

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

Gözde Ayşe Tataroğlu Özbulak<sup>1,2</sup>, Yash Raj Shrestha<sup>1</sup>, Jean-Paul Calbimonte<sup>2,3</sup>

<sup>1</sup>University of Lausanne

<sup>2</sup>University of Applied Sciences and Arts Western Switzerland HES-SO

<sup>3</sup>The Sense Innovation and Research Center

The ACM Conference on Information and Knowledge Management (CIKM) 2025

Seoul, Korea



# Application Domains for Spatio-Temporal Streaming Data



**Autonomous driving:** video streams capturing changing traffic scenes



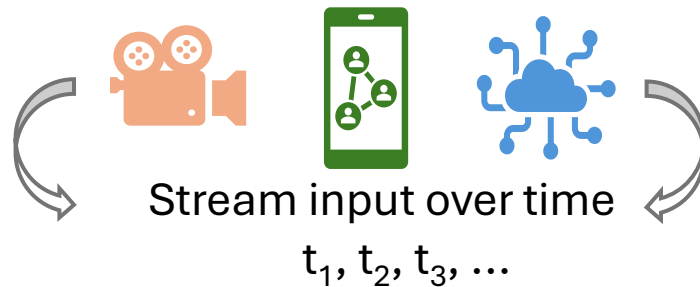
**Smart-city IoT:** sensor networks tracking environmental factors



**Healthcare monitoring:** wearable sensors recording patient states



**Social-media analysis:** detecting events from geotagged videos or posts

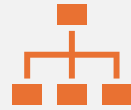


... video, sensors, social media logs, transaction records ...

# Background & Motivation

- Rapid growth of spatio-temporal data in video streams.
- In activity recognition, existing models emphasize visual perception rather than reasoning.
- Treating spatial and temporal cues separately limits contextual understanding of activities.
- Need for unified spatio-temporal reasoning to capture evolving activity relations.
- To scale the framework for dynamic and heterogeneous video data.
- Integrate semantic, spatial, and temporal cues for context-aware activity reasoning.
- Move beyond pixel-level perception toward structured relational understanding.

# Research Gap & Limitations



Absence of unified spatio-temporal representations in existing activity recognition models



Insufficient capacity for semantic and relational reasoning in dynamic activity contexts



Poor adaptability and generalization capacity of existing models across heterogeneous video sources



Lack of scalability in current frameworks for large, evolving spatio-temporal graphs

# Problem Statement & Challenges

## Problem Statement


Existing approaches cannot jointly reason over structured and dynamic spatio-temporal information, limiting contextual consistency and generalization.

## Key Challenges

- Joint Modelling
- Dynamic Representation
- Relational Reasoning
- Generalization
- Scalability

# Overview Proposed Method

Main Recognition Methods	Semantic-aware Context	Scalability	Pixel - wise Recognition	Dynamic Capabilities	Reasoning Capabilities
Traditional Recognition Methods	✗	✗	✓	<i>Partial</i>	✗
Knowledge Graph-driven Models	✓	✗	✓	<i>Partial</i>	<i>Partial</i>
Transfer Learning and Interaction Modeling	✗	✗	✓	✗	<i>Partial</i>
Dynamic KGs for Spatio-Temporal Recognition	✓	✗	<i>Partial</i>	<i>Partial</i>	✓
<b>Our Method — STKGNN</b>	✓	✓	✓	✓	✓

