Recall by offering stronger generalization and time sensitivity in both inductive and transductive scenarios. TemporalSAGE also achieves better recall and accuracy than StableGCN on complex datasets. However, both struggle with lower precision and recall on multi-source KGs.

7 Discussion and Conclusion

In this work, we proposed STKGNN, a robust and scalable approach for spatio-temporal reasoning by integrating STKGs into activity recognition tasks. With this approach, unlike traditional pixel-based methods, we capture both spatial and temporal dependencies and enable semantic reasoning through structured representations. We also propose an adaptive algorithm that transforms raw video data into semantically enriched STKGs using object-level information and temporal relationships.

To promote diversity and robustness, STKGs were constructed using three separate open-source datasets (UCF101, HMDB51, and Kinetics400), allowing evaluations across a wide range of activity types and graphical complexities. To this end, STKGNN comprises a series of increasingly specialized GNN architectures (StableGCN, TemporalSAGE, and FusionGAT) that have been designed to cope with the reasoning challenges posed by dynamic and multi-source video streams and each tailored to specific spatio-temporal demands. All these models consist of different scale of hierarchy which balances computational efficiency and scalability. Thereby, they enable the system to scale from simple structured graphs to complex, multi-source STKGs.

Our experimental results demonstrate that the proposed approach consistently achieves high accuracy across diverse reasoning environments. Among all developed models, FusionGAT attains the highest performance, particularly under inductive reasoning scenarios, where the challenge of generalization is most difficult. This proves that our models perform consistently across STKGs of varying size and complexity. These findings confirm the adaptability of our proposed framework and its ability to handle real-world variability in graph structure, temporal dependencies, and event meanings. Furthermore, the consistent results we obtain with metrics such as F1 score, precision, recall, and MRR express the reliability and semantic accuracy of the proposed models.

Apart from all these findings, we can say that our proposed reproducible STKG generation algorithm fills an important gap in the literature. This will enable researchers to create domain-specific STKGs (e.g., healthcare, autonomous systems, and security systems) from open-source video datasets in future work. In this context, as a future work, we aim to adapt this proposed approach to real-time and decentralized environments for reasoning over continuous video streams. Thus, the framework can be a basis for generalizable and explainable video-based intelligence systems.

Acknowledgements

This work was supported by the Swiss National Science Foundation through the StreamKG project with grant number 213369.

GenAI Usage Disclosure

No GenAI tools were used in any research stage of this work, neither the writing nor the data & code preparation.

References

[1] Linton C Freeman et al. 2002. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge* 1 (2002), 238–263.

[2] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2016), 2222–2232.

[3] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. 2021. A Survey on Deep Learning for Human Activity Recognition. *Comput. Surveys* 54, 8 (Oct. 2021), 1–34.

[4] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017), 1025–1035.

[5] Ajay Jaiswal, Peihao Wang, Tianlong Chen, Justin Rousseau, Ying Ding, and Zhangyang Wang. 2022. Old can be gold: Better gradient flow can make vanilla-gcns great again. *Advances in Neural Information Processing Systems* 35 (2022), 7561–7574.

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV] https://arxiv.org/abs/1705.06950

[7] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs.LG] https://arxiv.org/abs/1609.02907

[8] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, US, 2556–2563.

[9] Jiayu Li, Tianyun Zhang, Hao Tian, Shengmin Jin, Makan Fardad, and Reza Zafarani. 2020. SGCN: A graph sparsifier based on graph convolutional networks. In *Proc. 24th Pacific-Asia Conference, PAKDD*. Springer, Cham, 275–287.

[10] Qianyu Li, Jiebin Chen, Xiaoli Tang, Han Yu, and Hengjie Song. 2024. Modeling Time Decay Effect in Temporal Knowledge Graphs via Multivariate Hawkes Process. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, US, 1–8.

[11] Zhenyu Liu, Yaqiang Yao, Yan Liu, Yuening Zhu, Zhenchao Tao, Lei Wang, and Yuhong Feng. 2020. Learning dynamic spatio-temporal relations for human activity recognition. *IEEE Access* 8 (2020), 130340–130352.

[12] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. 2022. Visual knowledge graph for human action reasoning in videos. In *Proc. 30th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 4132–4141.

[13] Hieu V Nguyen and Li Bai. 2010. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*. Springer, Berlin, 709–720.

[14] Yangjun Ou, Li Mi, and Zhenzhong Chen. 2022. Object-relation reasoning graph for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, US, 20133–20142.

[15] Jipeng Qian, Suchi Li, and Xin Fu. 2024. Knowledge Graph Enhanced Dynamic Multi-Graph Convolutional Network for Traffic Origin-Destination Forecasting. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, US, 1–8.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE trans. on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.

[17] Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra, and Ajai Kumar. 2022. A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets. *Applied Artificial Intelligence* 36, 1 (2022), 2093705.

[18] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 [cs.CV] https://arxiv.org/abs/1212.0402

[19] Peng Su and Dejiu Chen. 2024. Adopting Graph Neural Networks to Analyze Human–Object Interactions for Inferring Activities of Daily Living. *Sensors* 24, 8 (2024), 2567.

[20] Yansong Tang, Yi Wei, Xumin Yu, Jiwen Lu, and Jie Zhou. 2020. Graph interaction networks for relation transfer in human activity videos. *IEEE Trans. on Circuits and Systems for Video Technology* 30, 9 (2020), 2872–2886.

[21] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. arXiv:1710.10903 [stat.ML] https://arxiv.org/abs/1710.10903

[22] Chun Wang, Shirui Pan, Guodong Long, Xingquan Zhu, and Jing Jiang. 2017. Mgae: Marginalized graph autoencoder for graph clustering. In *Proc. ACM Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 889–898.

[23] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, US, 14549–14560.

[24] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep reasoning with knowledge graph for social relationship understanding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden) *(IJCAI'18)*. AAAI Press, USA, 1021–1028.

[25] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. 2023. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, US, 6620–6630.

[26] Ruiyi Yang, Flora D Salim, and Hao Xue. 2024. SSTKG: Simple Spatio-Temporal Knowledge Graph for Intepretable and Versatile Dynamic Information Embedding. In *Proc. ACM Web Conference 2024*. Association for Computing Machinery, New York, NY, USA, 551–559.

[27] Sravani Yenduri, Vishnu Chalavadi, and C Krishna Mohan. 2022. STIP-GCN: Space-time interest points graph convolutional network for action recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, US, 1–8.

[28] Jingran Zhang, Fumin Shen, Xing Xu, and Heng Tao Shen. 2020. Temporal reasoning graph for activity recognition. *IEEE Transactions on Image Processing* 29 (2020), 5491–5506.