香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB

DSA Research Experiences for Undergraduates

Research Project

Section1: Faculty Information

Full Name	Yuxuan Liang	Tel	(020) 8833 5733
Thrust/Hub	System Hub / Thrust of Intelligent Transportation Info Hub / Thrust of Data Science and Analytics	Office	W2 L5 511
Email	yuxuanliang@hkust-gz.edu.cn		

Section2: Research Project Proposal

Project Title	Exploring Vision-Language Models for Spatio-Temporal Data Prediction
Project Description (max 800 words)	**1. Background** Spatio-temporal prediction, such as traffic flow forecasting and weather modeling, involves capturing complex interactions between temporal dynamics and spatial structures. While existing methods have made significant progress, we propose a novel direction that leverages the power of Vision-Language Models (VLMs) to enhance spatio-temporal forecasting. Our approach transforms raw spatio-temporal data into image-like frames, enabling the use of pre-trained VLMs (e.g., CLIP, Flamingo) to integrate visual patterns (e.g., traffic heatmaps) and semantic context (e.g., event descriptions) for more accurate and interpretable predictions. By bridging the gap between spatio-temporal data and multimodal learning, this project aims to unlock new possibilities for urban dynamics modeling and beyond. **2. Objectives** This project aims to: - Develop a spatio-temporal-to-visual mapping framework to convert raw spatio-temporal data (e.g., traffic grids) into image-like frames. - Design a multimodal fusion architecture integrating pre-trained VLMs for joint visual-textual reasoning. - Validate the framework on public datasets (e.g., PeMS traffic data) to achieve state-of-the-art prediction accuracy with interpretable outputs. **3. Methodology** Step 1: Spatio-Temporal Visual Encoding - Convert spatio-temporal tensors (time × height × width × channels) into multi-channel image sequences. (Example: Map traffic flow grids to RGB images where channels encode speed, density, and direction.)

	- Use video pretrained models (e.g., TimeSformer) to extract spatio-temporal features. Step 2: Textual Context Injection - Generate text prompts from metadata (e.g., "Weekday morning rush hour with light rain"). - Align visual features and text embeddings via CLIP-style contrastive learning. Step 3: Multimodal Fusion - Combine visual features, text embeddings, and raw temporal signals using gated cross-attention. - Train a lightweight predictor (e.g., MLP or 1D-CNN) to generate future frames. **4. Innovation** - First integration of VLMs for spatio-temporal prediction, leveraging both visual priors and semantic context. - Interpretable outputs: Visualize attention maps to explain how text prompts (e.g., "accident on Highway 101") influence predictions. **5. Expected Outcomes** - A lightweight library for spatio-temporal visual-textual data conversion. - A benchmark comparison against ST-GCN, ConvLSTM, and ST-Transformer on traffic datasets. - A research paper (targeting CIKM or KDD workshops) and open-source code.
Proposed Research Duration	Start Date: _1__ / _June___ / 2025___ End Date: __1__ / _December_ / ___2025_
Student/Researcher Duties	Phase 1: Data Preprocessing & Visualization - Convert raw spatio-temporal data (e.g., CSV files) into image frames using Python/OpenCV. - Annotate text prompts based on metadata (e.g., weather, holidays). Phase 2: Model Implementation - Fine-tune pre-trained VLMs (e.g., CLIP-ViT) on spatio-temporal image sequences. - Implement fusion modules (e.g., cross-attention) using PyTorch. Phase 3: Evaluation & Visualization - Compare prediction accuracy (MAE/RMSE) against baselines. - Generate Grad-CAM visualizations to interpret model decisions.

	Phase 4: Paper writing & Deployment - Write a paper following the standard format of conference. - Develop and deploy a visual interactive system.
Technical Skills Required	☑Python ☑Machine Learning ☑Big Data ☐ R ☑ Deep Learning ☐ SQL ☐ C/C++ ☐ Other: _____
Preferred Student/Researcher Background	Academic: Undergraduate students in Computer Science, Data Science, or GIS-related fields. Experience: - Coursework or projects in machine learning (especially CV/NLP). - Hands-on experience with PyTorch. Bonus: - Prior exposure to spatio-temporal prediction (e.g., traffic forecasting). - Familiarity with linux system for large-scale training.
Maximum Number of Students/Researchers	☐ 1 ☑ 2

Section3: Pre-Application Research Exposure Meeting

Faculty members are encouraged to schedule a Research Exposure Meeting to introduce students to their projects.

Preferred Date	Each Friday
Preferred Time	14:00-15:00
Meeting Mode	☑In-Person ☐ Online
Venue (if in-person)	W2-511
Meeting Link (if online)	