OUTLINE

Introduction

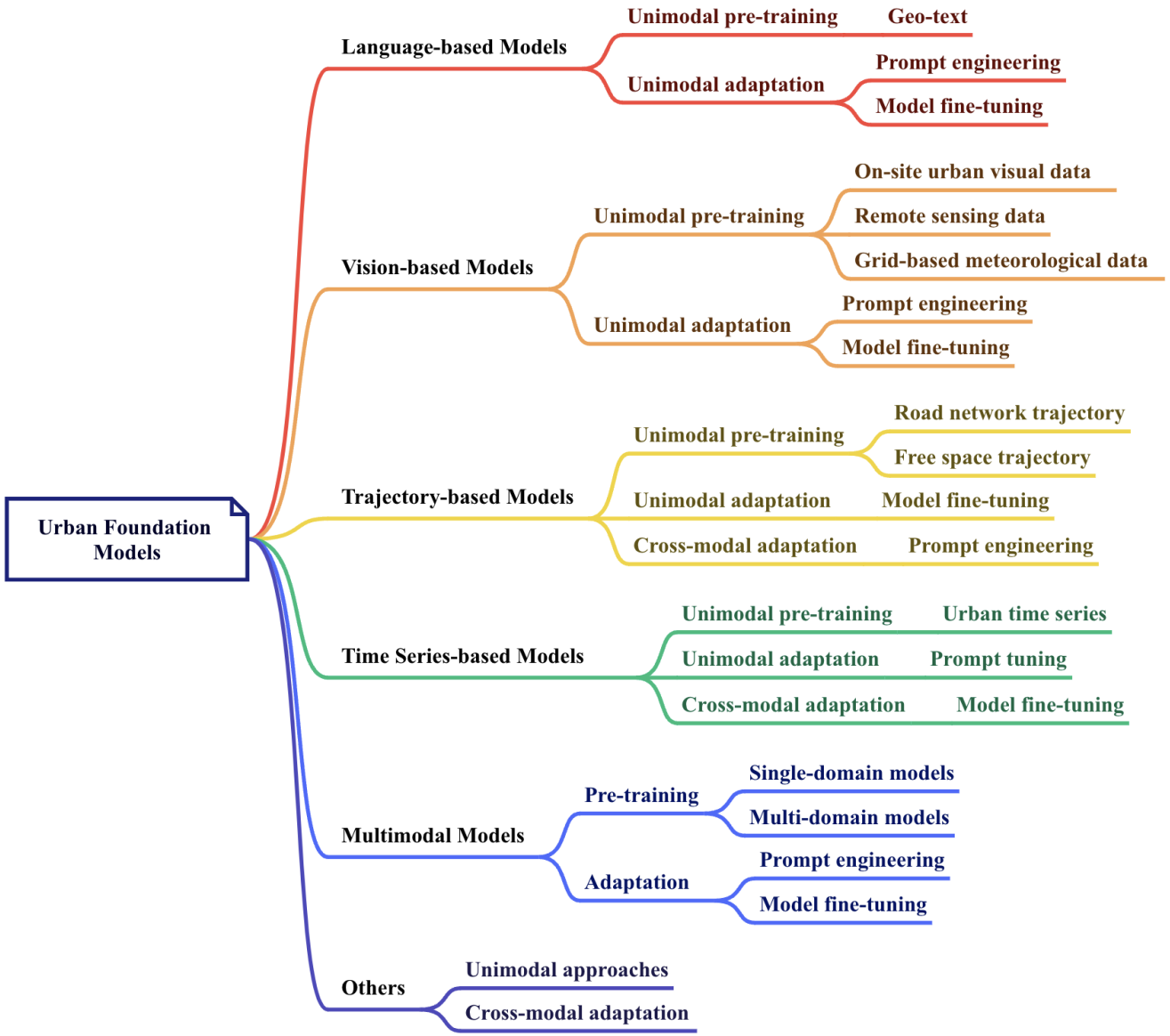
Challenges of Building UFMs

Overview of UFMs

Prospects of UFMs

Summary

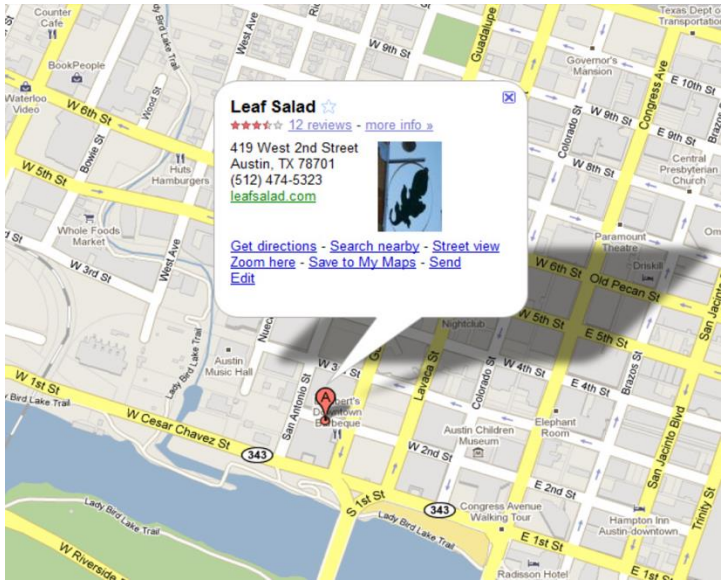
A Data-Centric Taxonomy of UFMs



Language-based UFMs

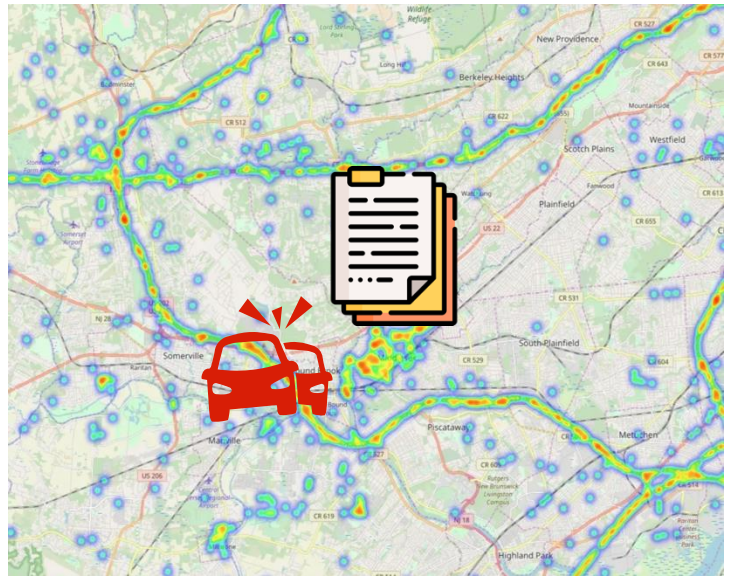
Text Data in Cities

- Unique properties of city-related textual data
  - Associated with specific locations, e.g., landmarks, GPS coordinates
  - Exhibit time-dependent nature, e.g., reports of city events



Geo-tagged POI reviews

Ride-hailing dialogs

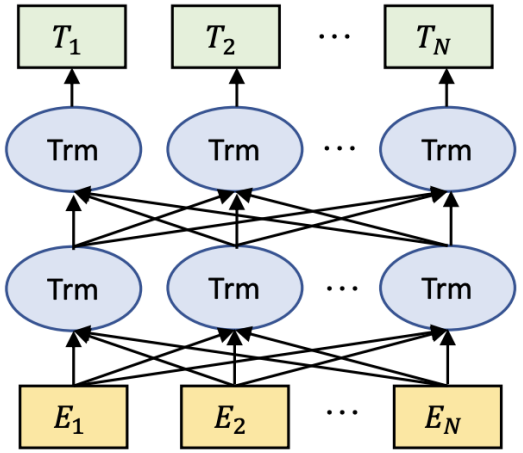


Traffic accident reports

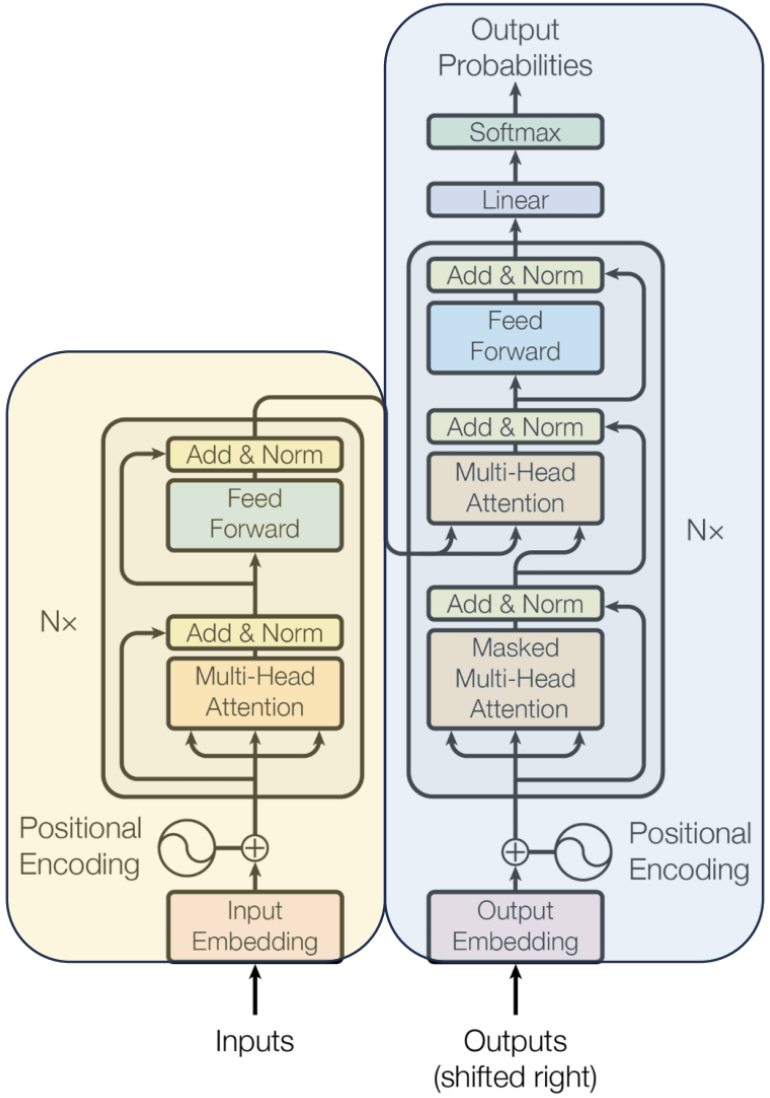
It is critical to understand the spatio-temporal dynamics when handling urban textual data.

Pre-trained Language Models (PLMs)

BERT (bidirectional)
Oct 2018
Masked word prediction

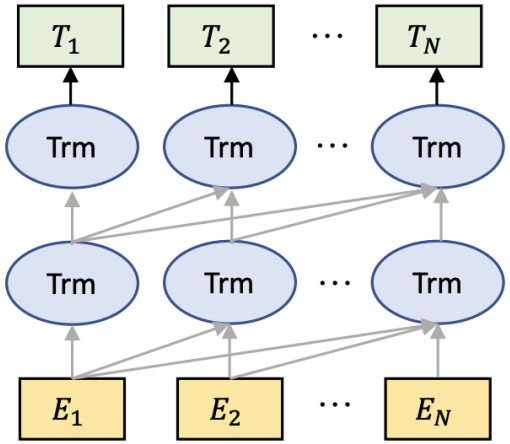


Google



Transformer Architecture

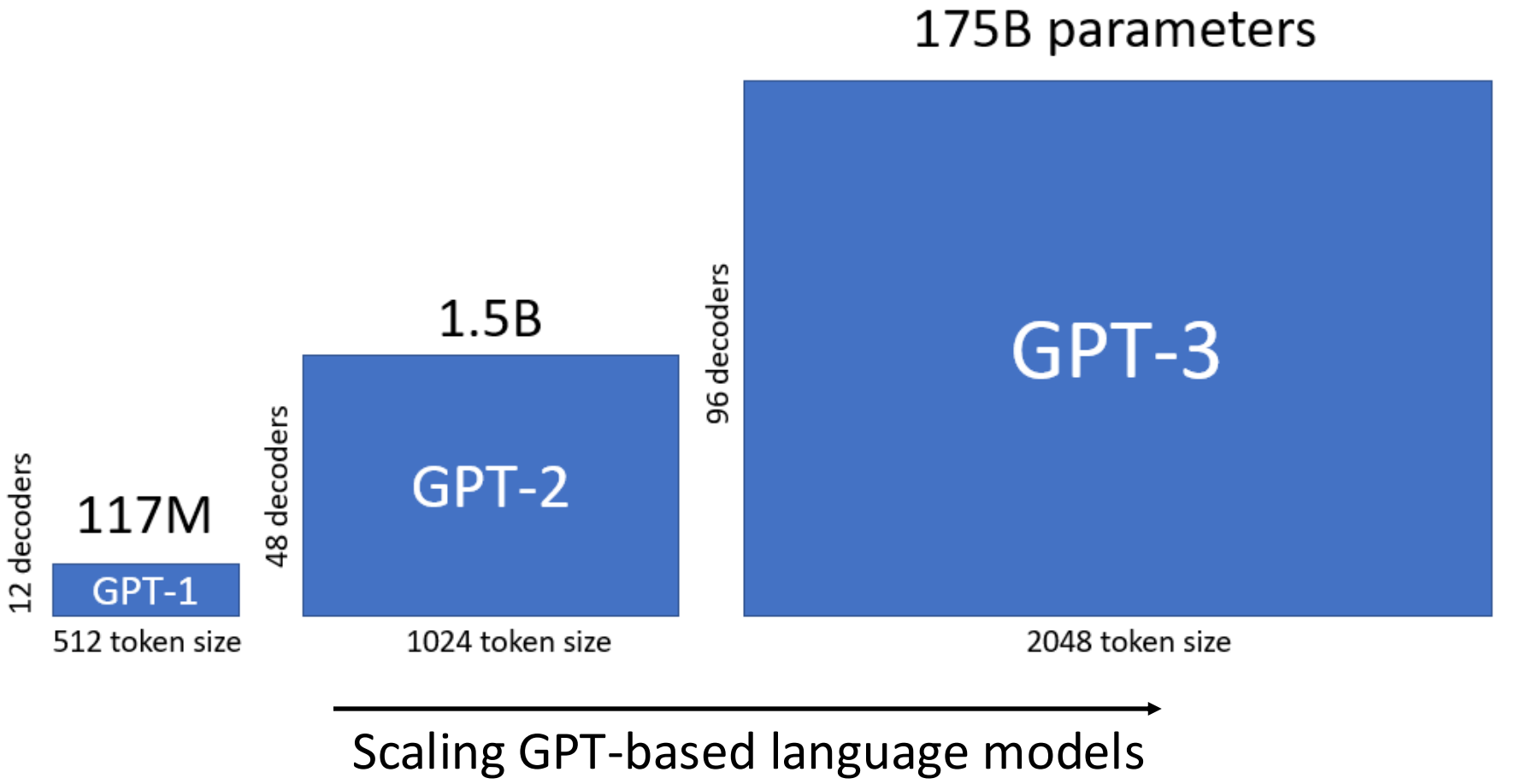
GPT (unidirectional)
Jun 2018
Next word prediction



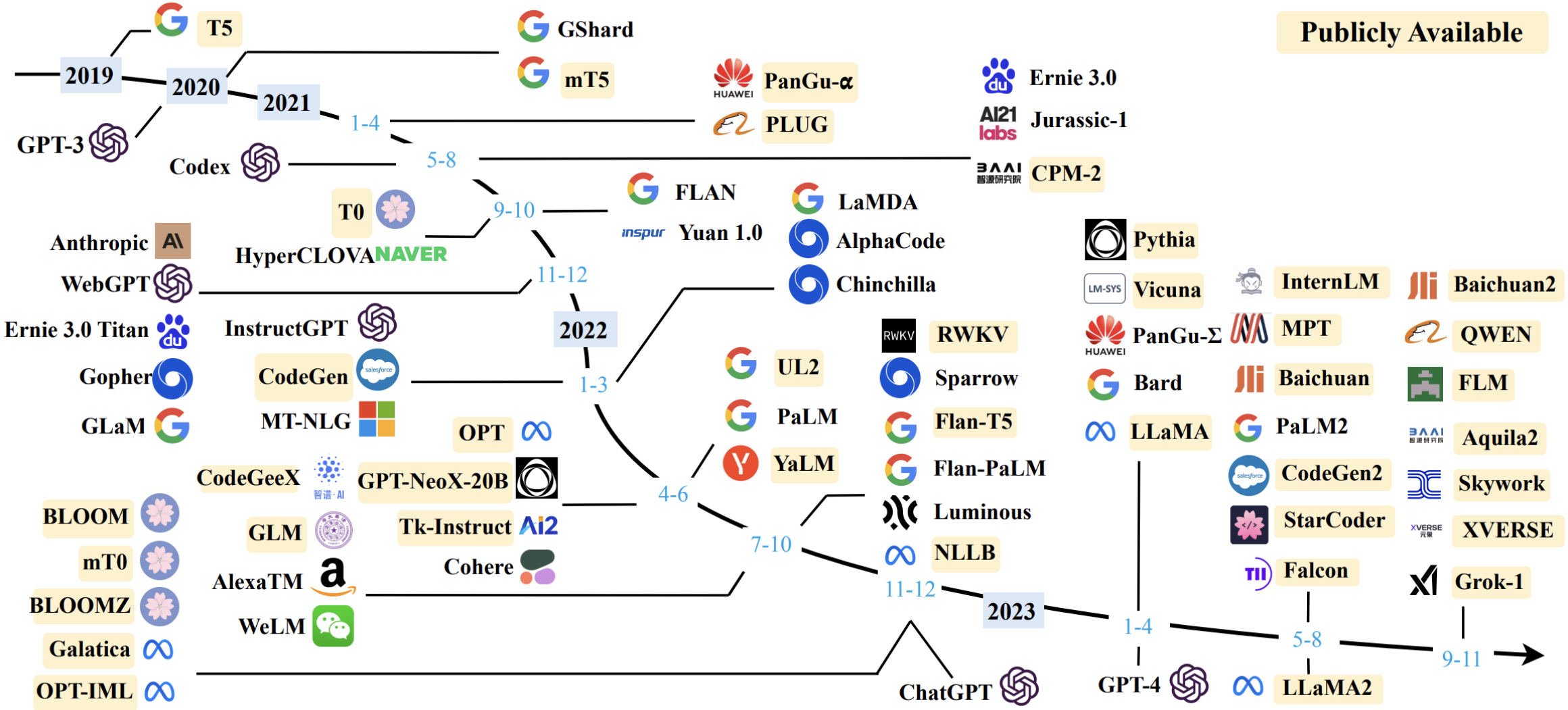
OpenAI

Large Language Models (LLMs)

- Scaling GPT-based models leads to impressive capabilities
  - GPT-1/2: pre-training+fine-tuning → solve various NLP tasks
  - GPT-3: scaling model size→ solve real-world tasks using prompts



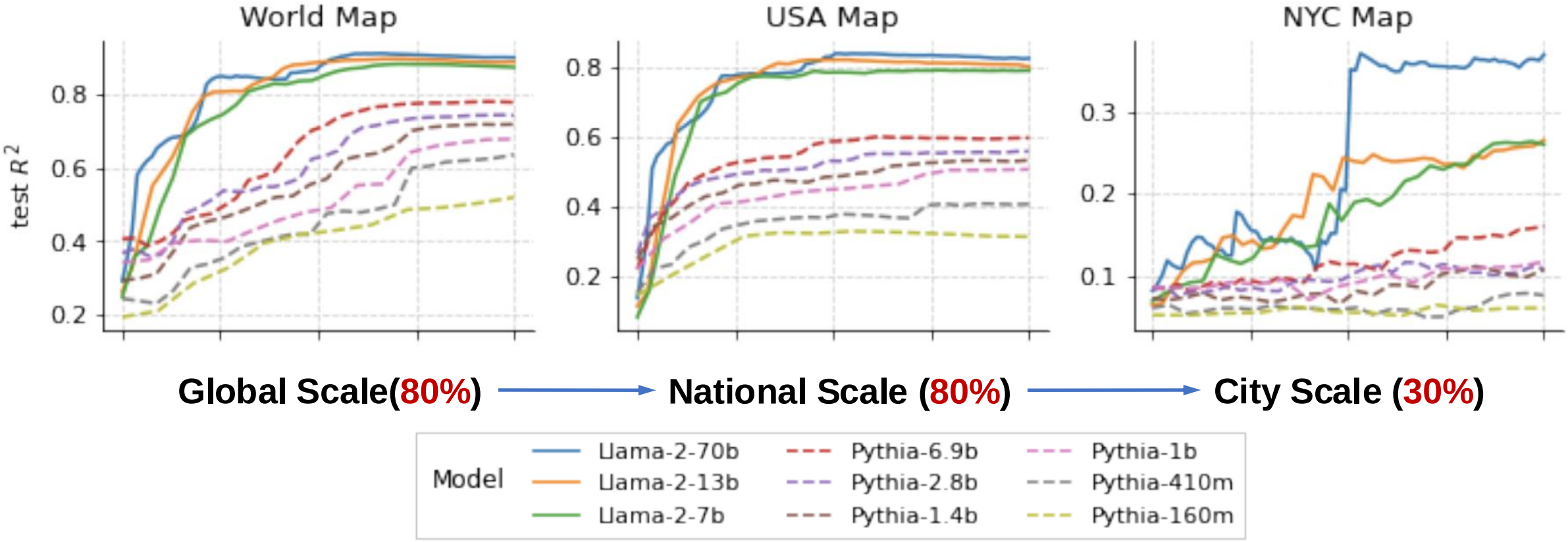
Large Language Models (LLMs)



Zhao, Wayne Xin, et al. A Survey of Large Language Models. arXiv 2023.

Why Need Language-based UFMs?