  
 Bridging  
 limitations  
 to  
 proposed  
 framework



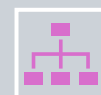
Video Input Stream



Feature Extraction for Node & Edge Construction



Temporal & Spatial Linking



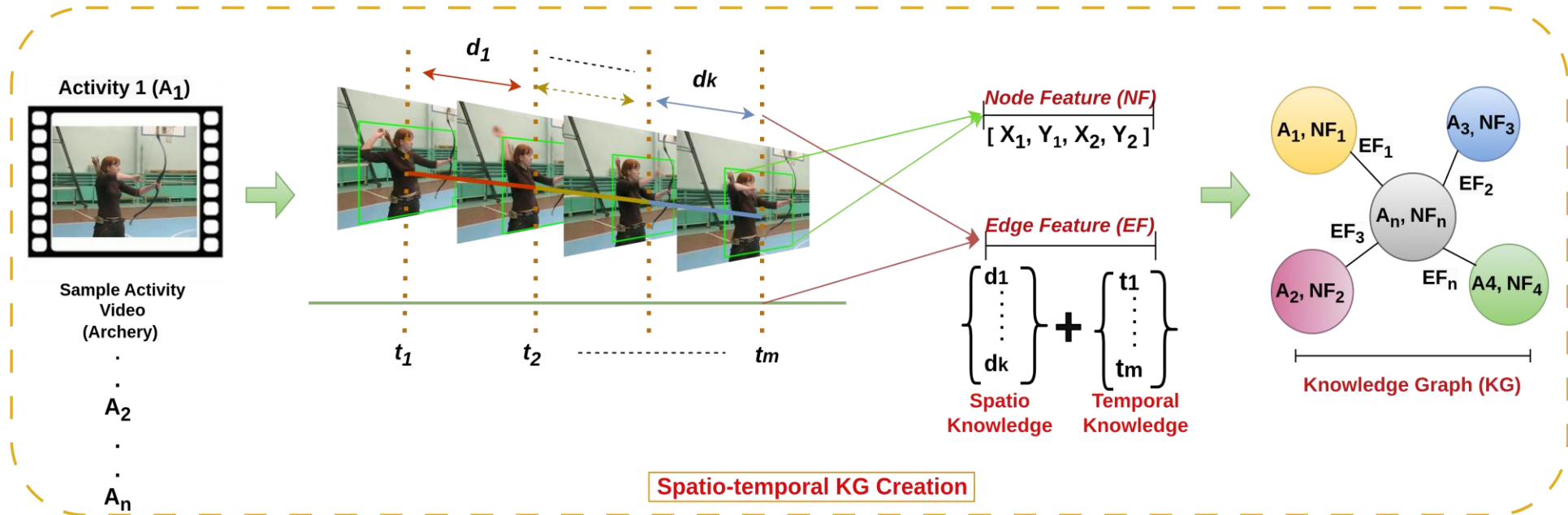
Knowledge Graph Assembly (STKG)



Reasoning & Inference (STKGNN)

- Spatio-temporal Knowledge Graph (STKG)
- Spatio-temporal Knowledge Graph Neural Network (STKGNN)

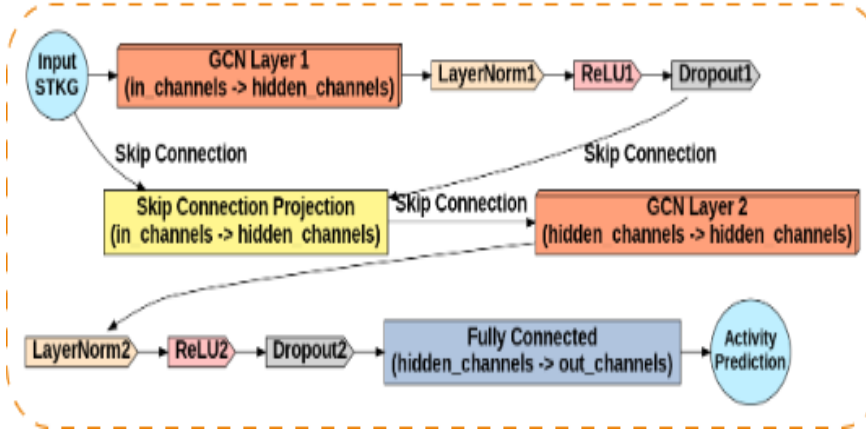
# Spatio-temporal Knowledge Graph (STKG) Construction