- The accuracy of POI location prediction for pre-trained language models



- Observation: language models often lack real-world urban knowledge, e.g., spatial entities within a city

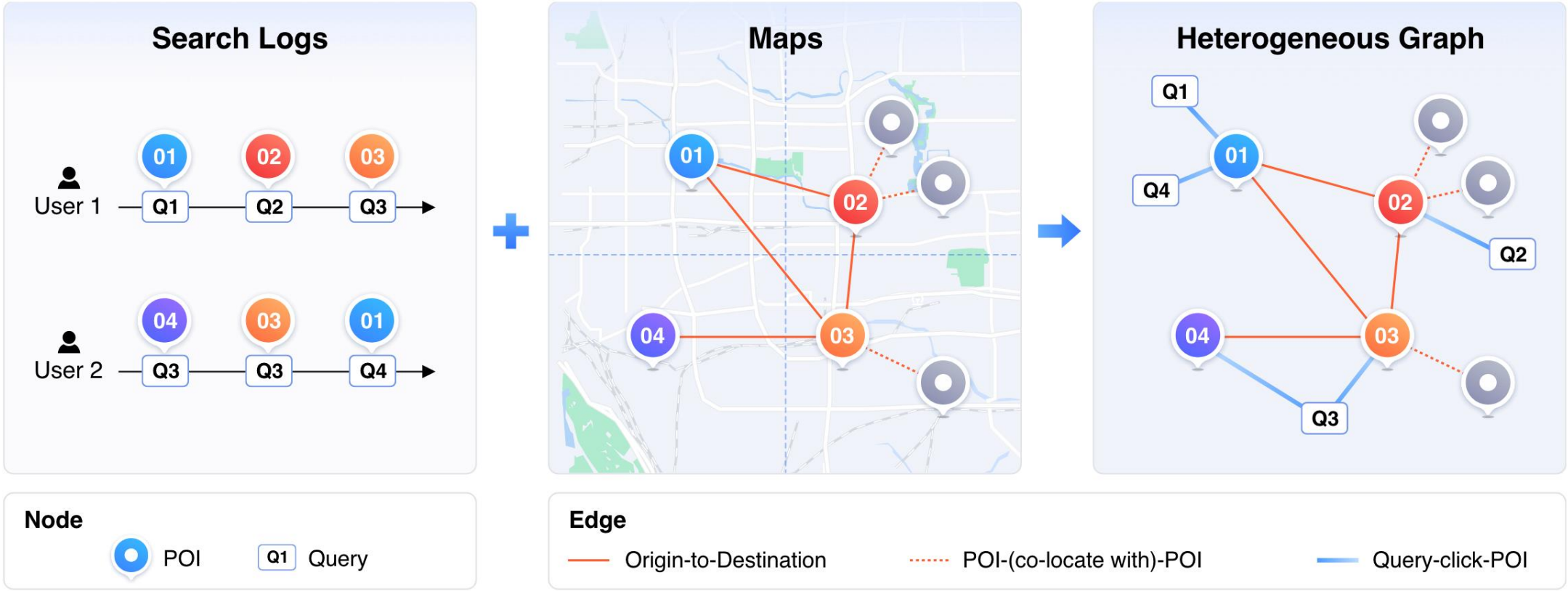
Gurnee Wes, et al. Language Models Represent Space and Time. ICLR 2024.

Language-based UFMs

- **Language-based Pre-training**
  - *Geo-text*
- **Language-based Adaptation**
  - Prompt engineering
  - Model fine-tuning

Pre-training on Geo-text

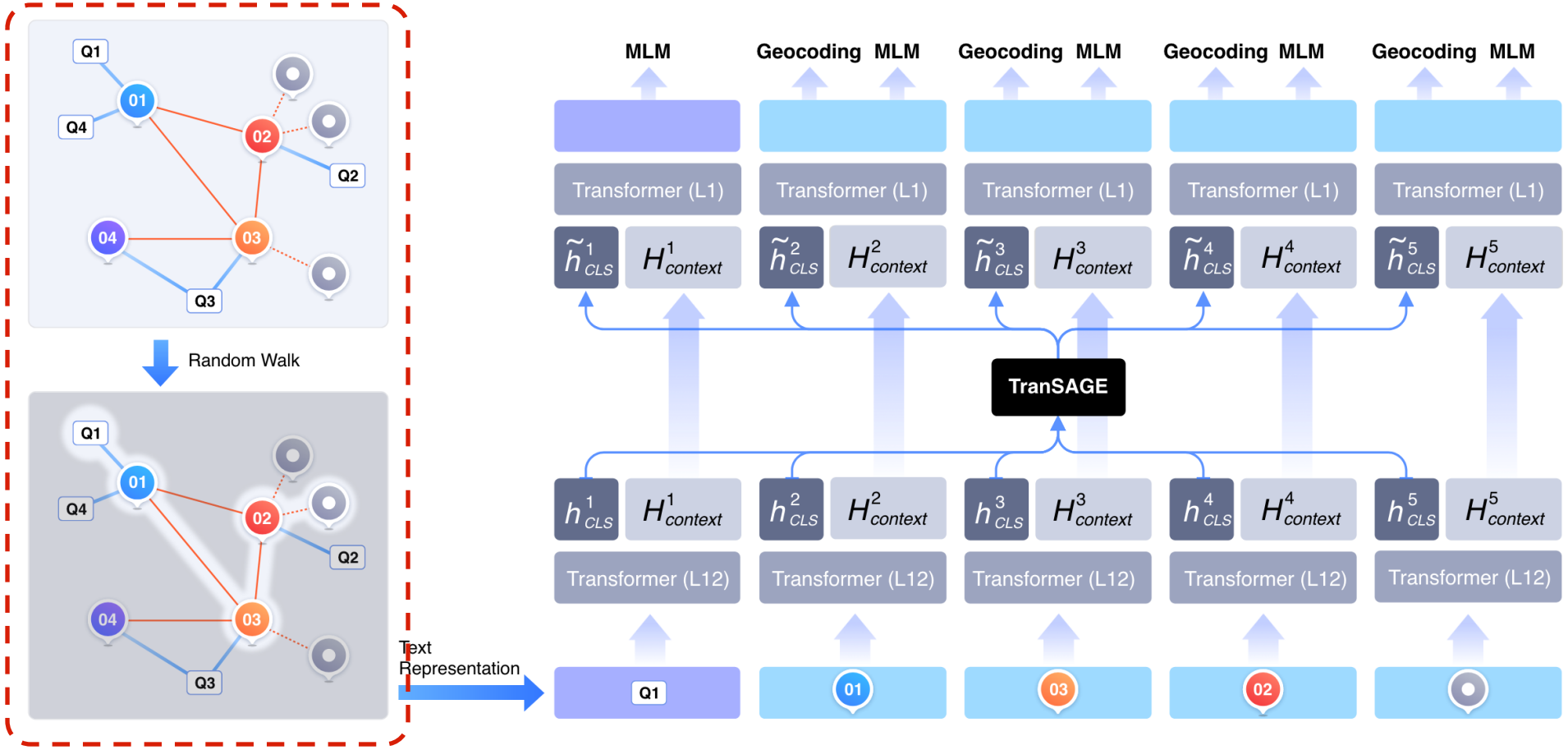
- ERNIE-GeoL:
  - Step 1: construct a heterogenous graph that contains POI and user query nodes, based on search logs in online maps



Huang, Jizhou, et al. A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. KDD 2022.

Pre-training on Geo-text

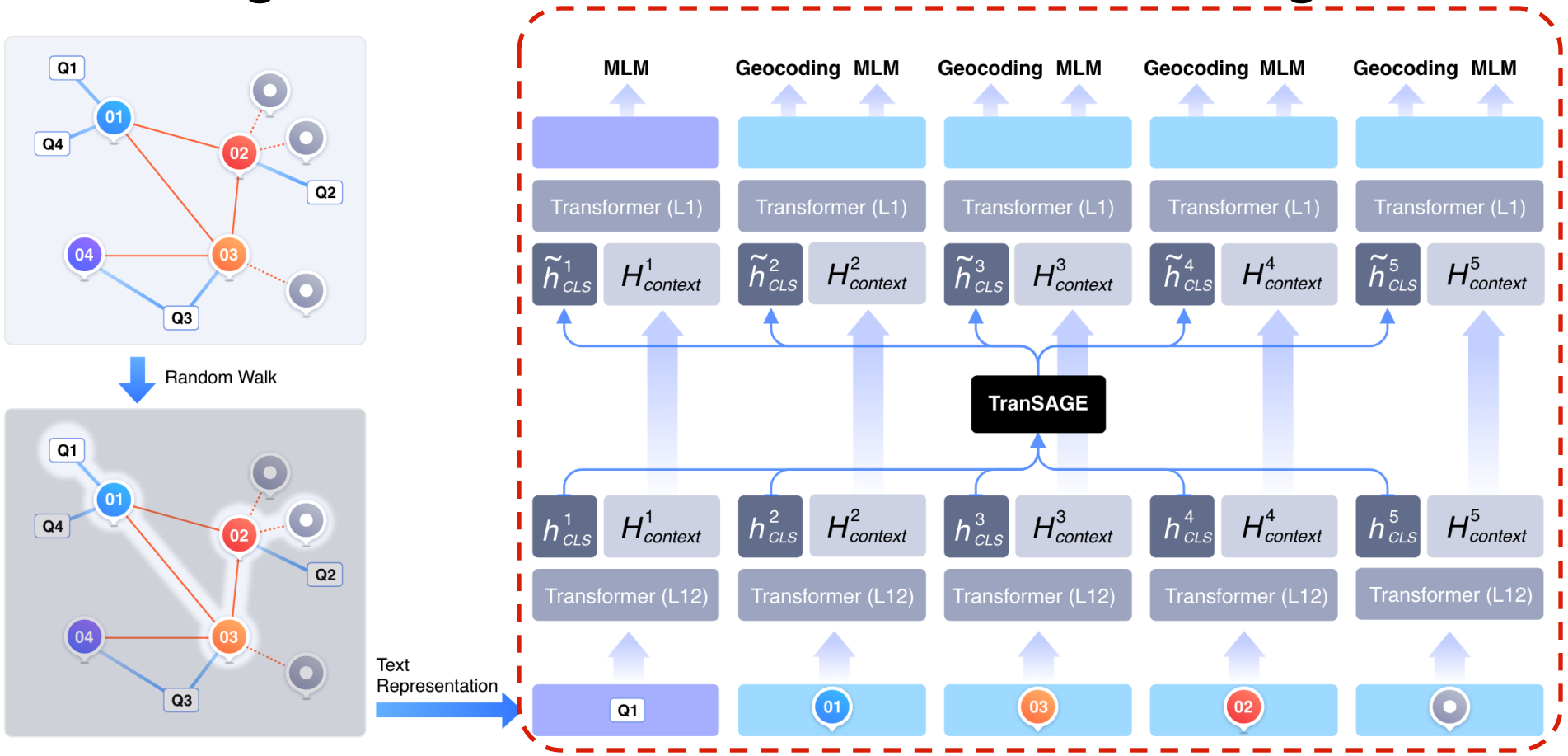
- ERNIE-GeoL
  - Step 2: create a pre-training dataset by performing random walk sampling on the heterogeneous graph



Huang, Jizhou, et al. A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. KDD 2022.

Pre-training on Geo-text

- ERNIE-GeoL
  - Step 3: pre-train a Transformer via masked word prediction and geo-coding
    - Geo-coding: translate human-readable address into geo-coordinate



Huang, Jizhou, et al. A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. KDD 2022.

Language-based UFMs

- **Language-based Pre-training**
  - Geo-text
- **Language-based Adaptation**
  - *Prompt engineering*
  - Model fine-tuning

Prompt Engineering

- Model pre-training
  - Pros: can memorize extensive urban knowledge
  - Cons: resource-intensive
- Prompt engineering
  - Using task description or few-shot examples to steer generalist LLMs
  - Flexible, no additional training required

Task description

Question: Below is the coordinate information and related comments of a point of interest: · · · .
Please answer the category of this point of interest.
Options: (1) xxxx, (2) xxxx, (3) xxxx, · · · .
Please answer one option.
Answer: The answer is option (

Zero-shot prompting

Li, Wenbin, et al. "STBench: Assessing the Ability of Large Language Models in Spatio-Temporal Analysis." arXiv 2024.

Prompt Engineering

- In-context prompting: few-shot

Given the coordinate information and related comments of a point of interest: 41.40338, 2.17403, Sagrada Familia, Barcelona. Please answer the category of this point of interest.
Options: (1) Museum, (2) Church, (3) Park, (4) Restaurant
Please answer one option.
Answer: The answer is option (2) Church.

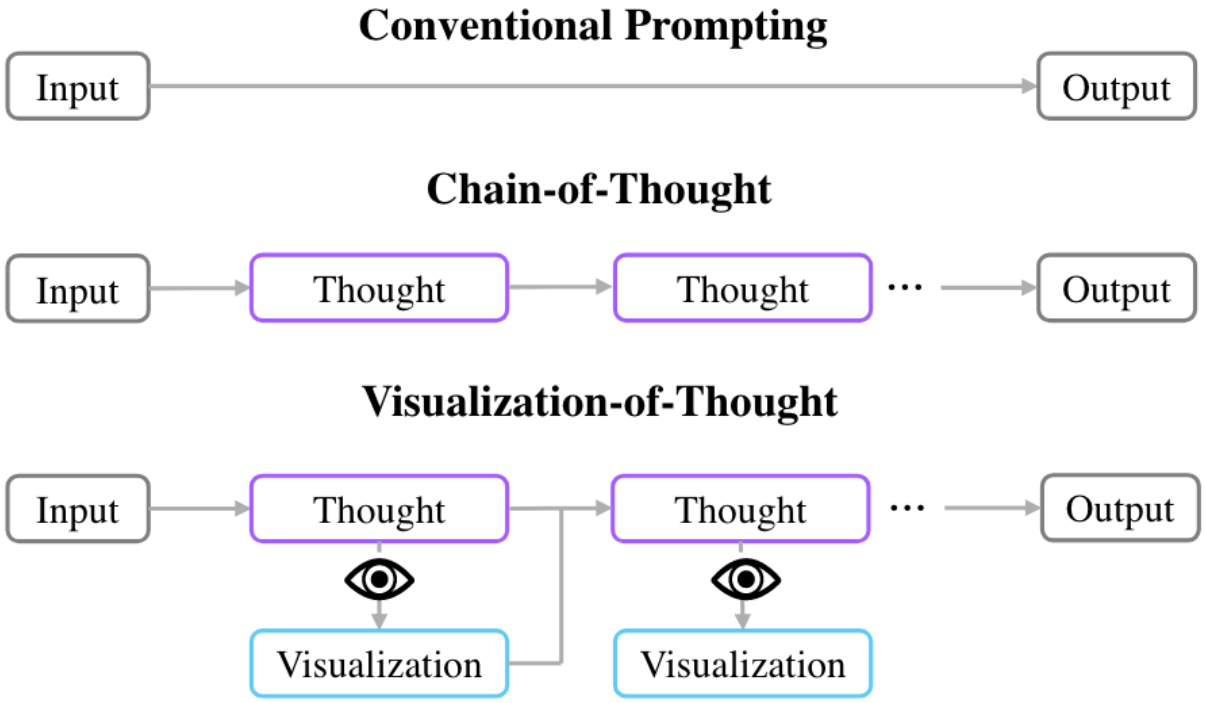
Given the coordinate information and related comments of a point of interest: 51.5074, -0.1278, Big Ben, London. Please answer the category of this point of interest.
Options: (1) Historical Monument, (2) Theater, (3) Shopping Mall, (4) School
Please answer one option.
Answer: The answer is option (1) Historical Monument.

Few-shot examples help language models familiarize response style

Question: Below is the coordinate information and related comments of a point of interest: · · · .
Please answer the category of this point of interest.
Options: (1) xxxx, (2) xxxx, (3) xxxx, · · · .
Please answer one option.
Answer: The answer is option (

Prompt Engineering

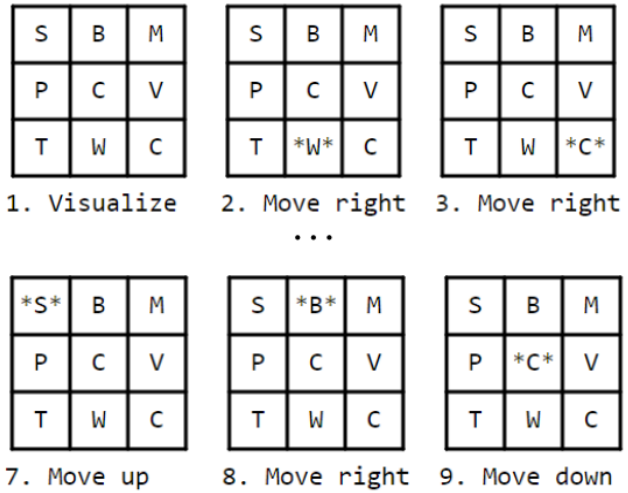
- Chain-of-Thought (CoT) prompting
  - Prompt LLMs to output a series of intermediate reasoning steps, i.e., thoughts
- Visualization-of-thought for spatial reasoning



Natural Language Navigation

You have been given a 3 by 3 square grid. Initially, you are at the bottom-left corner…find a cassette player…go right…a wool, go right…a conch, go up…a moving van, go left…a confectionery store, go left…a pot pie, go up…a siamang, go right…a black-and-white colobus, go right…a minivan. Now you have all the information on the map. You start at where the cassette player is located, then you go right by one step, go right…go up…go left…go left…go up…go right…go down by one step. What will you find?

Visualize the state after each reasoning step.



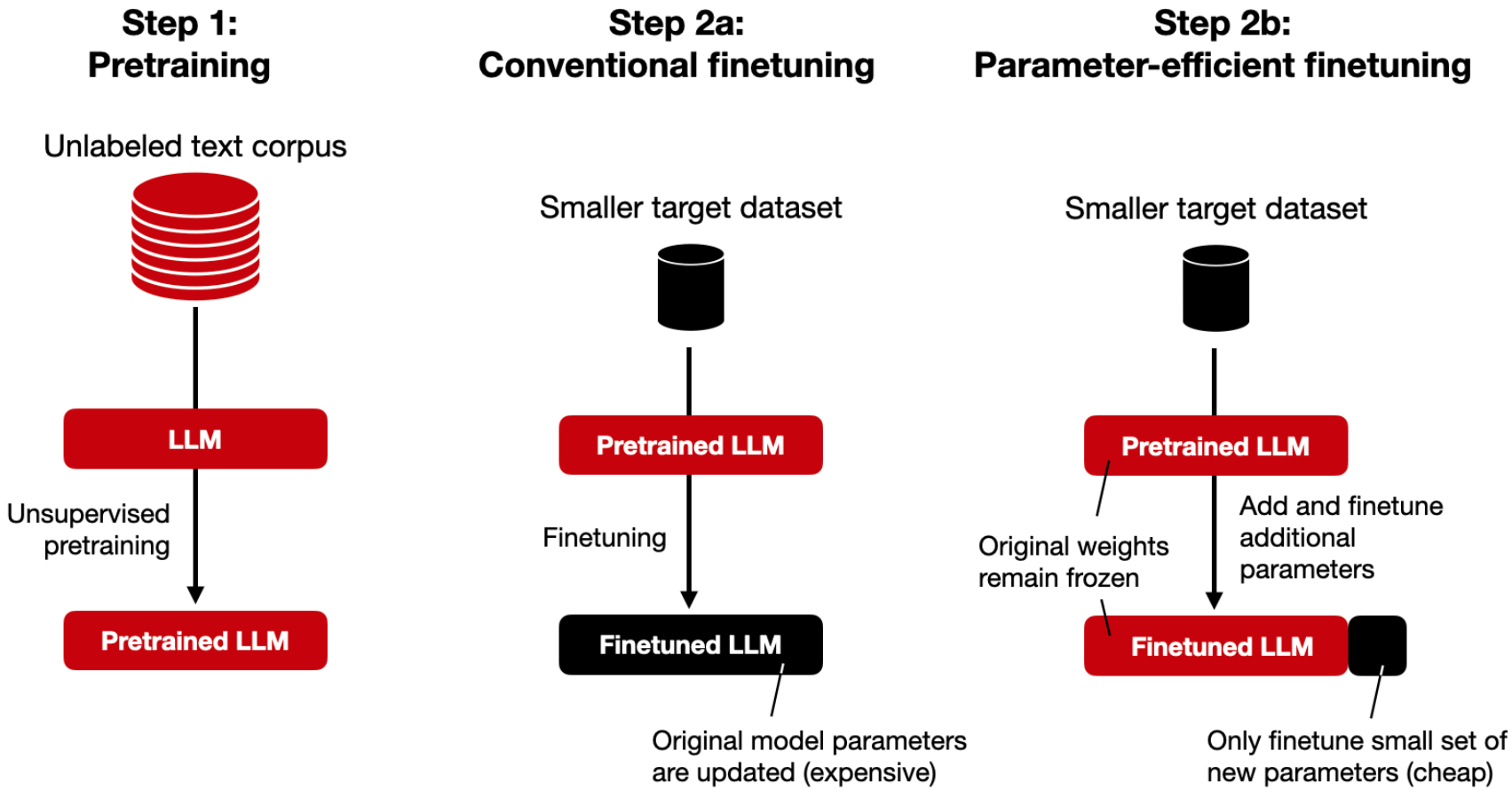
Wu, Wenshan, et al. "Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models." arXiv 2024.

Language-based UFMs

- **Language-based Pre-training**
  - Geo-text
- **Language-based Adaptation**
  - Prompt engineering
  - *Model fine-tuning*

Model Fine-tuning

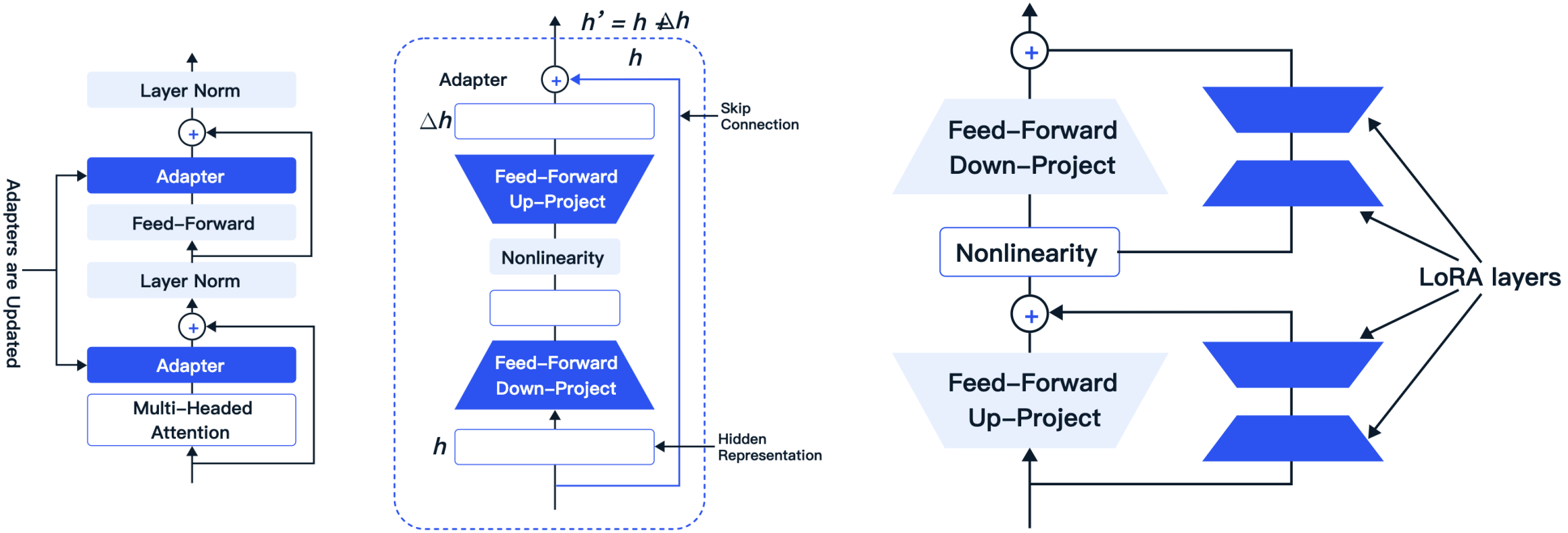
- Prompts are usually difficult to create and sensitive to small changes
- Model fine-tuning: encode and elicit specific knowledge for downstream tasks



https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters

Model Fine-tuning

- Parameter-efficient fine-tuning
  - Adapter: only learn the inserted layers rather than the entire model
  - LoRA: fine-tune additional low-rank matrices within the model



Adapter

Low-Rank Adaptation (LoRA)

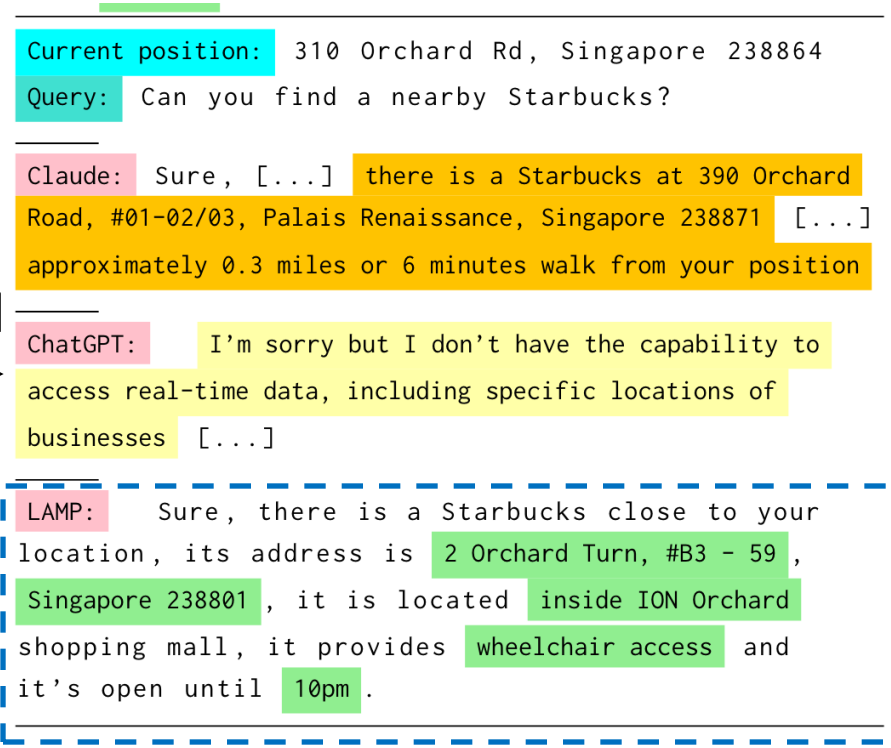
https://www.leewayhertz.com/parameter-efficient-fine-tuning/#Parameter-efficient-fine-tuning-techniques

Model Fine-tuning

- Urban knowledge memorization
  - LAMP: fine-tuned on POI search tasks, allowing it to accurately recall spatial entities within a city

Type	Query example
Name search	Hi LAMP, tell me where is {*POI_Name*} located
Category search	Please help me finding a nearby {*POI_Category*}
Type search	Can you please point out a highly rated restaurant in the area?
Type search	Can you please point out a nearby restaurant that offers {*food_type*} food?

Fine-tune LLM



Provide correct answer without hallucination

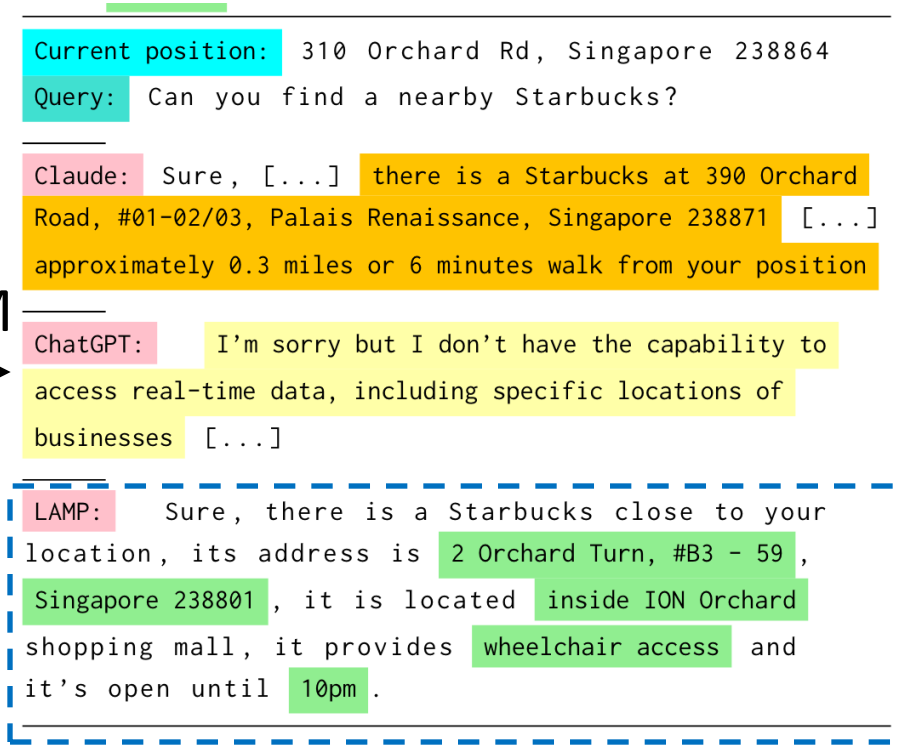
Balsebre, Pasquale et al. LAMP: A Language Model on the Map. arXiv 2024.

Model Fine-tuning

- Key steps of LAMP
  - Step 1: generated synthetic user queries and response for each POI

Type	Query example
Name search	Hi LAMP, tell me where is {*POI_Name*} located
Category search	Please help me finding a nearby {*POI_Category*}
Type search	Can you please point out a highly rated restaurant in the area?
Type search	Can you please point out a nearby restaurant that offers {*food_type*} food?

Fine-tune LLM

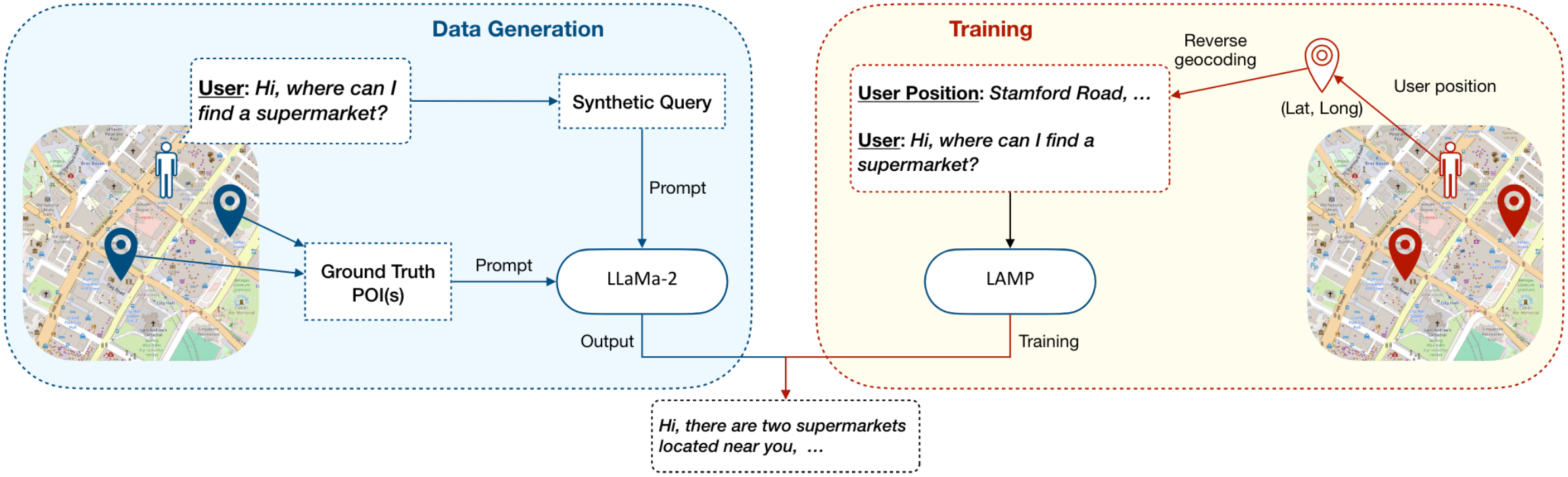


Provide correct answer without hallucination

Balsebre, Pasquale et al. LAMP: A Language Model on the Map. arXiv 2024.

Model Fine-tuning

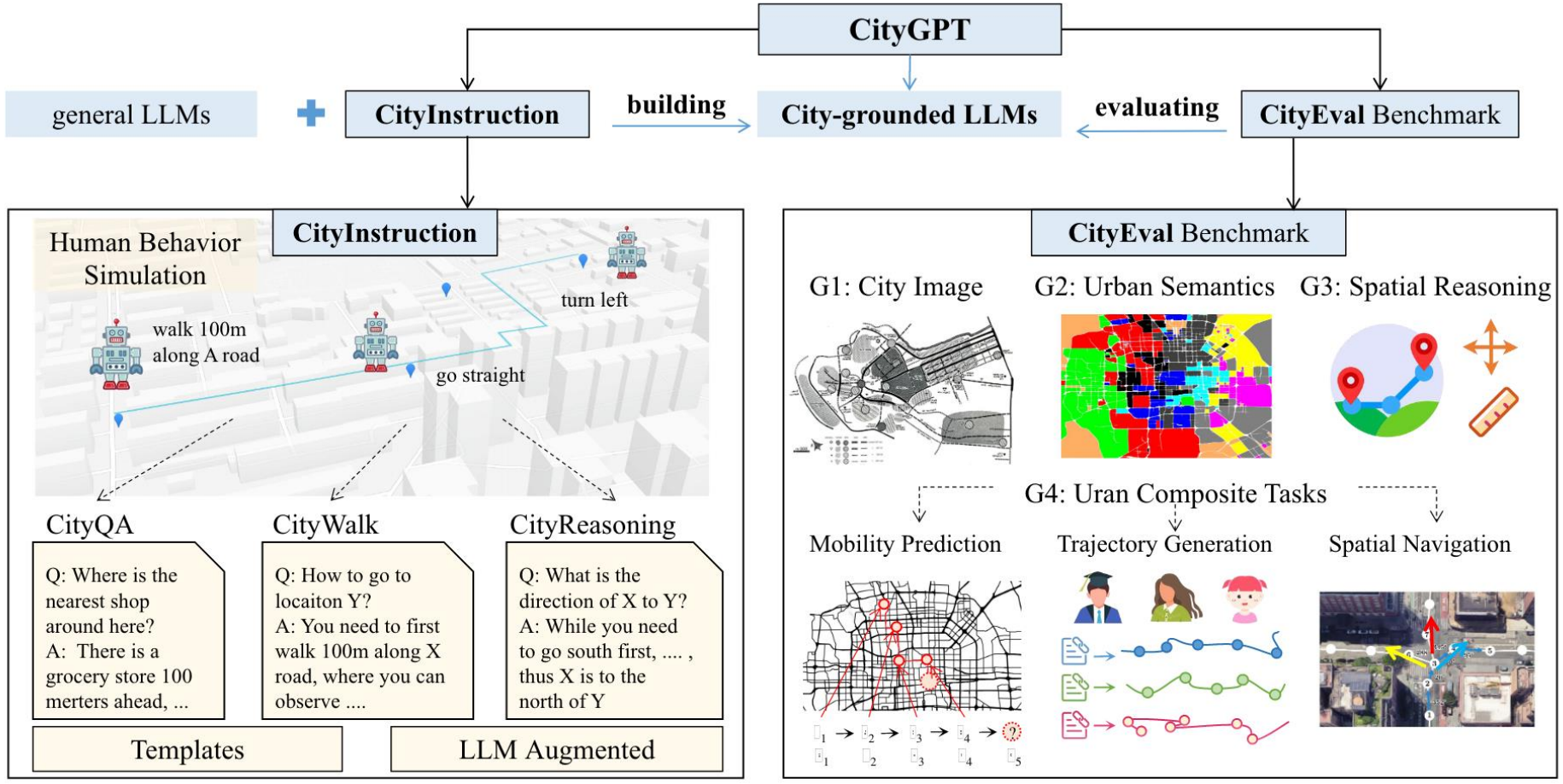
- Key steps of LAMP
  - Step 1: generated synthetic user queries and response for each POI
  - Step 2: associate each query-response pair with a user position (address)
  - Step 3: fine-tuning LLM on the generated data



Balsebre, Pasquale et al. LAMP: A Language Model on the Map. arXiv 2024.

Model Fine-tuning

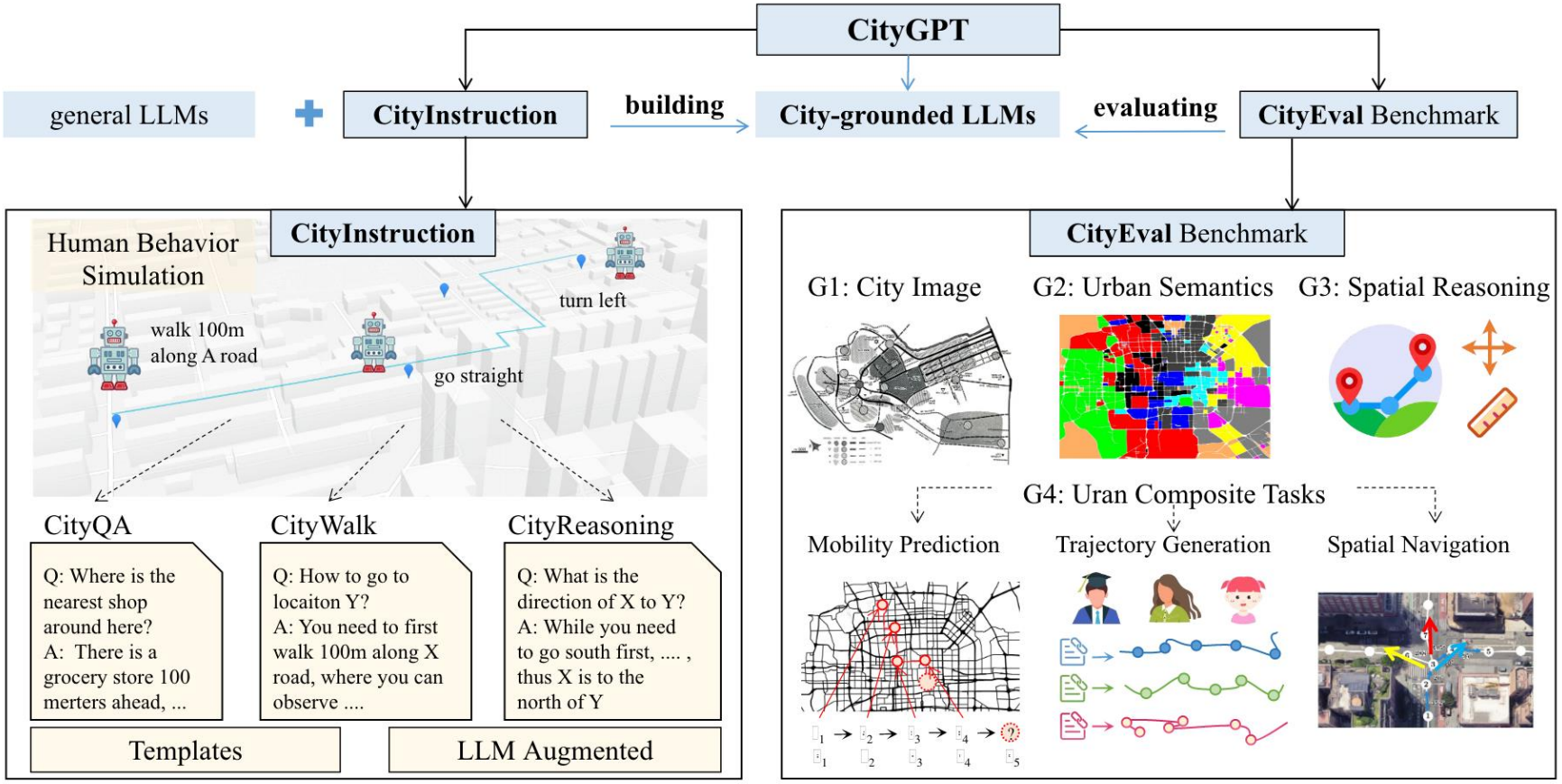
- Spatial reasoning - CityGPT
  - Step 1: randomly select two locations and obtain their navigation path



Feng, Jie, et al. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. arXiv 2024.

Model Fine-tuning

- Spatial reasoning - CityGPT
  - Step 2: generate the explicit reasoning steps from navigation path and translate them into text format



Feng, Jie, et al. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. arXiv 2024.

Summary & Opportunities

- Generalist language models lack sufficient real-world urban knowledge
  - Injecting urban-specific knowledge is the key for building effective language-based UFMs

- Pre-training: knowledge memorization
- Prompt engineering: zero-shot, few-shot, and chain-of-thought
- Model fine-tuning: knowledge memorization & spatial reasoning

- Next step - going beyond basic language capabilities
  - Spatio-temporal reasoning + acting
  - Tool utilization

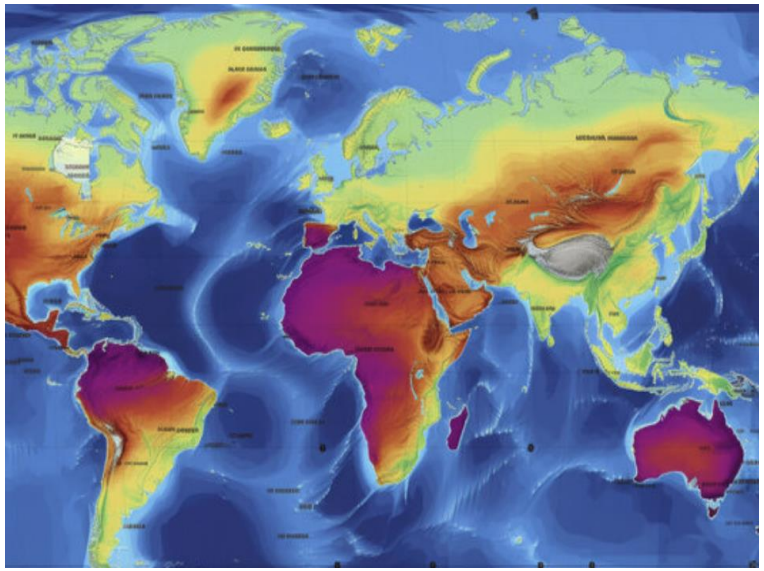
Vision-based UFMs

Visual Data in Cities

- The wide adoption of camera and satellite technologies have collected a vast volume of visual data in urban space

On-site visual data

Remote sensing (RS) data

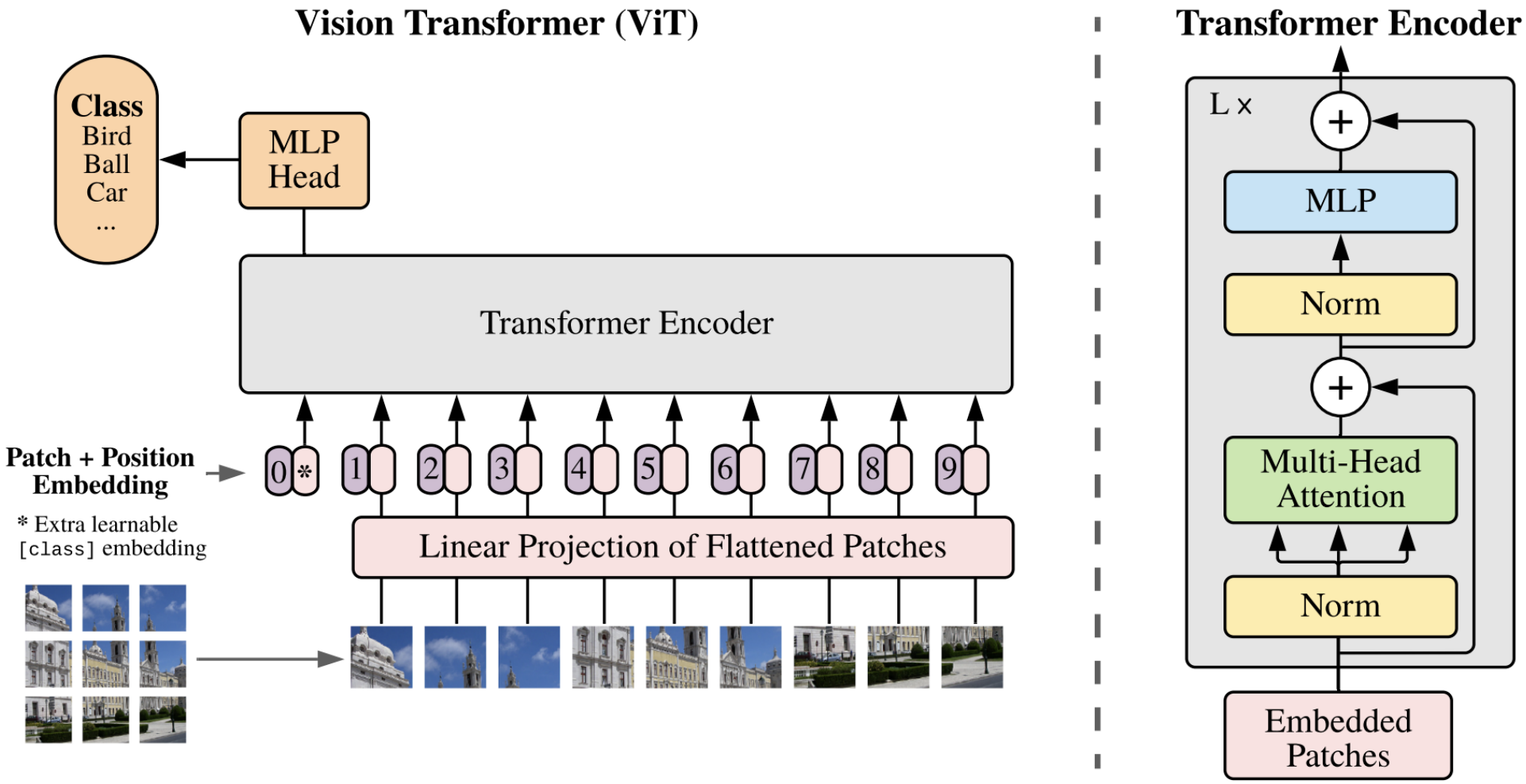


Urban raster data

How to build vision-based UFMs for urban scenarios?