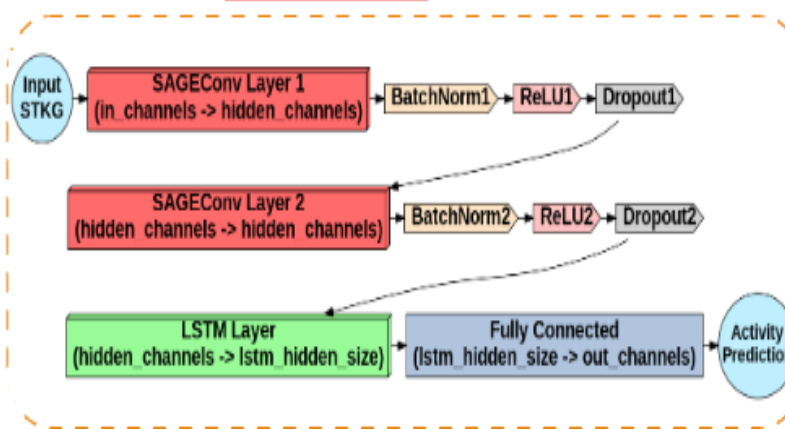
The module shows the STKG construction, where each node  $A_i$  represents an activity element with node features (NF) derived from detected object coordinates, and edges (EF) represent spatio-temporal relationships based on spatial distance  $dk$  and temporal intervals  $tn$ .

# Proposed GNN Models

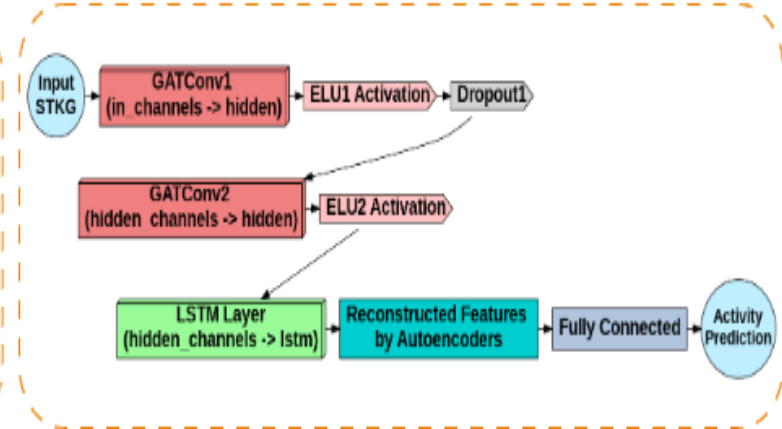
**StableGCN**



**TemporalSAGE**







**FusionGAT**



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



# Video-Based Benchmark Datasets Used for Deriving STKG

<u>Dataset</u>	<u>Temporal Coverage</u>	<u>Scene / Source Characteristics</u>	<u>Structural Complexity</u>	<u>Representative Scenario</u>
<b>HMDB-STKG</b>	Short clips (2–5s)	Single camera, isolated human actions	Low	Single-action recognition 
<b>UCF-STKG</b>	Short sequences (5–10s)	Multiple short scenes, limited transitions	Low–Middle	Scene-level activity detection 
<b>Kinetics-STKG</b>	Long sequences (10–30s)	Diverse video sources, multiple concurrent actions	Middle	Cross-scene event reasoning 
<b>MS-STKG (Small / Medium / Large)</b>	Multi-temporal (short → long)	Fused from HMDB, UCF, and Kinetics; heterogeneous nodes and relations	High → Very High	Multi-source fusion Benchmark 

**NOTE:** Each dataset contains 15 activity classes and 120 videos. The three MS-STKG variants are constructed by combining samples from HMDB, UCF, and Kinetics:

• **Small:** 15 classes, 60 videos • **Medium:** 15 classes, 120 videos • **Large:** 30 classes, 150 videos