Vision-based Models

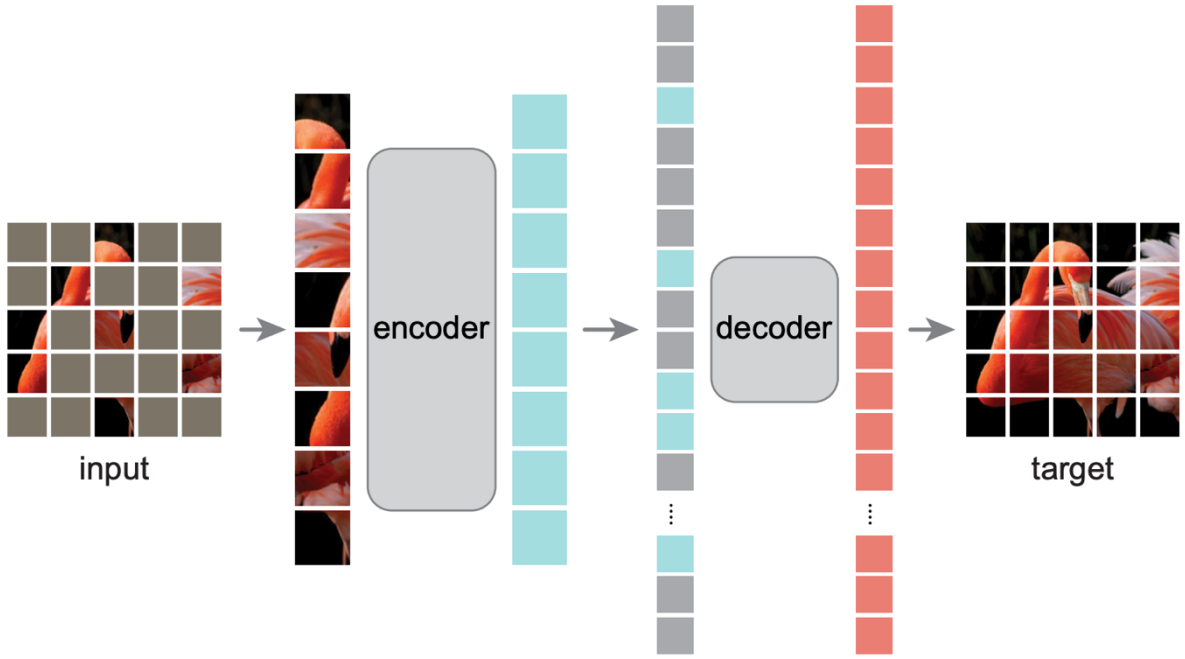
- Vision Transformer (ViT) is the most popular encoder for visual data



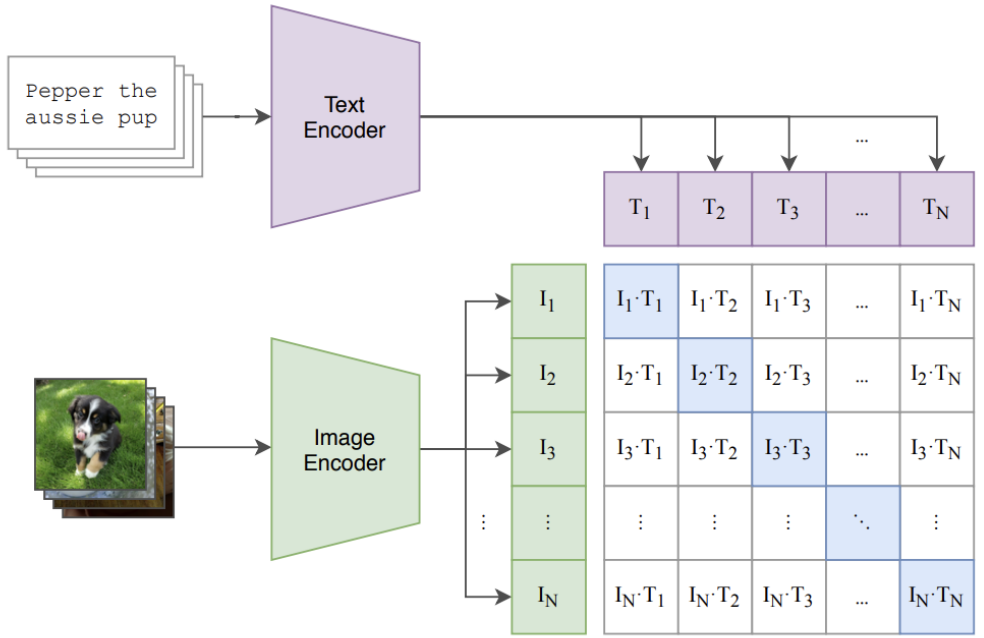
Dosovitskiy, Alexey, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2020.

Vision-based Models

- Two common ways for pre-training vision transformer
  - Self-supervision - masked patch prediction
  - Natural language supervision - CLIP



Masked patch prediction



CLIP

He, Kaiming, et al. Masked Autoencoders are Scalable Vision Learners. ICCV 2022.
Radford, Alec, et al. Learning Transferable Visual Models from Natural Language Supervision. ICML 2021.

Vision-based UFMs

Vision-based Pre-training

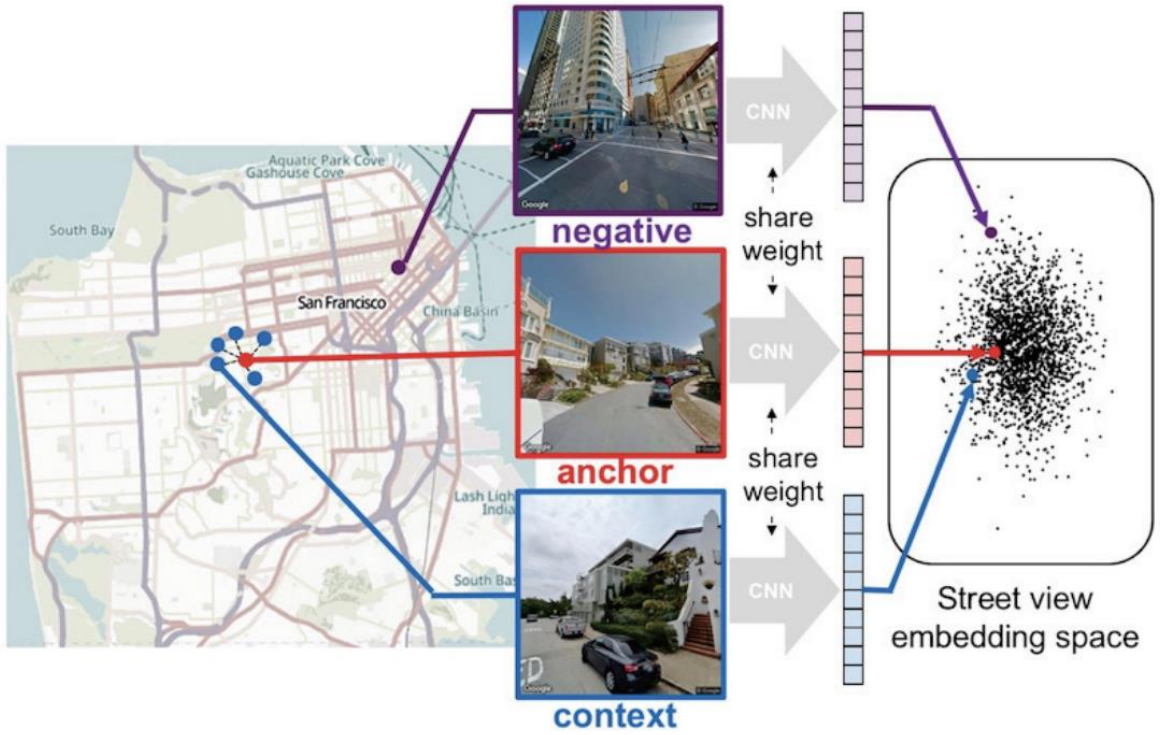
- *On-site visual data*
- *Remote sensing data*
- *Urban raster data*

Vision-based Adaptation

- Prompt engineering
- Model fine-tuning

On-Site Visual Data

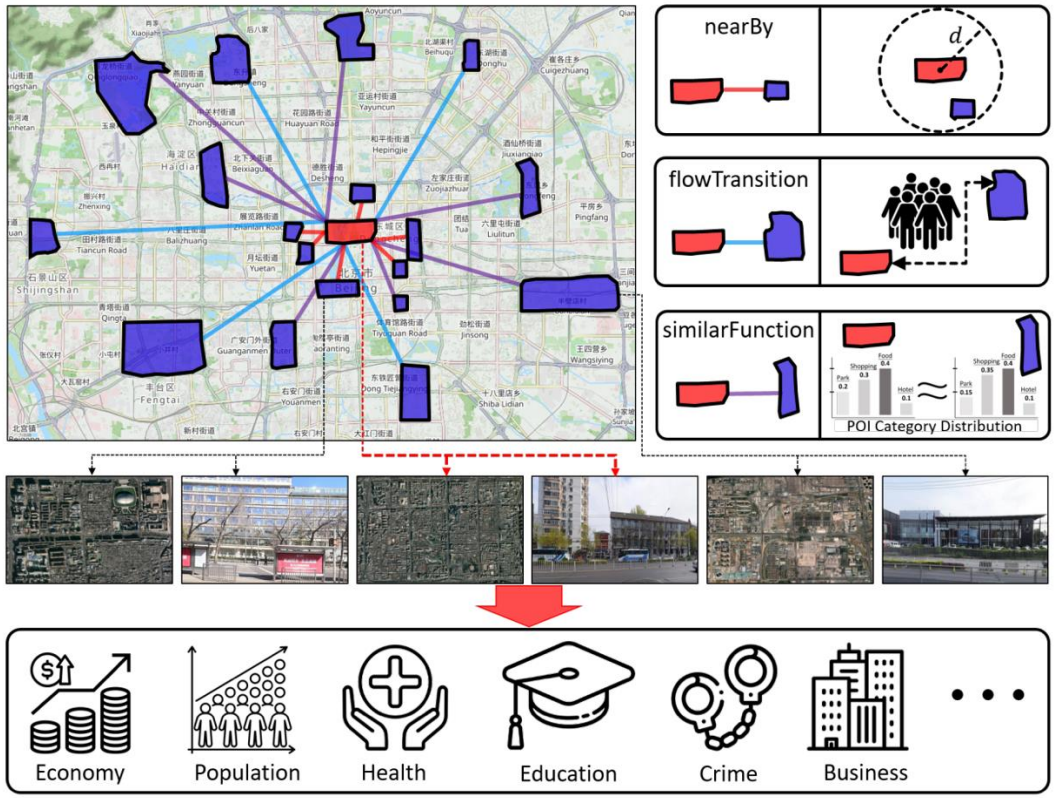
- On-site visual data is closely tied to specific geographic locations
- Spatial-aware contrastive pre-training
  - Urban2Vec: enforce spatially adjacent images to be similar in latent feature space

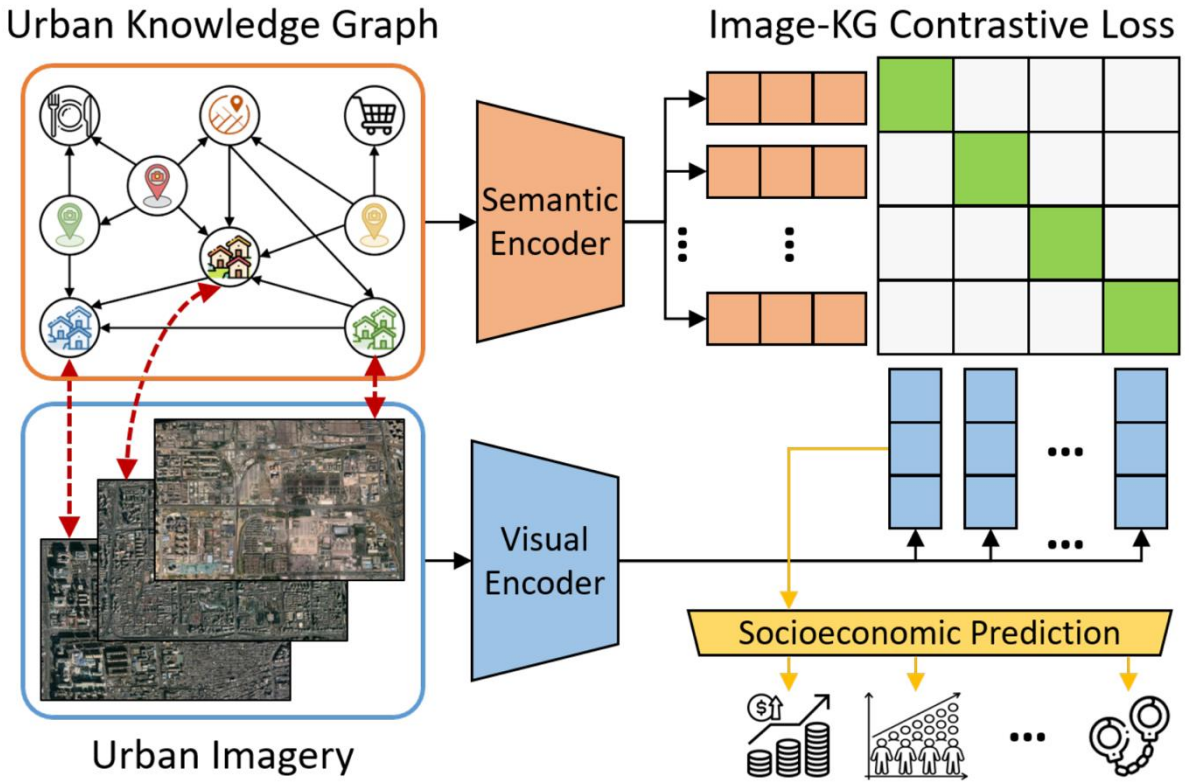


Wang, Zhecheng, et al. Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding. AAAI 2020.

On-Site Visual Data

- Knowledge graph enhanced pre-training
  - KnowCL: maximize the mutual information between street-view images and their corresponding knowledge graph embedding

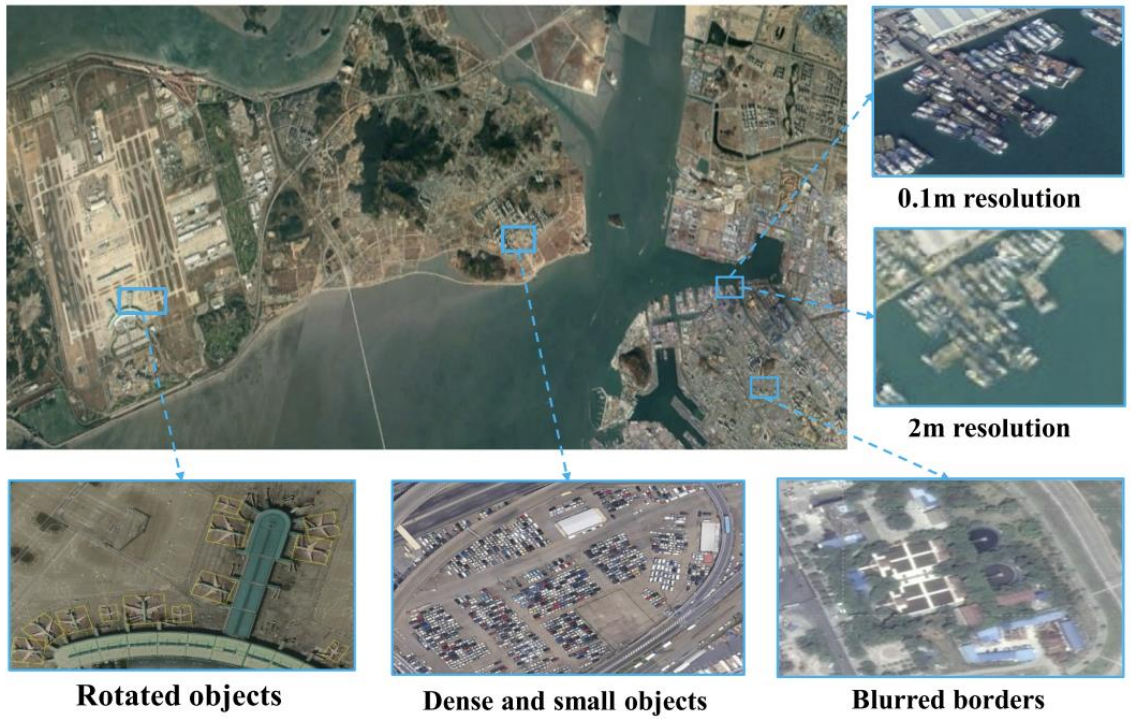




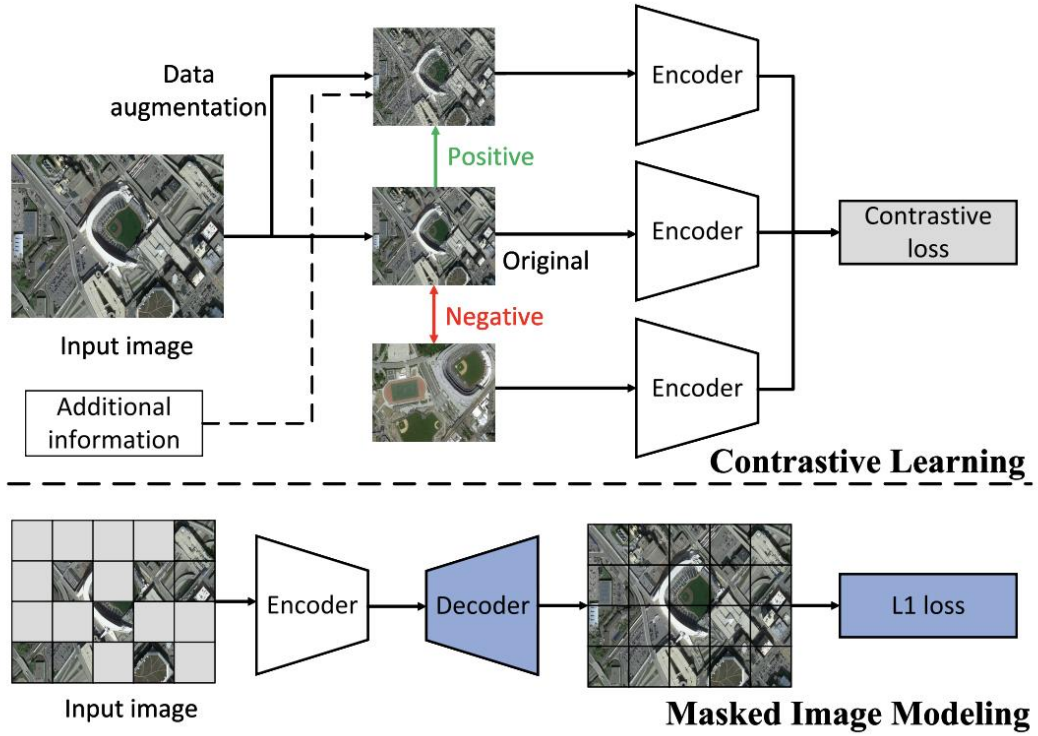
Liu, Yu, et al. Knowledge-infused Contrastive Learning for Urban Imagery-based Socioeconomic Prediction. WWW 2023.

Remote Sensing Data

- Unique properties
  - e.g., rotated objects, dense and small objects, …
- Pre-training on remote sensing data
  - Should take these unique properties into account



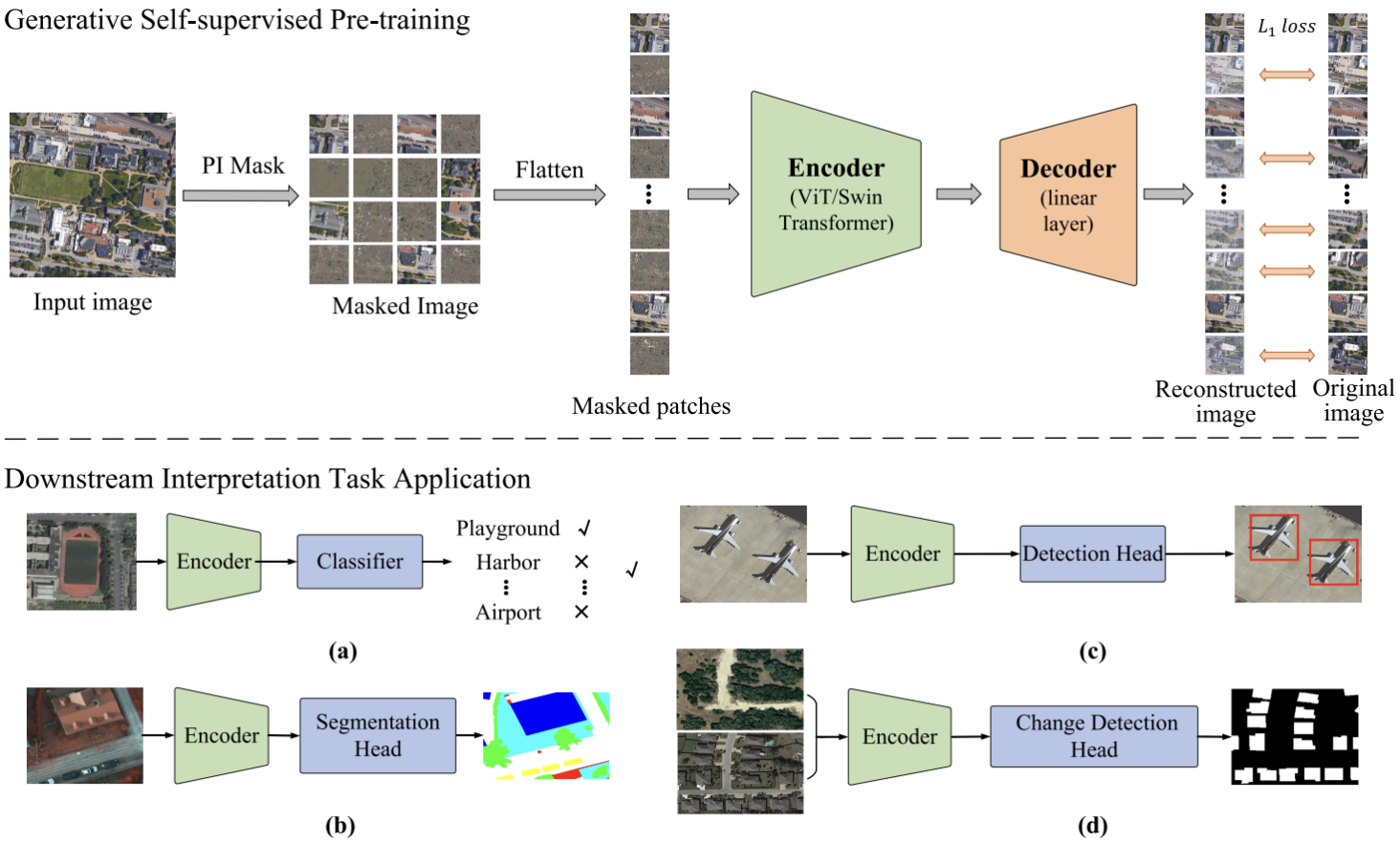
Unique properties



Pre-training paradigm

Remote Sensing Data

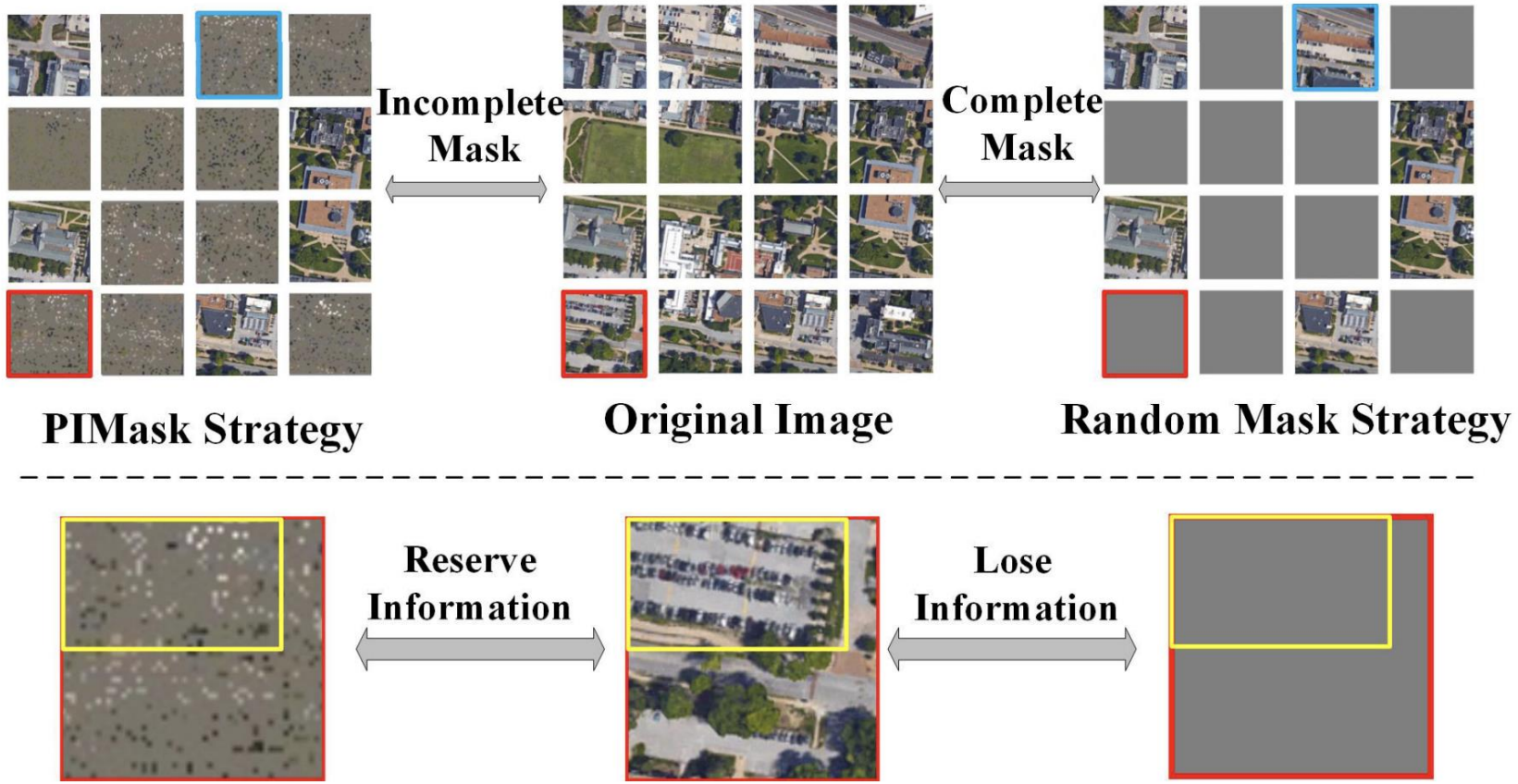
- Masked patch prediction
  - RingMo: encodes a masked RS image with an image encoder, and then decodes the masked portion to reconstruct the entire image



Yao, Fanglong, et al. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling. IEEE TGRS, 2023.

Remote Sensing Data

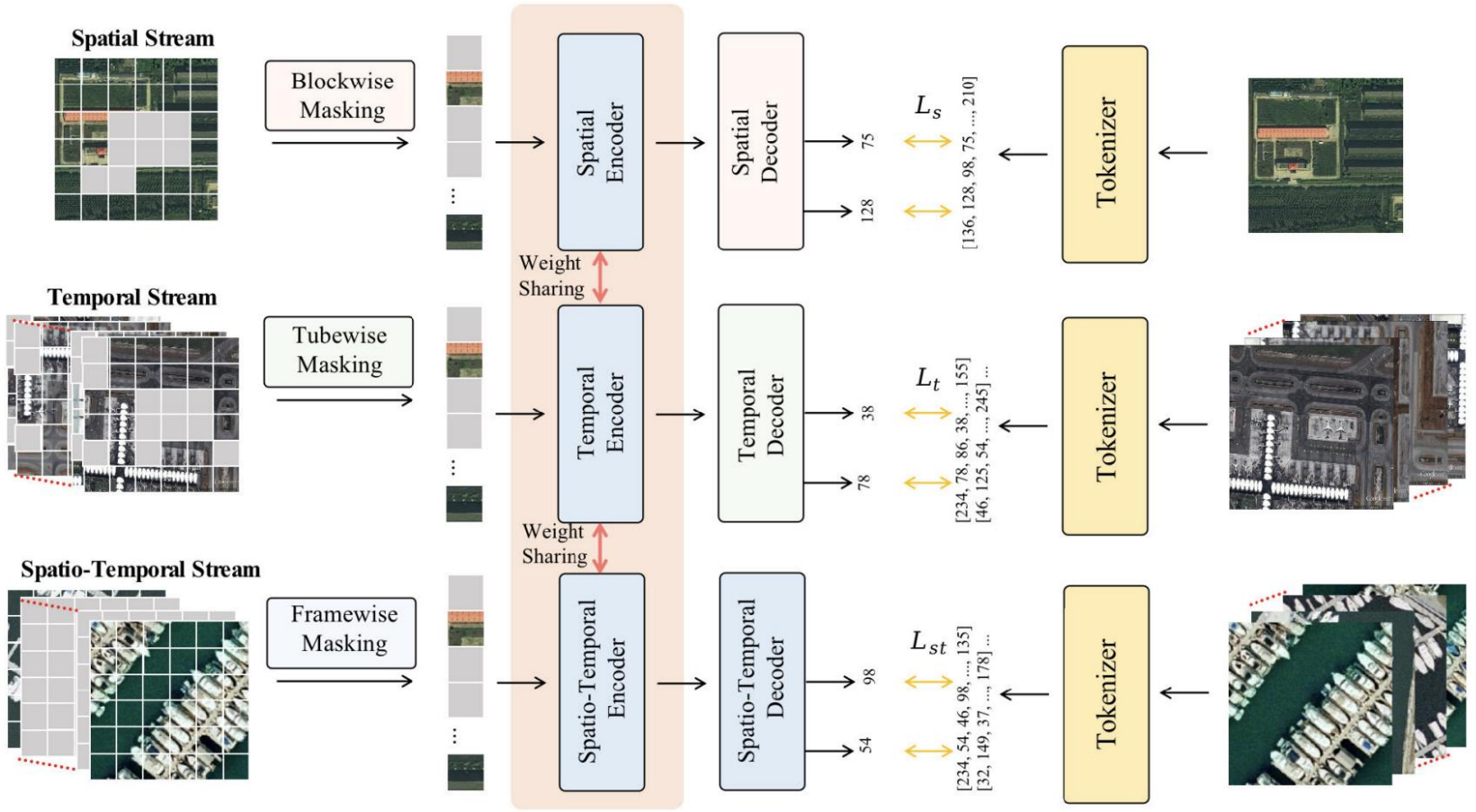
- Potential issue
  - Masking strategy overlooks small objects within the image patch
- RingMo randomly reserves some pixels in masked patches



Yao, Fanglong, et al. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling. IEEE TGRS, 2023.

Remote Sensing Data

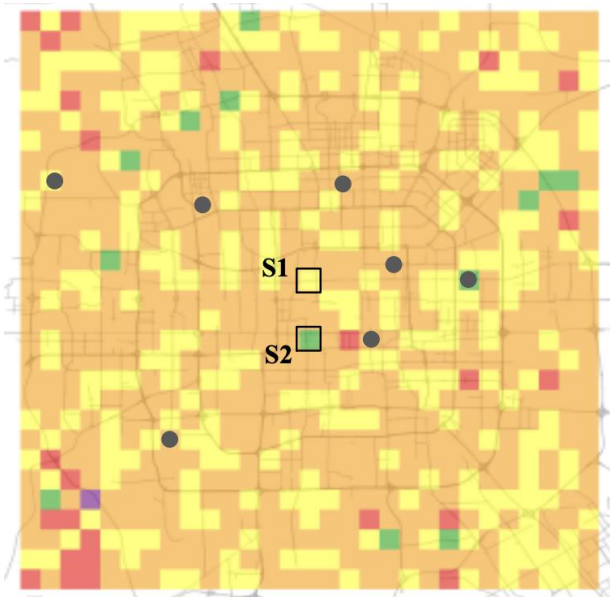
- Extending RingMo to 3D remote sensing data
  - RingMo-Sense: blockwise, tubewise, and framewise masking



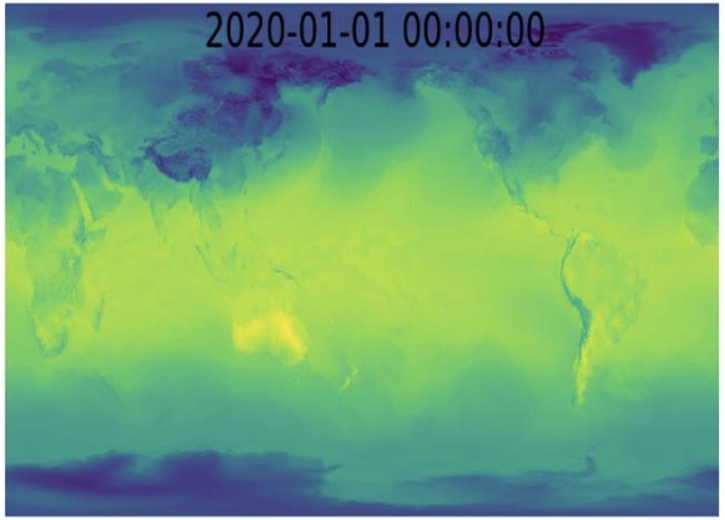
Yao, Fanglong, et al. RingMo-Sense: Remote Sensing Foundation Model for Spatiotemporal Prediction via Spatiotemporal Evolution Disentangling. IEEE TGRS, 2023.

Urban Raster Data

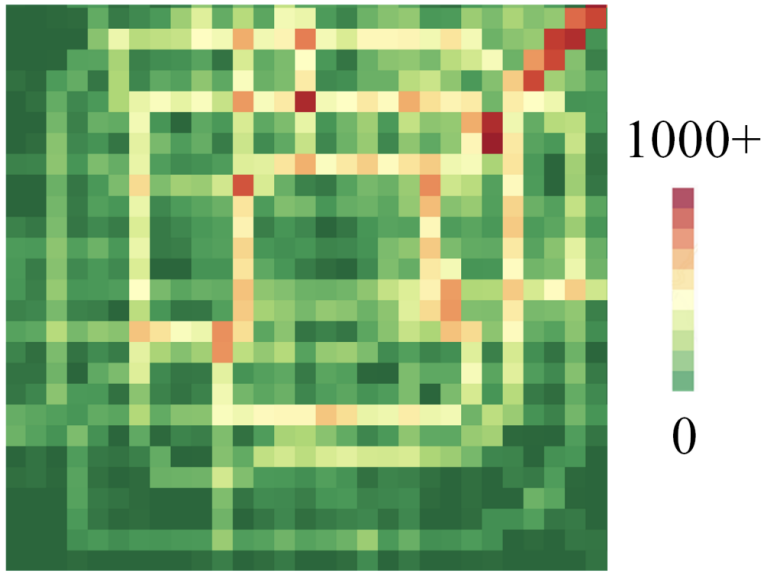
- Geospatial time series formatted as regular grid structure in urban space
  - Exhibit unique time series patterns, e.g., trend and cyclic behaviors



Air Pollution
[Zheng et al. 2013]



Raster-based
Weather Data

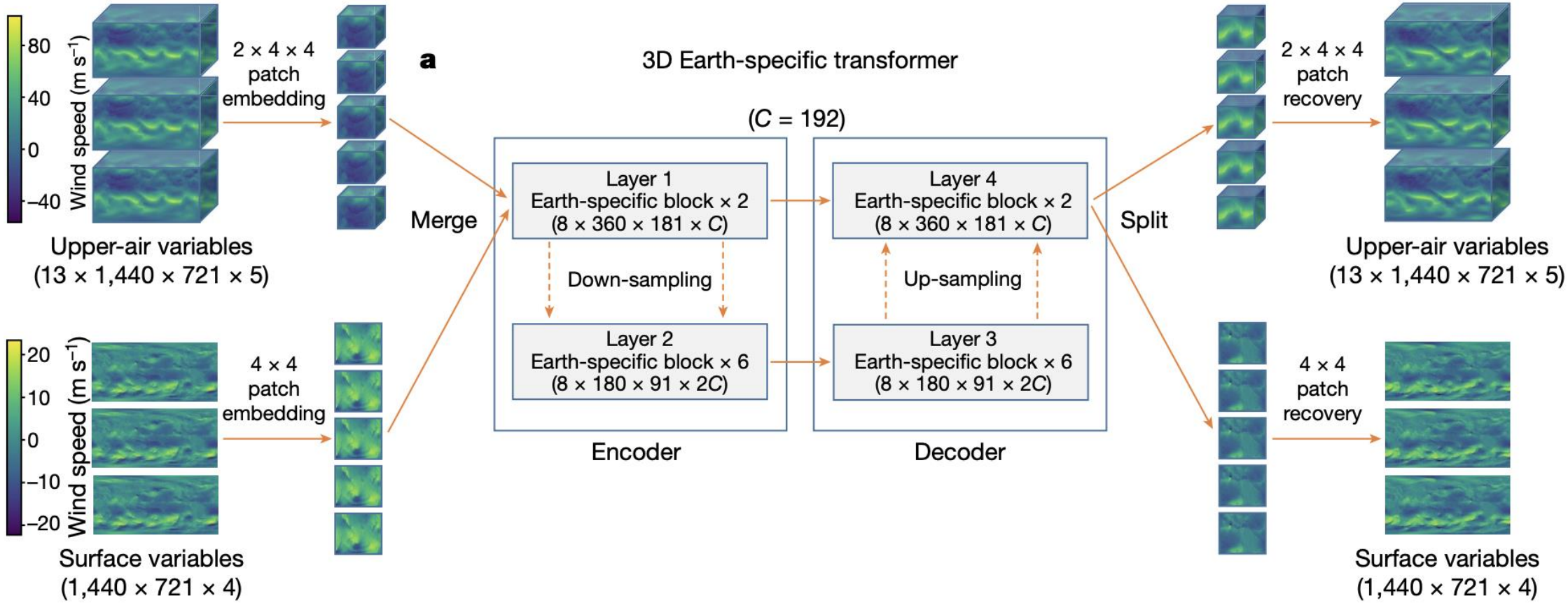


Taxi Flow
[Zhang et al. 2017]

H. Hersbach et al., The ERA5 Global Analysis, in Quarterly Journal of the Royal Meteorological Society, 2020.

Urban Raster Data

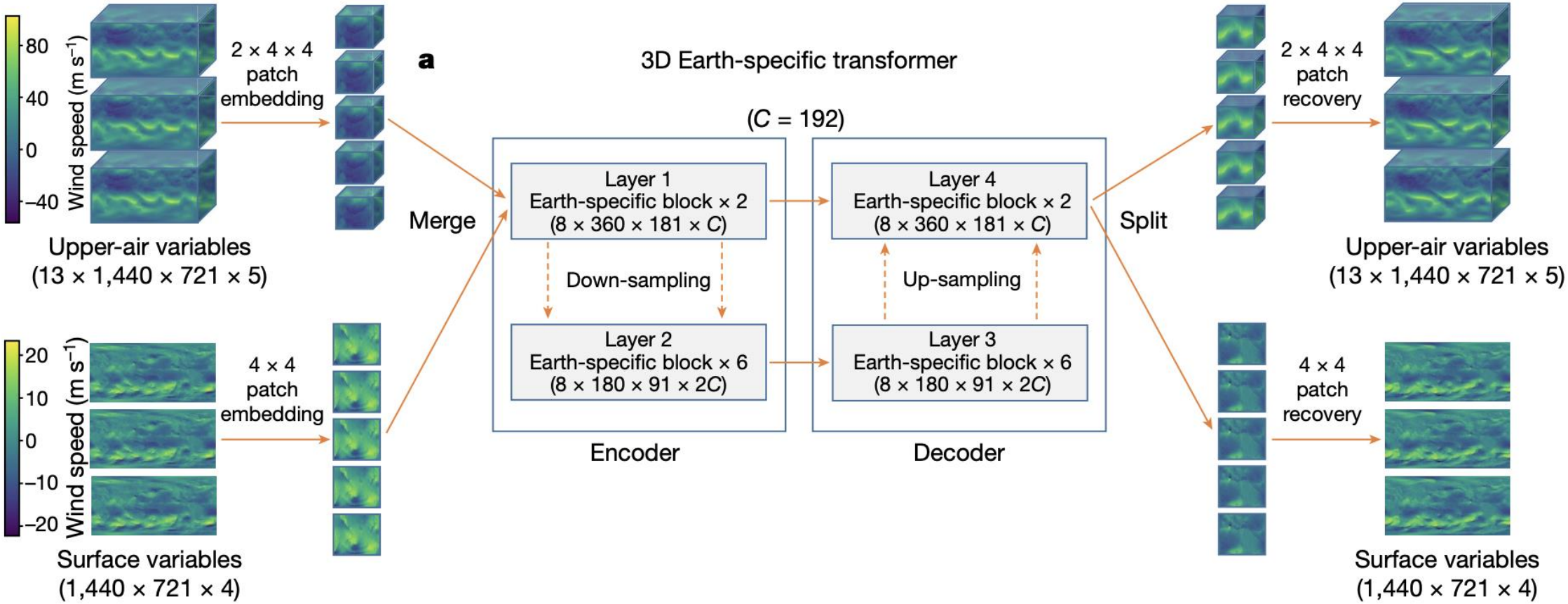
- Pre-training on raster-based weather data
  - Pangu-Weather: represent surface and upper-air weather variables as 3D gridded data, and divide 3D data into 3D patches



Bi, Kaifeng, et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature 2023.

Urban Raster Data

- Pre-training on raster-based weather data
  - Pangu-Weather: 3D patches are then passed through a series of transformer blocks for future prediction



Bi, Kaifeng, et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature 2023.