# Performance of Models Various Scaled STKGs and Reasoning Settings

DATASET	StableGCN Inductive $\pm$ Std	StableGCN Transductive $\pm$ Std	TemporalSAGE Inductive $\pm$ Std	TemporalSAGE Transductive $\pm$ Std	FusionGAT Inductive $\pm$ Std	FusionGAT Transductive $\pm$ Std
HMDB-STKG	83.18 $\pm$ 0.63	94.09 $\pm$ 0.18	86.50 $\pm$ 0.27	95.17 $\pm$ 0.15	87.34 $\pm$ 0.24	98.10 $\pm$ 0.12
UCF-STKG	80.87 $\pm$ 0.24	90.73 $\pm$ 0.12	82.66 $\pm$ 0.21	93.09 $\pm$ 0.10	84.60 $\pm$ 0.44	96.73 $\pm$ 0.16
Kinetics-STKG	70.77 $\pm$ 0.98	82.49 $\pm$ 0.27	70.98 $\pm$ 0.44	86.67 $\pm$ 0.11	74.37 $\pm$ 0.64	92.38 $\pm$ 0.17
MS-STKG-Medium	80.71 $\pm$ 0.34	90.91 $\pm$ 0.15	81.88 $\pm$ 0.39	93.10 $\pm$ 0.09	84.15 $\pm$ 0.68	96.40 $\pm$ 0.18
MS-STKG-Large	75.90 $\pm$ 0.17	85.55 $\pm$ 0.21	78.48 $\pm$ 0.39	89.43 $\pm$ 0.20	81.67 $\pm$ 0.23	95.07 $\pm$ 0.21
MS-STKG-Small	86.79 $\pm$ 0.49	95.47 $\pm$ 0.21	87.83 $\pm$ 0.46	96.48 $\pm$ 0.19	88.45 $\pm$ 0.72	97.79 $\pm$ 0.23

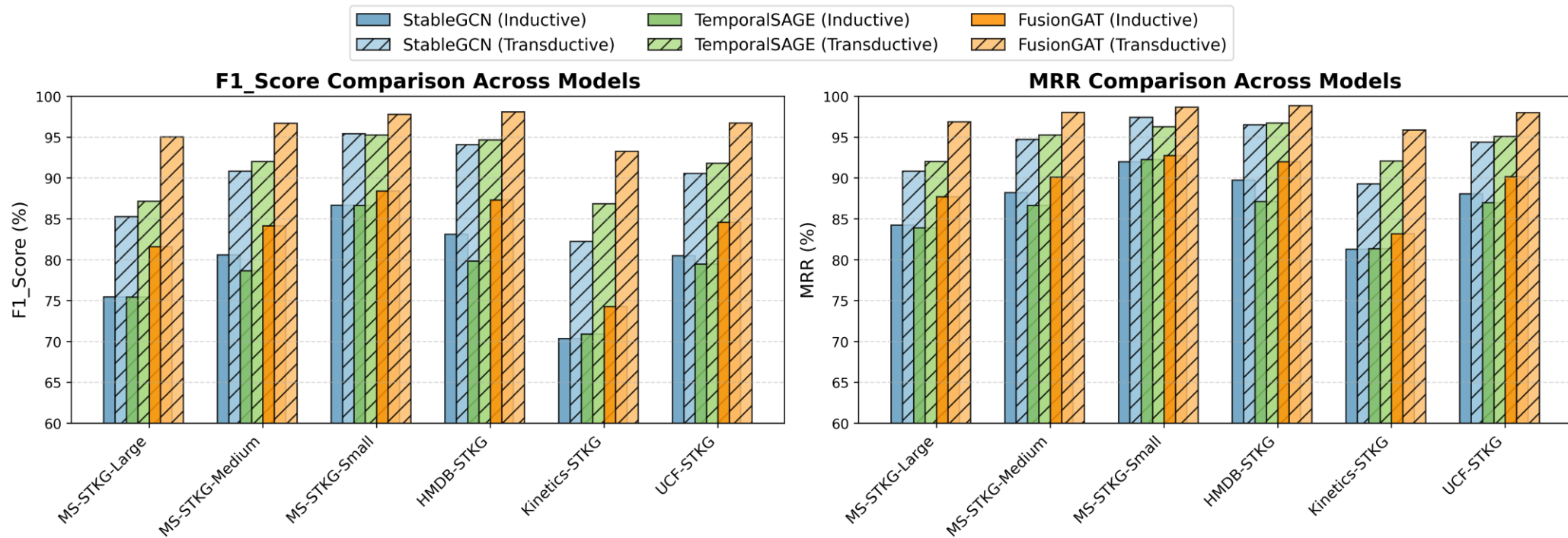
Provides a static graph reasoning baseline