Vision-based UFMs

Vision-based Pre-training

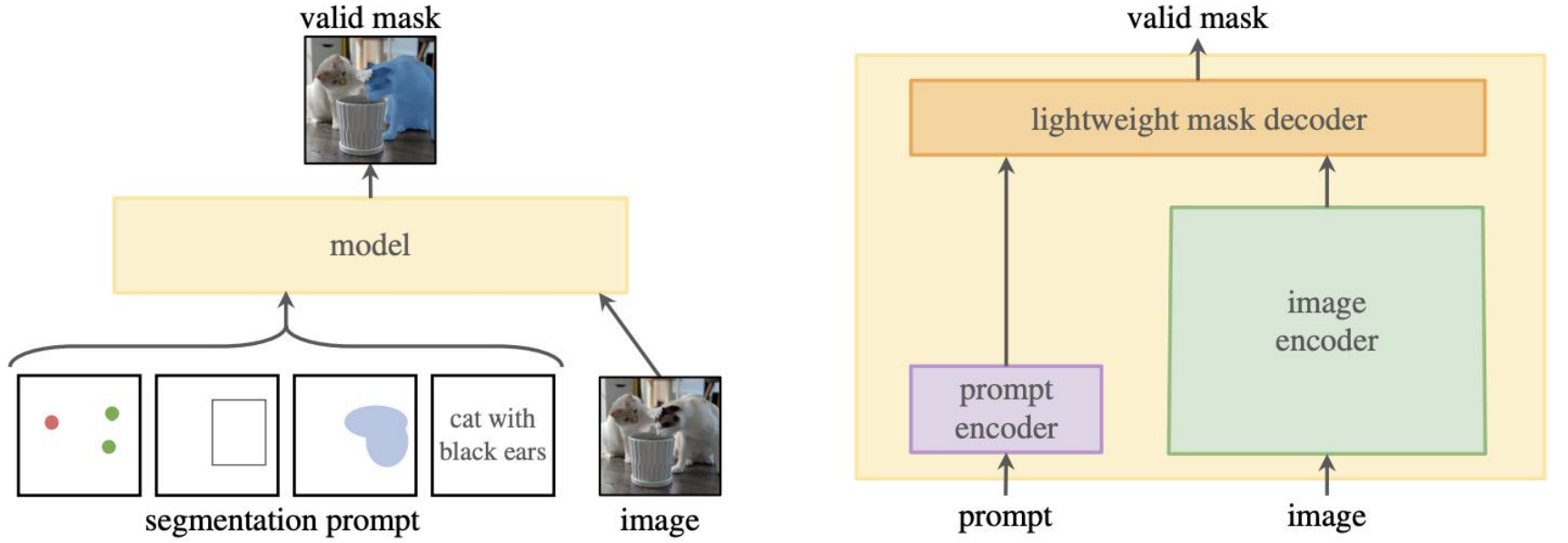
- On-site visual data
- Remote sensing data
- Urban raster data

Vision-based Adaptation

- *Prompt engineering*
- *Model fine-tuning*

Prompt Engineering

- What are the visual prompts?
- Segment Anything Model (SAM)
  - A pre-trained semantic segmentation model
  - Allow users to use various types of prompts, such as points, boxes, or coarse masks, to specify certain parts of an image for segmentation

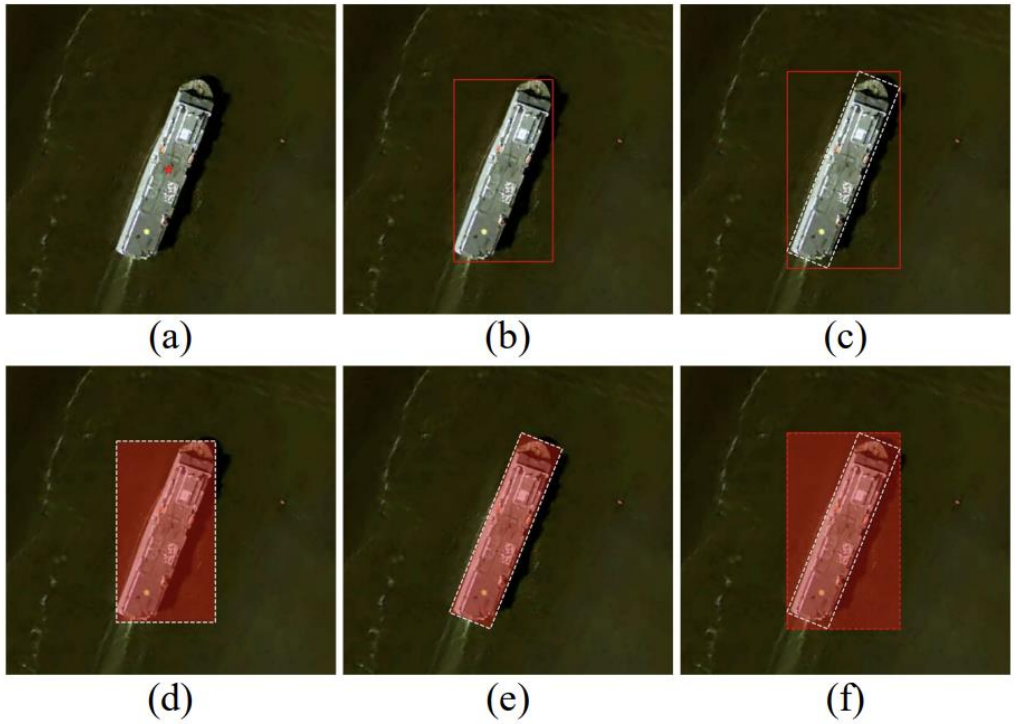


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**)

Kirillov, Alexander, et al. Segment Anything. CVPR 2023.

Prompt Engineering

- Hand-crafted prompts
  - SAMRS: design six basic prompts depending on the characteristics of RS image data



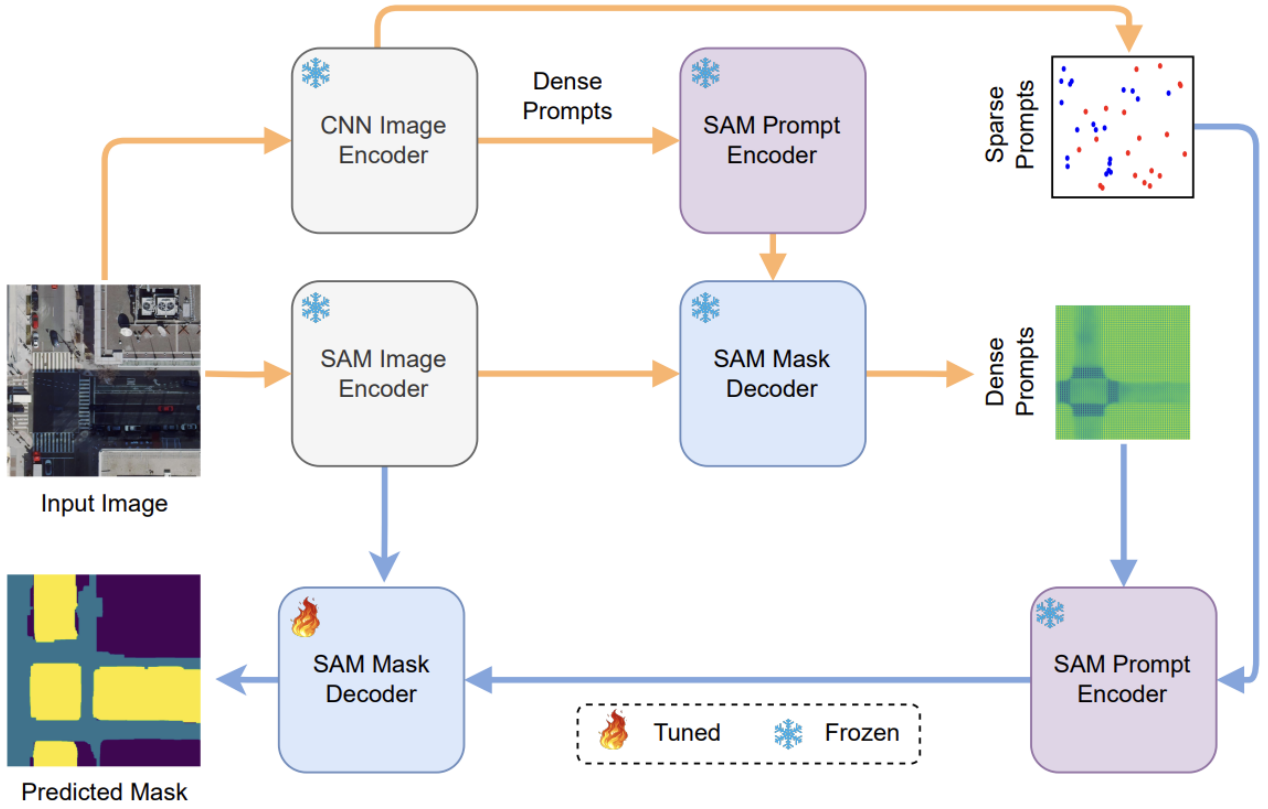
Rotated objects

- RS images contain rotated objects: rotated bounding boxes as box prompts

Wang, Di, et al. SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. NeurIPS 2023.

Prompt Engineering

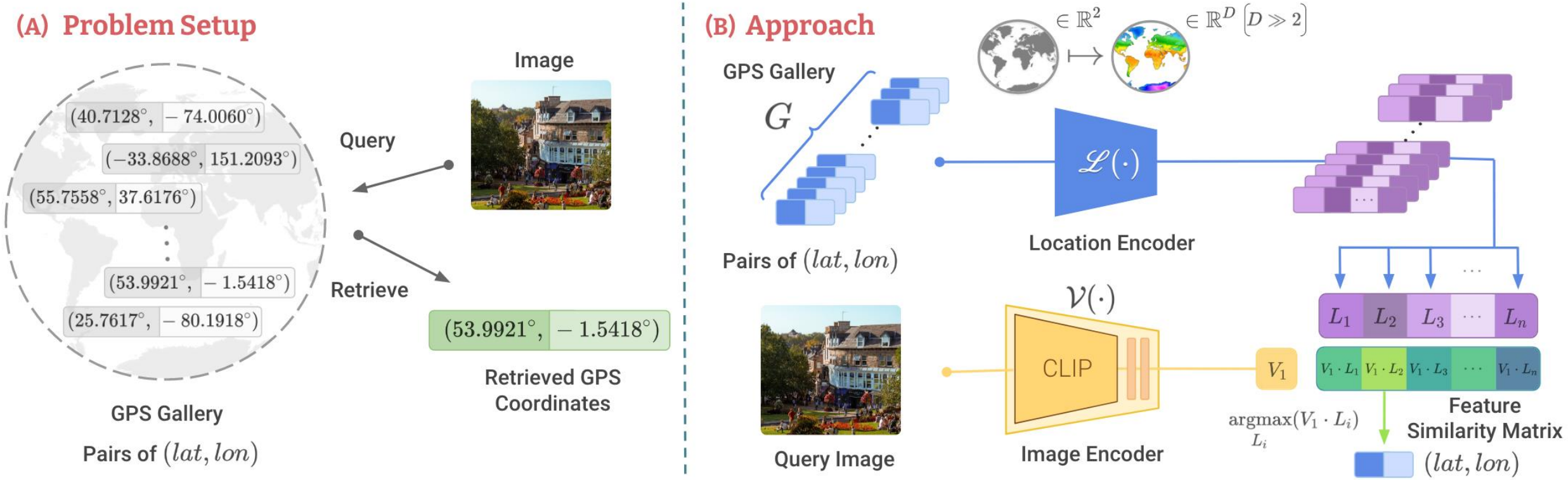
- Automated prompts
  - GeoSAM: generate the coarse mask prompt from SAM itself, and the point prompt from a pre-trained CNN segmentation model



Sultan, Rafi Ibn, et al. GeoSAM: Fine-tuning SAM with Sparse and Dense Visual Prompting for Automated Segmentation of Mobility Infrastructure. arXiv 2023.

Model Fine-tuning

- GeoCLIP
  - Fine-tune CLIP for image-to-GPS retrieval
  - Explicitly align image and its GPS coordinates in latent feature space



Vivanco Cepeda et al. GeoCLIP: CLIP-inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. NeurIPS 2023.

Summary & Opportunities

- Pre-training: contrastive learning, masked patch prediction
- Prompt engineering: leverage zero-shot capability of SAM
- Model fine-tuning: fine-tune CLIP for image-location retrieval

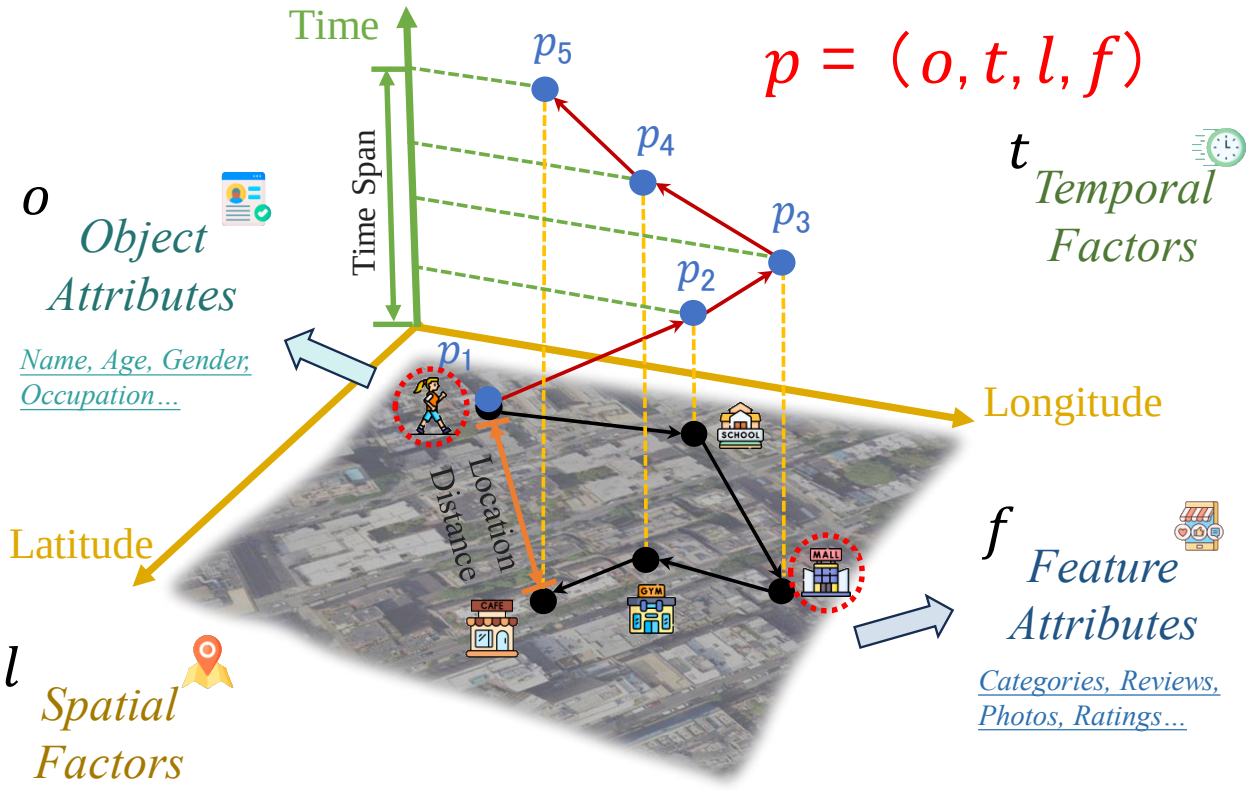
- Existing studies are still task-specialized with relatively small datasets
  - lack general-purpose utility in diverse urban scenarios

- Potential directions
  - Cross-domain pre-training, e.g., street-view and remote sensing images
  - Pre-training with cross-modal supervision, e.g., POI, time series, trajectories

Trajectory-based UFMs

Trajectory Data in Cities

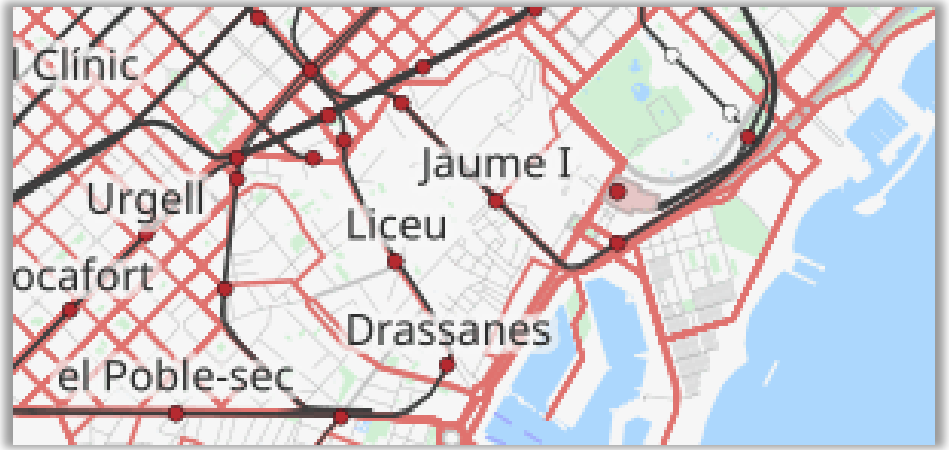
- Trajectory data plays a crucial role in urban environments.



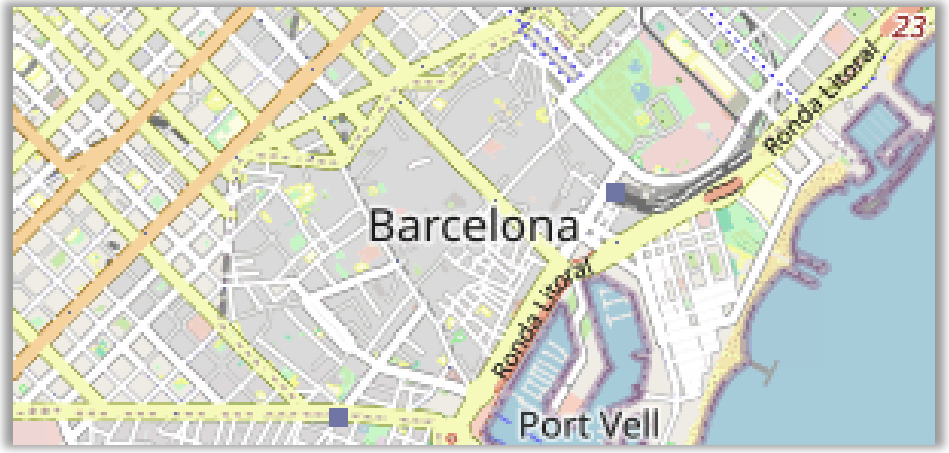
Chen, Wei, et al. Deep learning for trajectory data management and mining: A survey and beyond. arXiv 2024.

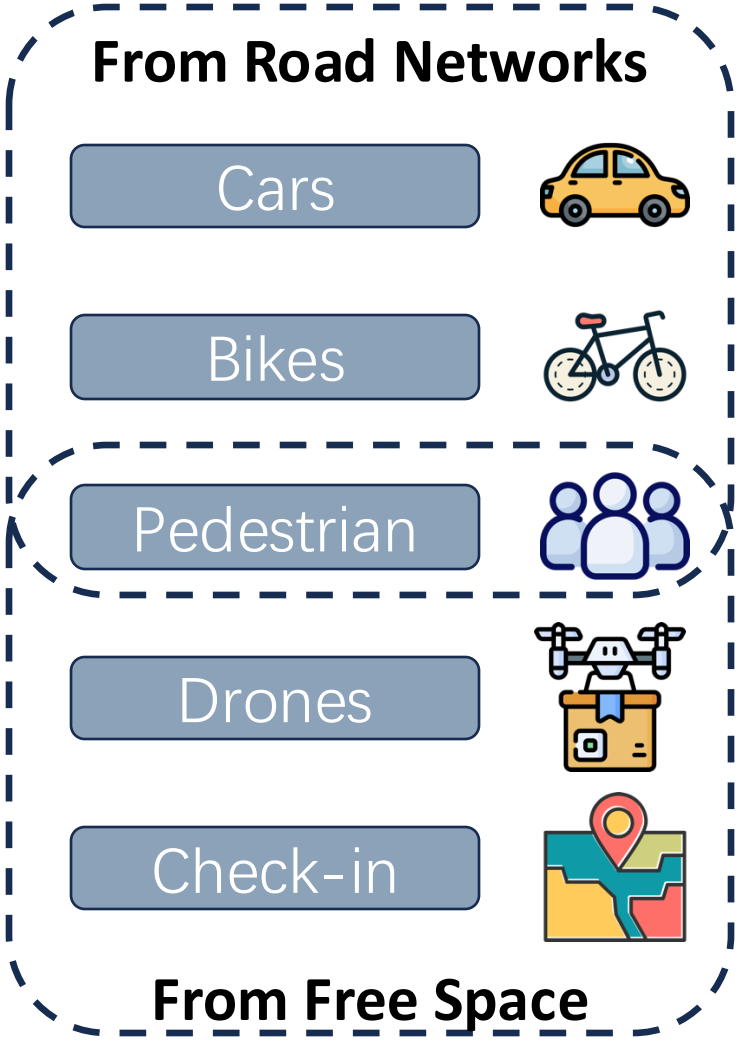
Trajectory Data in Cities

- Sources of Trajectory Data
  - From road network



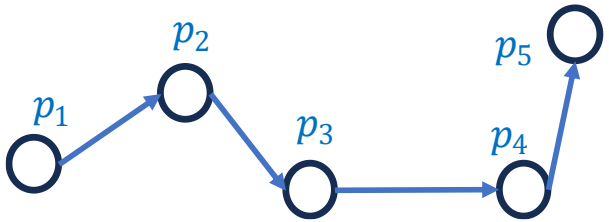
  - From free space





DL Paradigms for Trajectory Modeling

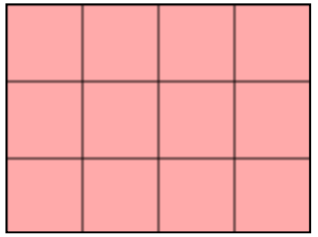
- Representation format of trajectory data



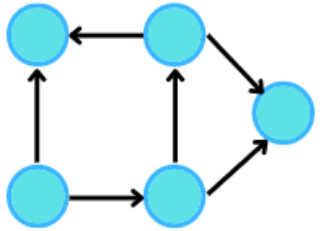
Raw Trajectory

$(P_1, \ldots, P_5)$

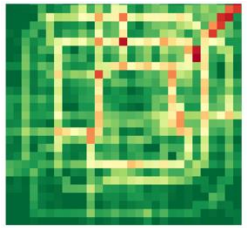
Sequence



Matrix

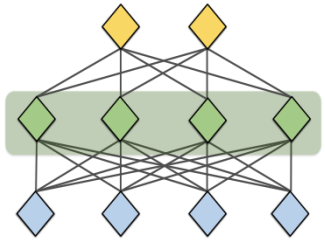


Graph

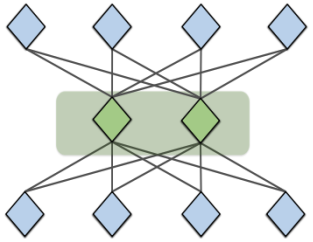


Image

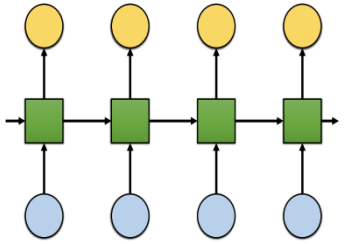
- Trajectory data modeling



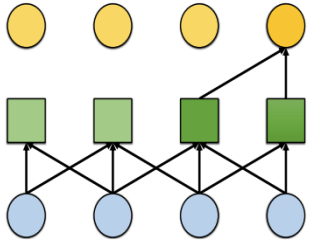
(a) FCs



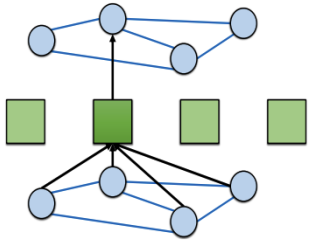
(b) AEs



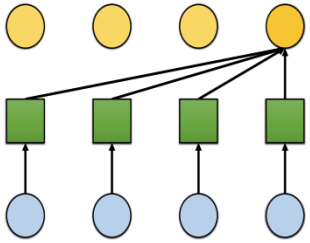
(c) RNNs



(d) CNNs



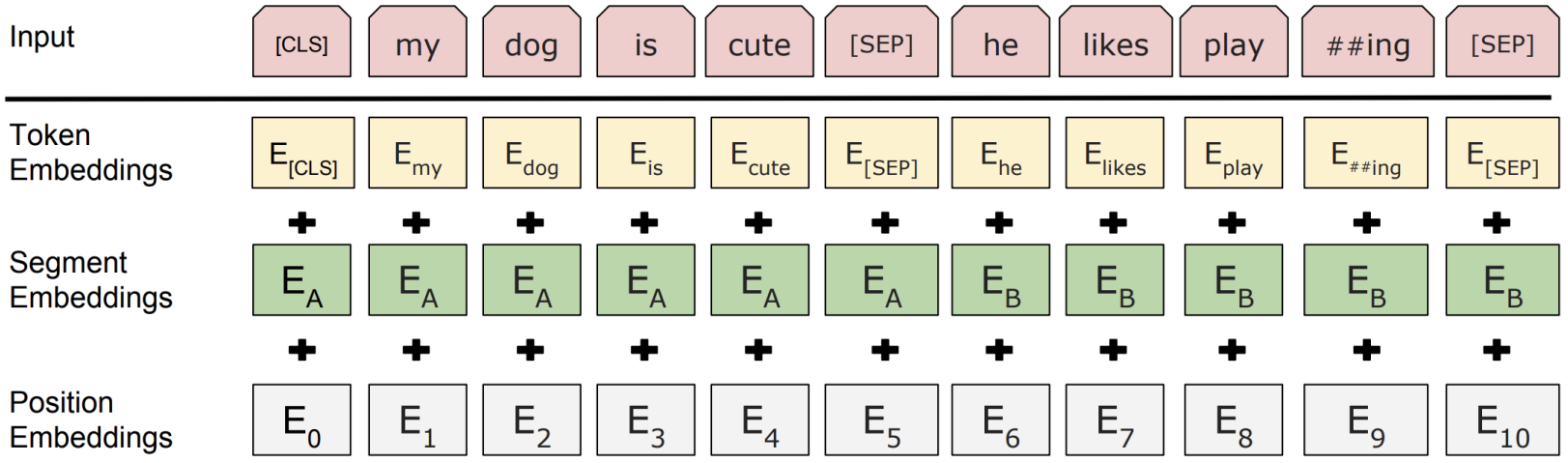
(e) GNNs

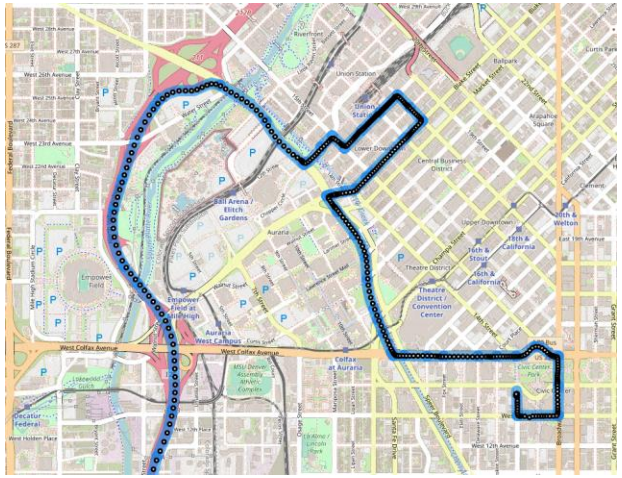


(f) Transformer

Limitations of Language-based Foundation Models

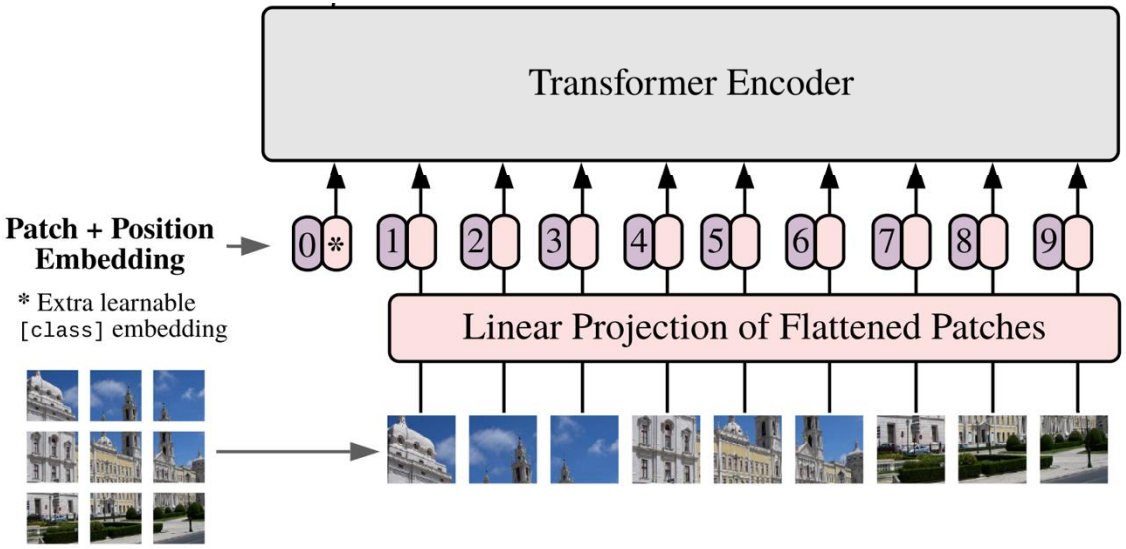
- Why trajectory-based urban foundation models are necessary?
  - Designed for discrete text sequences.
  - Effective for capturing semantic and contextual relationships between words.
  - Challenge: Cannot effectively model continuous spatial-temporal sequence or geographic constraints (e.g., road networks).

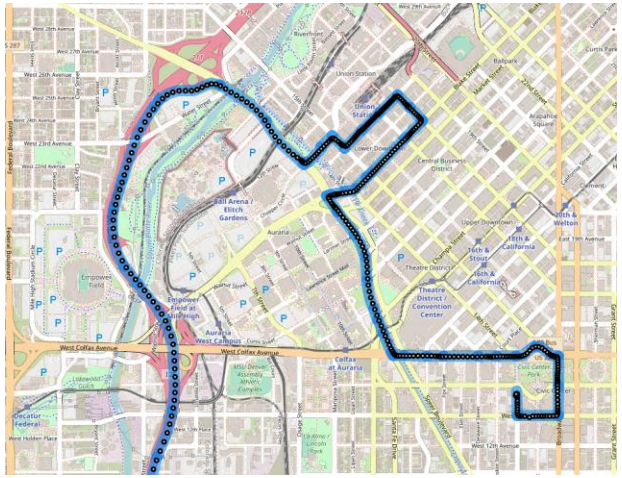




Limitations of Vision-based Foundation Models

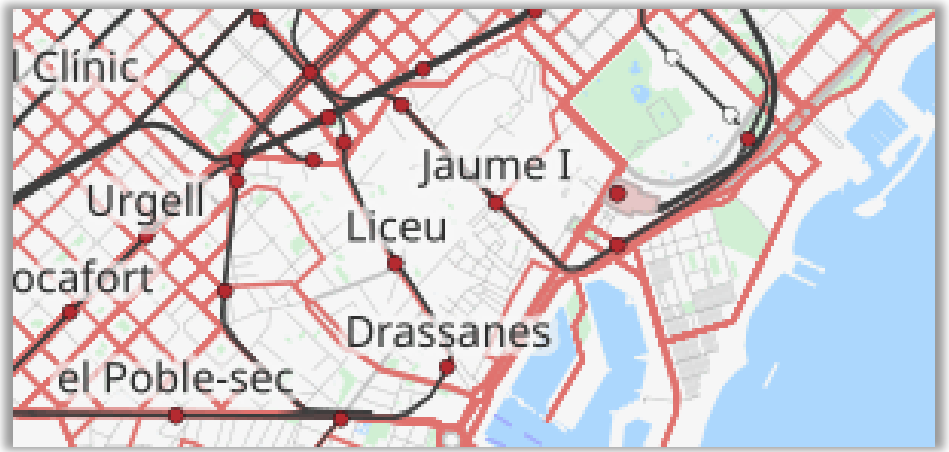
- Why trajectory-based urban foundation models are necessary?
  - Focus on extracting pixel-level visual information from images and videos.
  - Effective for identifying objects and scenes.
  - Challenge: Cannot reason about complex spatial relationships that are not directly related to visual inputs.



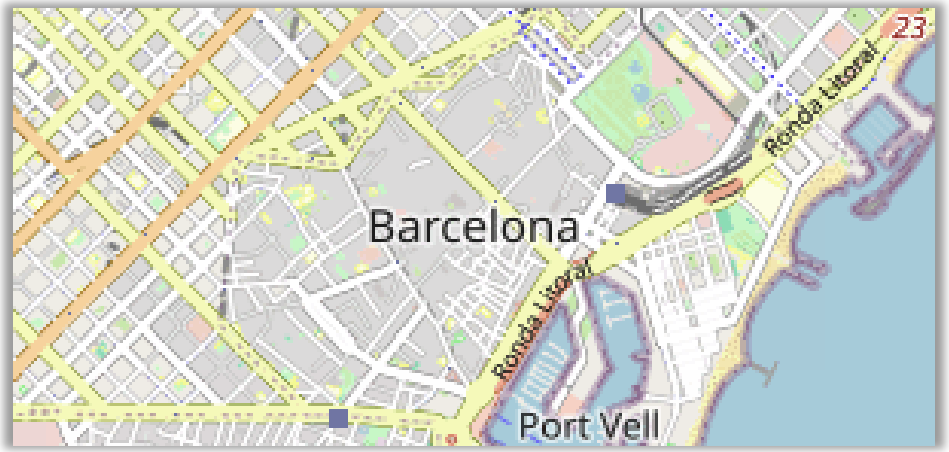


Trajectory-based UFMs

- **Trajectory-based Pre-training**
  - *Road network trajectory*
  - *Free space trajectory*
- **Trajectory-based Adaptation**
  - Model fine-tuning
- **Cross-modal Adaptation**
  - Prompt engineering



Road Network Trajectory

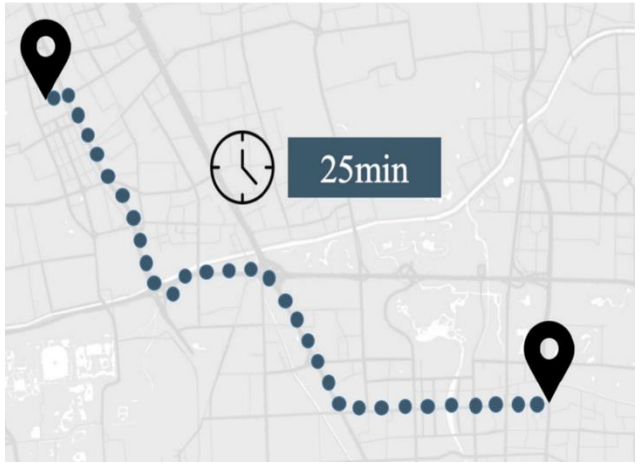


Free Space Trajectory

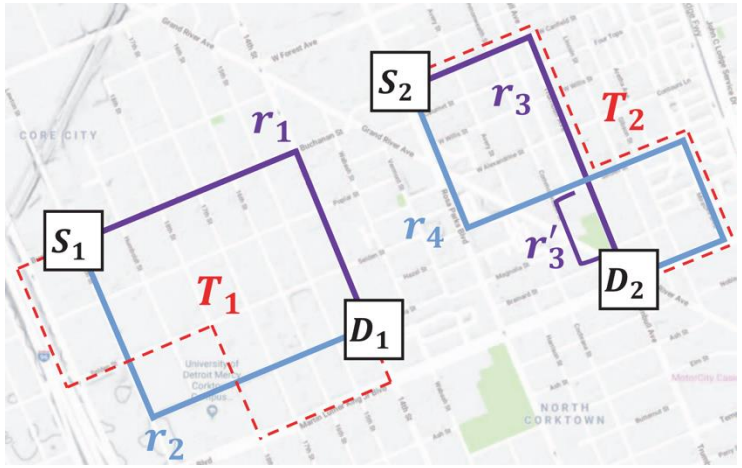
Trajectory-based Pre-training

- Goal: Learning low-dimensional embeddings of trajectories, paths, and locations (POIs).

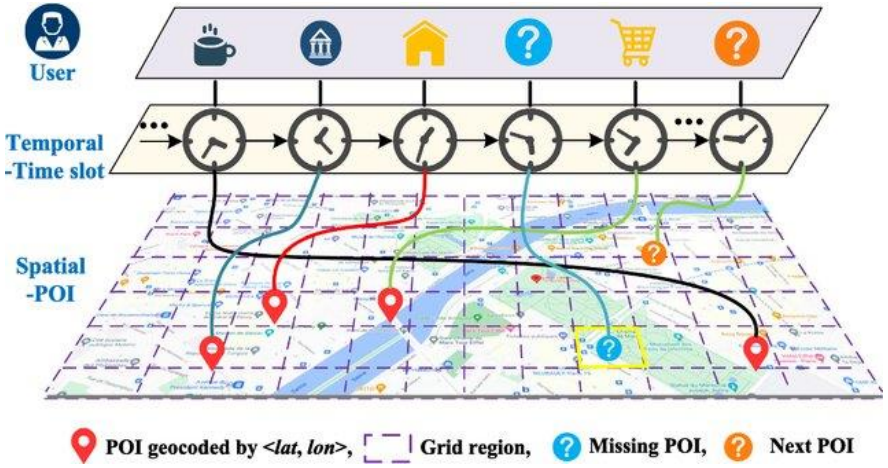
- Downstream Tasks:



Travel Time Estimation



Anomaly Detection

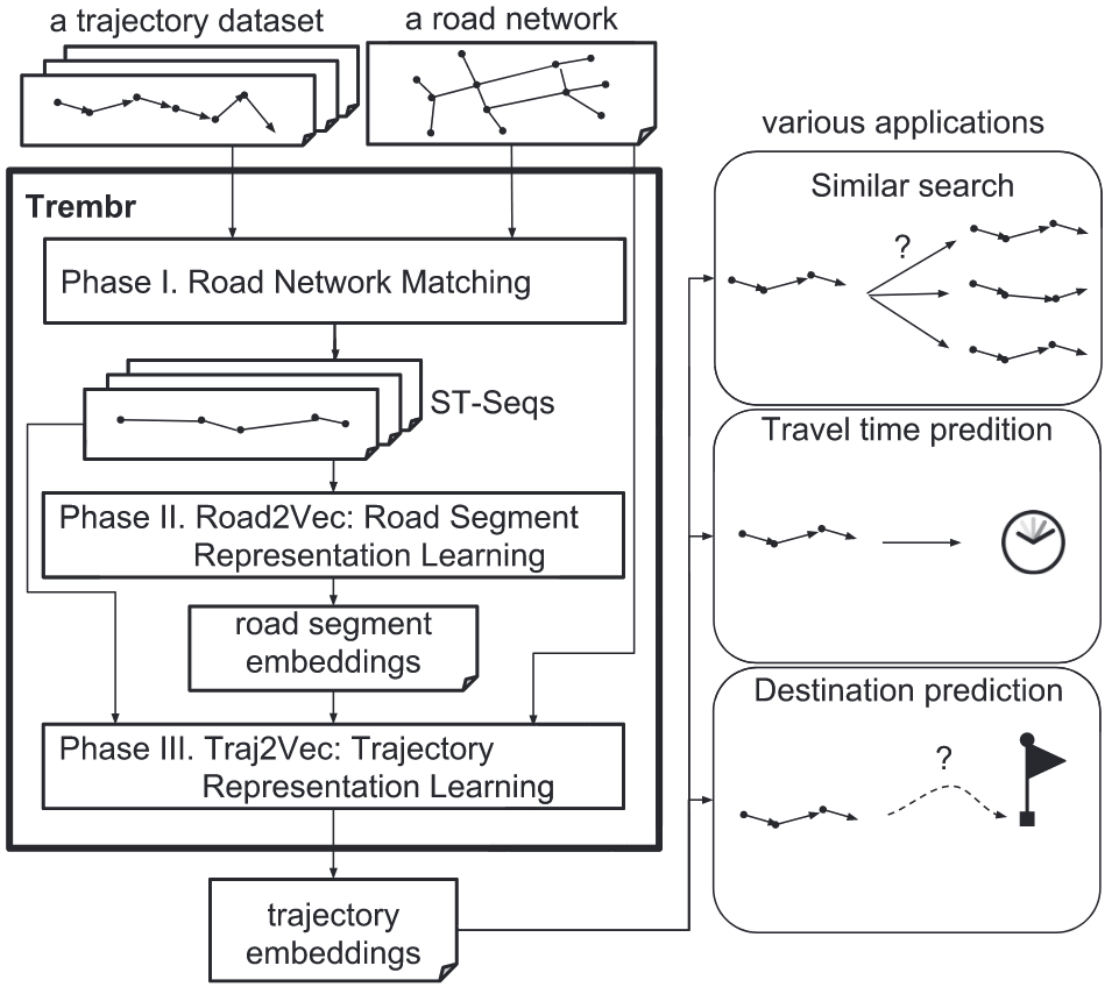


Next Location/POI Prediction

Trajectory User Linking

Pre-training on Road Network Trajectory - Trembr

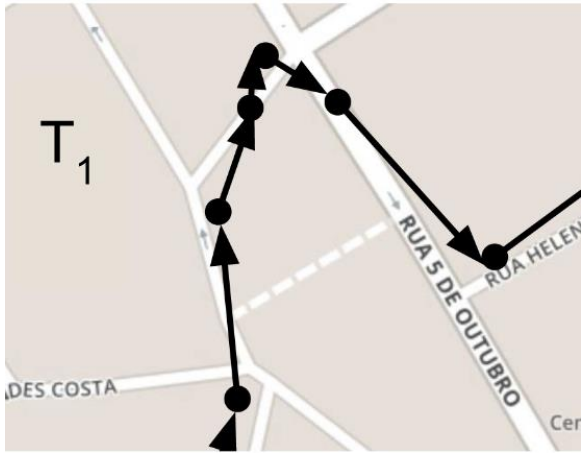
- Leverage road network constraints with RNN-based models to learn more generalized trajectory representations.



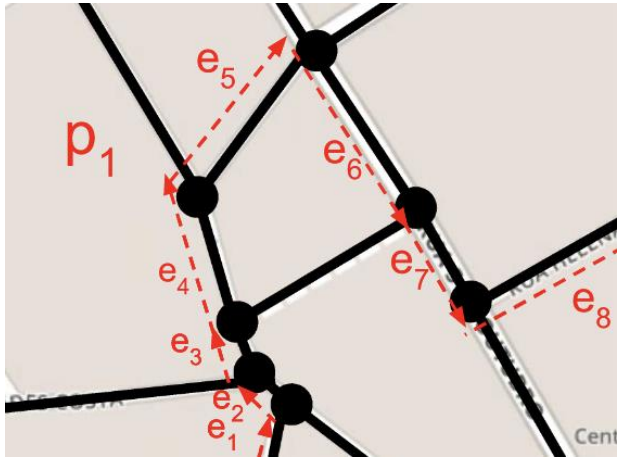
Fu, Tao-Yang, et al. Trembr: Exploring road networks for trajectory representation learning. TIST 2020.

Pre-training on Road Network Trajectory - Trembr

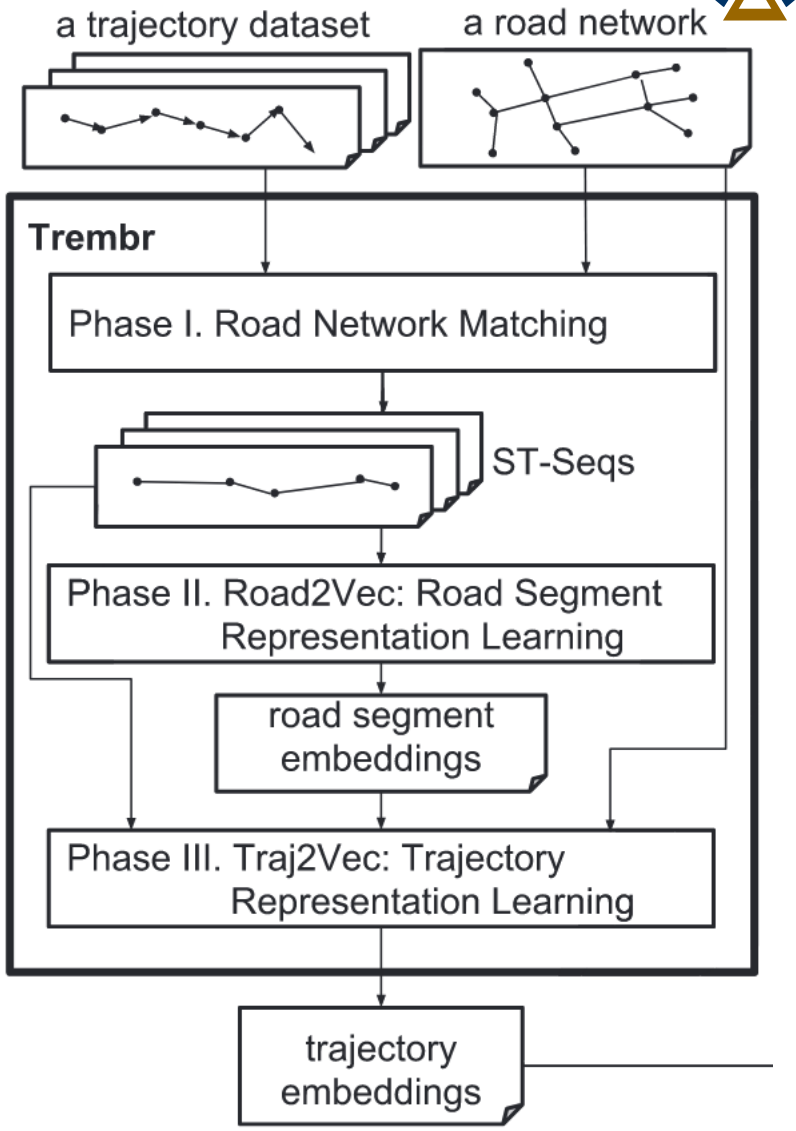
- Road2Vec: learn representations of road segments.
- Traj2Vec: learn trajectory embeddings by encoding both the spatial and temporal properties of trajectories.