Learns how structural patterns change over time

Combines spatial, temporal, and semantic signals for unified reasoning

**Inductive Reasoning:** Generalizing to unseen nodes or scenes    **Transductive Reasoning:** Completing relations within known graph

# Cross-Metric Validation



**F1 and MRR confirm stable reasoning patterns cross our models over multi-scale STKGs beyond accuracy**

# Baseline Comparison

MODELS	Top-1 Accuracy (%)
FusionGAT(Ours)	85.07
TemporalSAGE(Ours)	80.30
StableGCN(Ours)	79.38
STIP-GCN *	72.07
TRG **	70.51
AKU ***	67.42

**STIP-GCN \*** : Capture local motion cues but lacking semantic temporal reasoning.

**TRG \*\*** : Partially models temporal relations, but misses spatial semantics.

**AKU \*\*\*** : Multi-modal early fusion approach, yet without explicit reasoning.

## Comparison of our proposed models and closely related GNN based baselines on the MS-STKG-Medium dataset

\* 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.

\*\* 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.

\*\*\* 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.

# Key Contributions

---

Scalable framework (STKGNN) for structured spatio-temporal reasoning

---

Adaptive STKG construction pipeline capturing spatial, temporal, and semantic dependencies

---

Hierarchical graph-level three novel models which provide generalizable temporal and relational reasoning

---

Comprehensive evaluation under both inductive and transductive settings across heterogeneous data

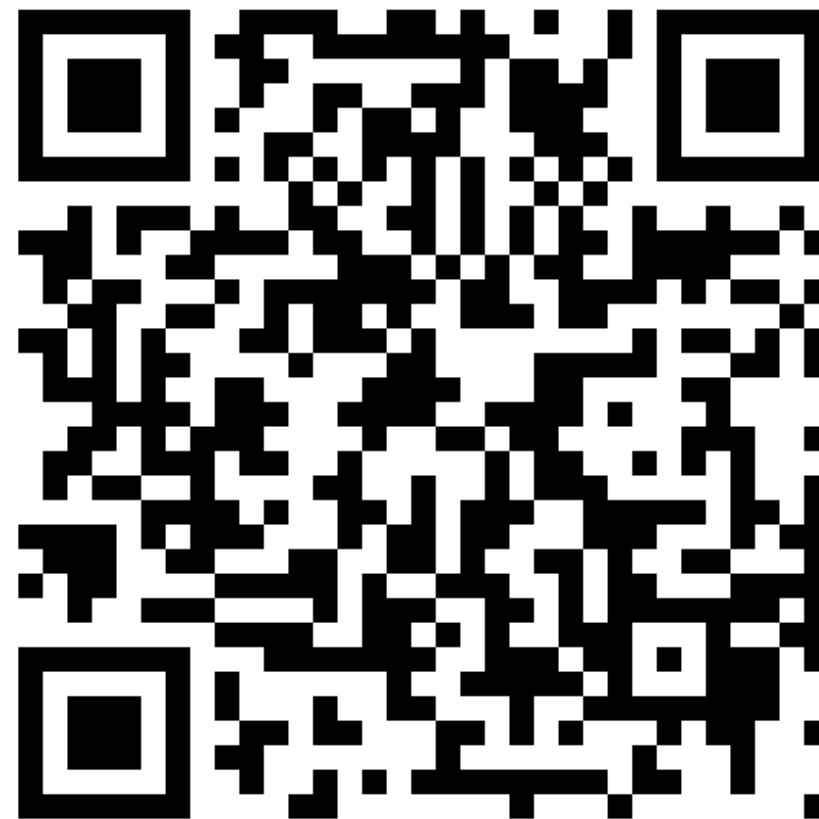
# Future Work

Extend the framework toward real-time and cross-domain reasoning across multiple video sources

Conduct computational efficiency and transparency analysis to assess scalability and real-time potential

**Thank you!**

**Please scan the QR Code for  
all Reproducible Source**



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

Gözde Ayşe Tataroğlu Özbulak, Yash Raj Shrestha, Jean-Paul Calbimonte

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