Raw Trajectory



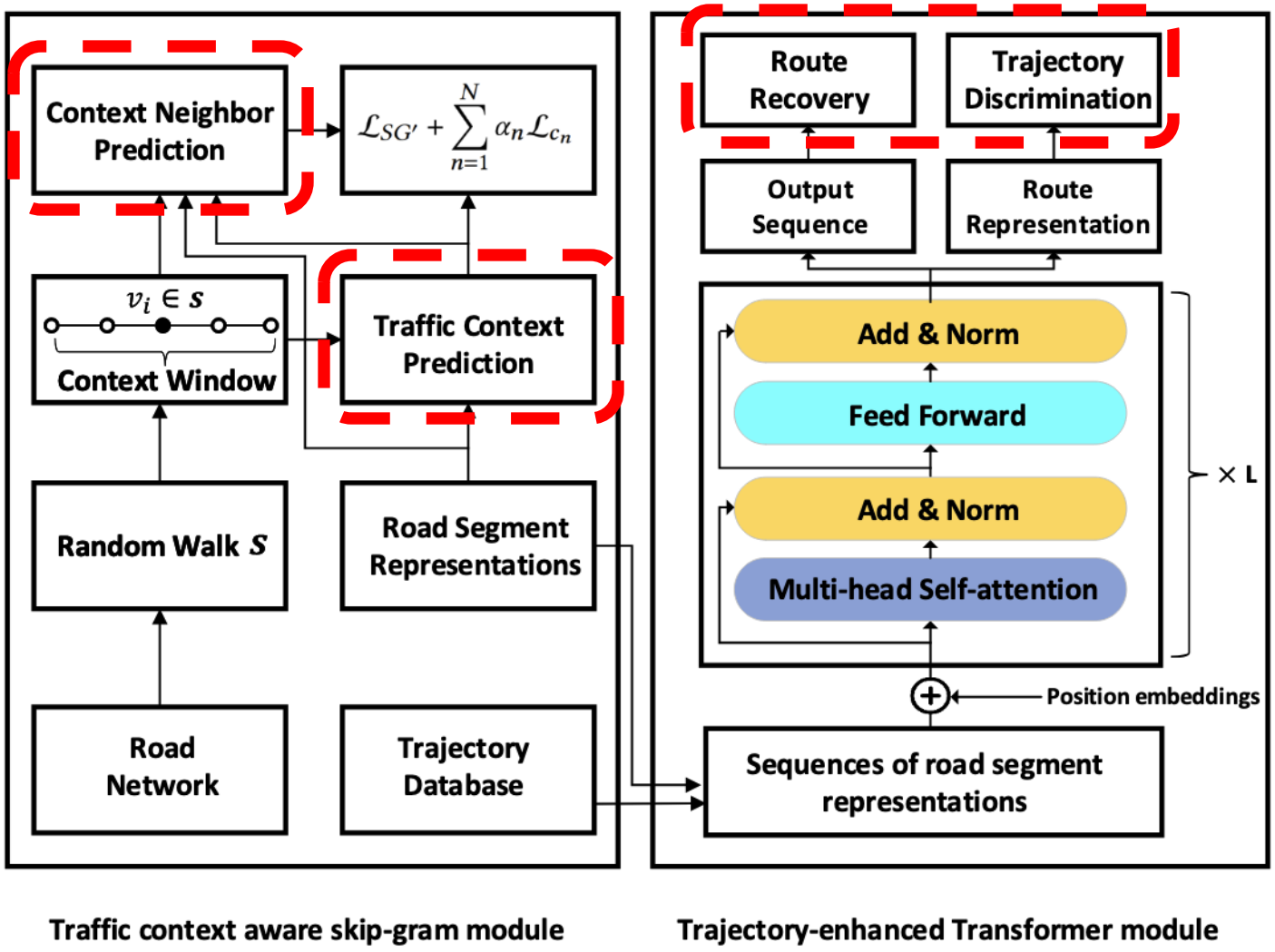
ST Sequence



Fu, Tao-Yang, et al. Trembr: Exploring road networks for trajectory representation learning. TIST 2020.

Pre-training on Road Network Trajectory - Toast

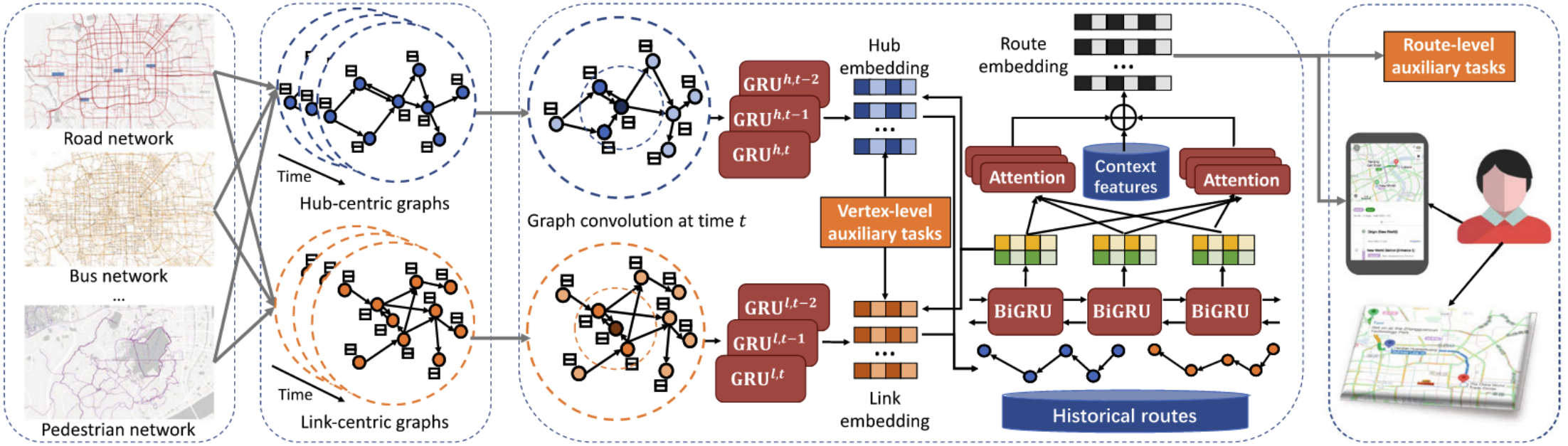
- Method:
  - Incorporating an auxiliary task of predicting traffic context features.
  - Encoding transition patterns and high-order dependencies between road segments.



Chen, Yile, et al. Robust road network representation learning: When traffic patterns meet traveling semantics. CIKM 2021.

Pre-training on Road Network Trajectory - HMTRL

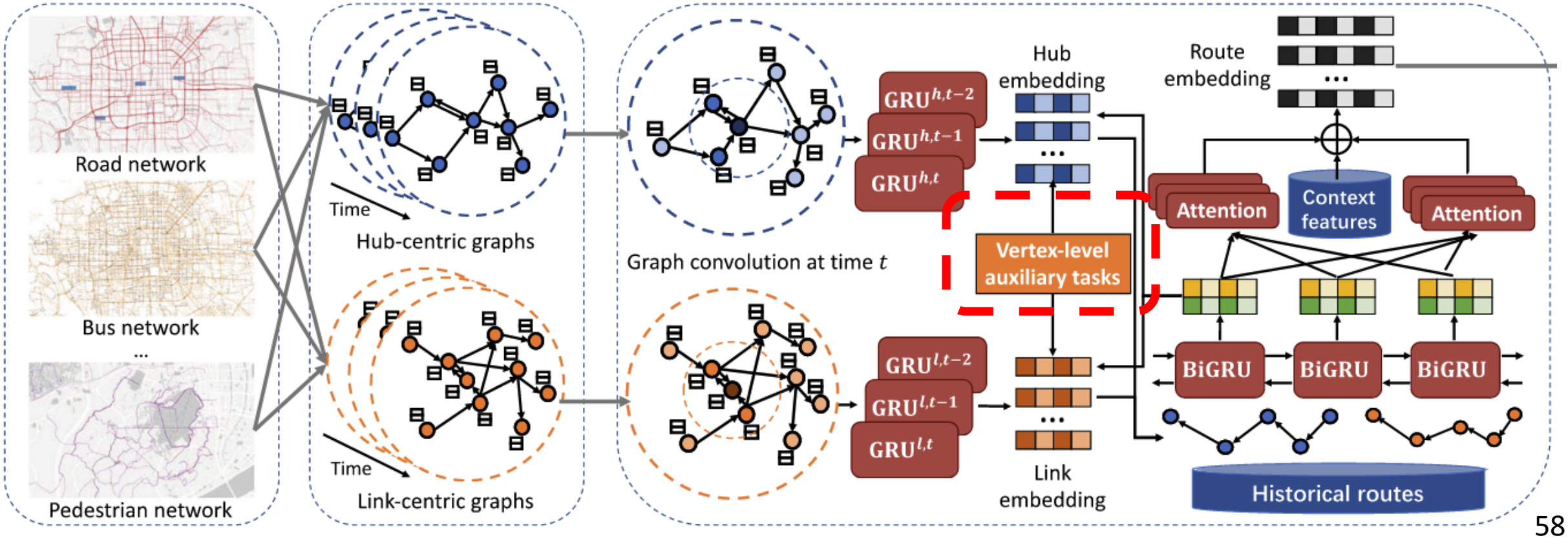
- Motivation:
  - Primarily focuses on single-modal representations.
  - Fail to capture the spatiotemporal dependencies and route coherence across diverse transportation mode.



Liu, Hao, et al. Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training. VLDB 2023.

Pre-training on Road Network Trajectory - HMTRL

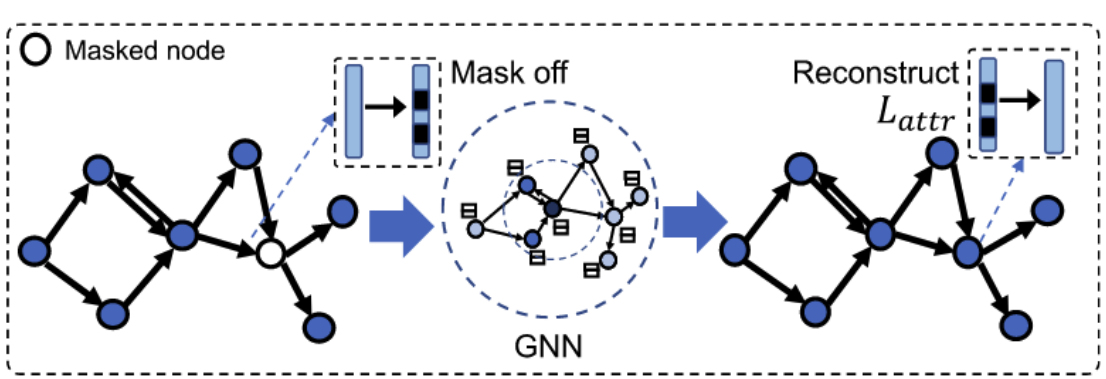
- Method:
  - Time-dependent Multi-view Transportation Graphs
  - Hierarchical Multi-Task Route Representation Learning



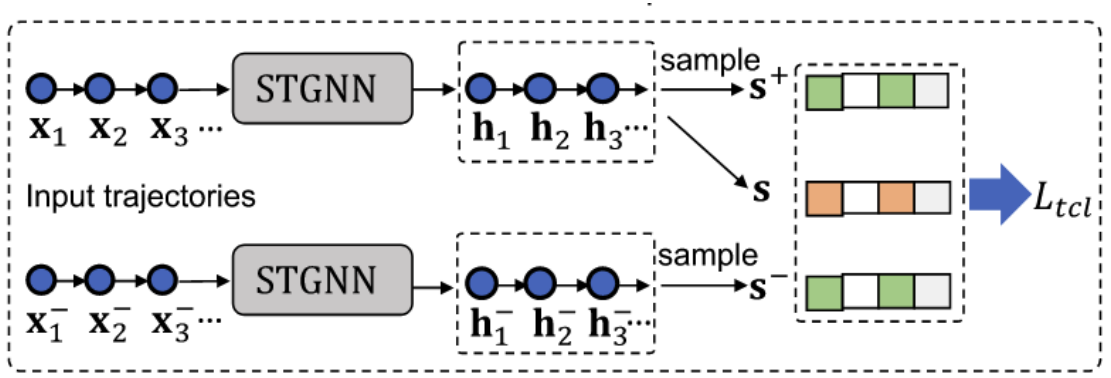
Liu, Hao, et al. Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training. VLDB 2023.

Pre-training on Road Network Trajectory - HMTRL

- Spatiotemporal Pre-train
  - Randomly mask node attributes in the transportation network graph, and predict masked features based on the observed attributes and graph structure.
  - Contrast the trajectory and its local parts based on mutual information maximization to maximize the agreement between nodes within the same trajectory while minimizing the agreement with nodes from other trajectories.



(a) Masked attribute prediction

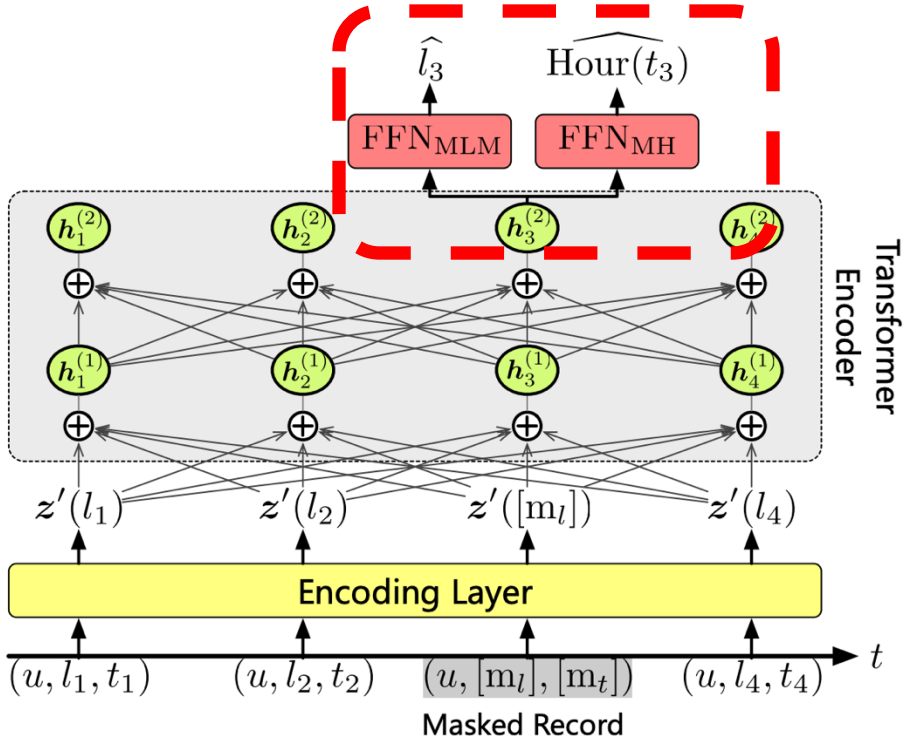


(b) Trajectory Contrastive Learning

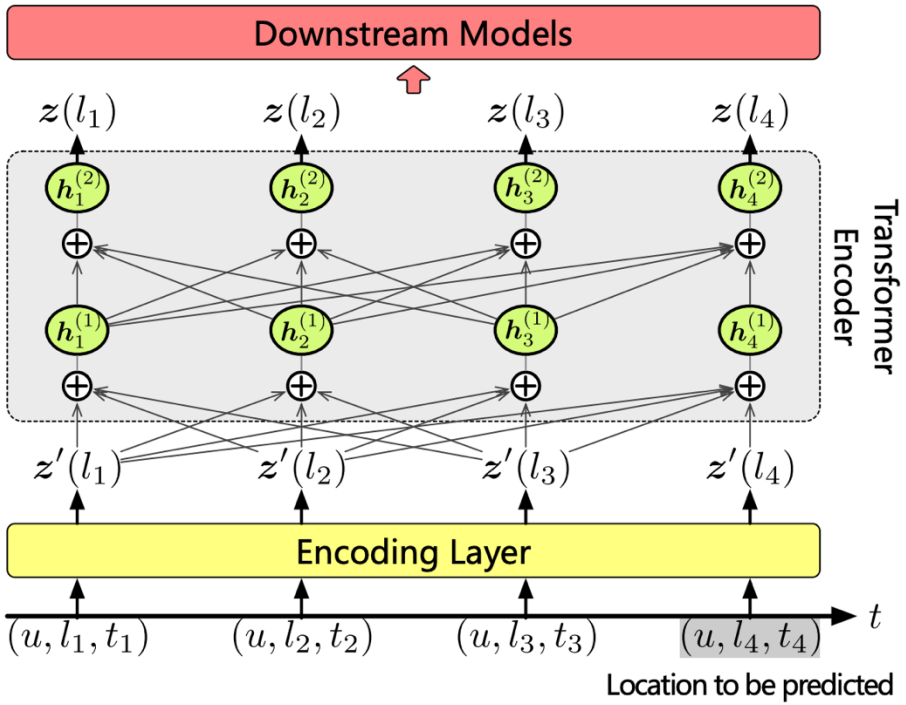
Liu, Hao, et al. Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training. VLDB 2023.

Pre-training on Free Space Trajectory - CTLE

- Generates embeddings based on the sequence of nearby visited locations of each location within a trajectory.
  - Masked Location Prediction
  - Masked Hour Prediction



(b) Pre-training



(c) Calculating contextual embeddings

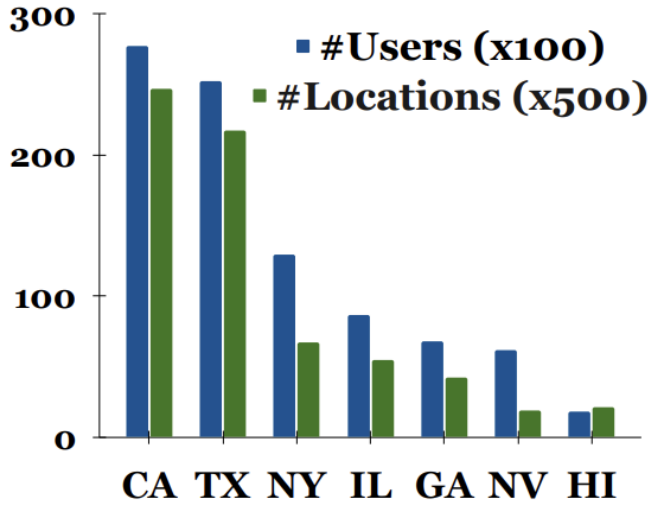
Lin, Yan, et al. Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction. AAAI 2021.

Trajectory-based UFMs

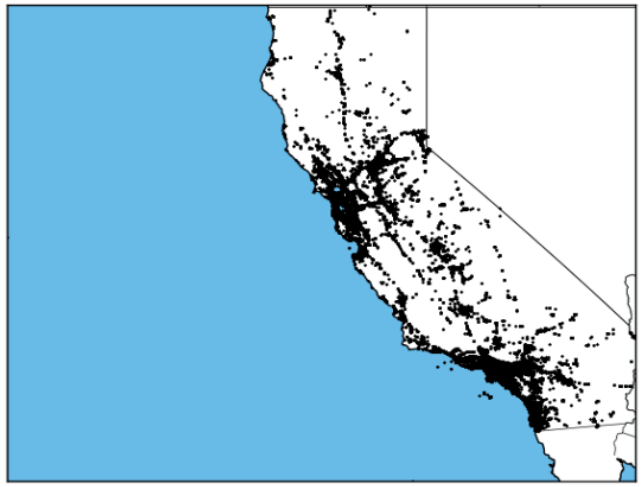
- **Trajectory-based Pre-training**
  - Road network trajectory
  - Free space trajectory
- **Trajectory-based Adaptation**
  - *Model fine-tuning*
- **Cross-modal Adaptation**
  - Prompt engineering

Model Fine-tuning - Axolotl

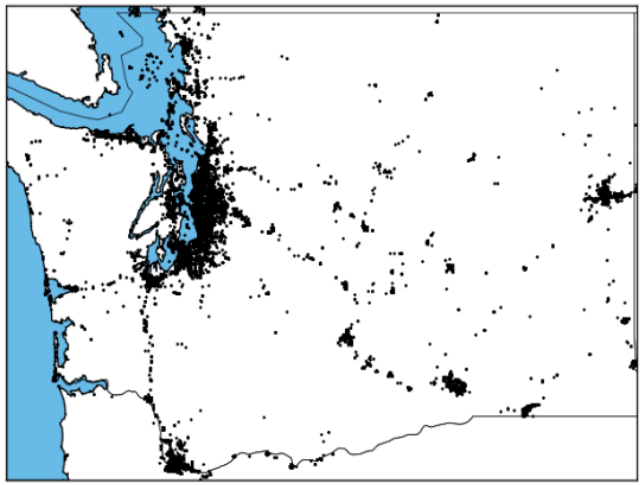
- Directly addresses the challenge of data scarcity for POI recommendation through fine-tuning.
- Fine-tuning in data-scarce regions involves minimal adjustments to the model parameters learned from the source region, with adapting these pre-trained parameters to better suit the target region's unique user-location dynamics.



(a) State-wise Data



(b) California

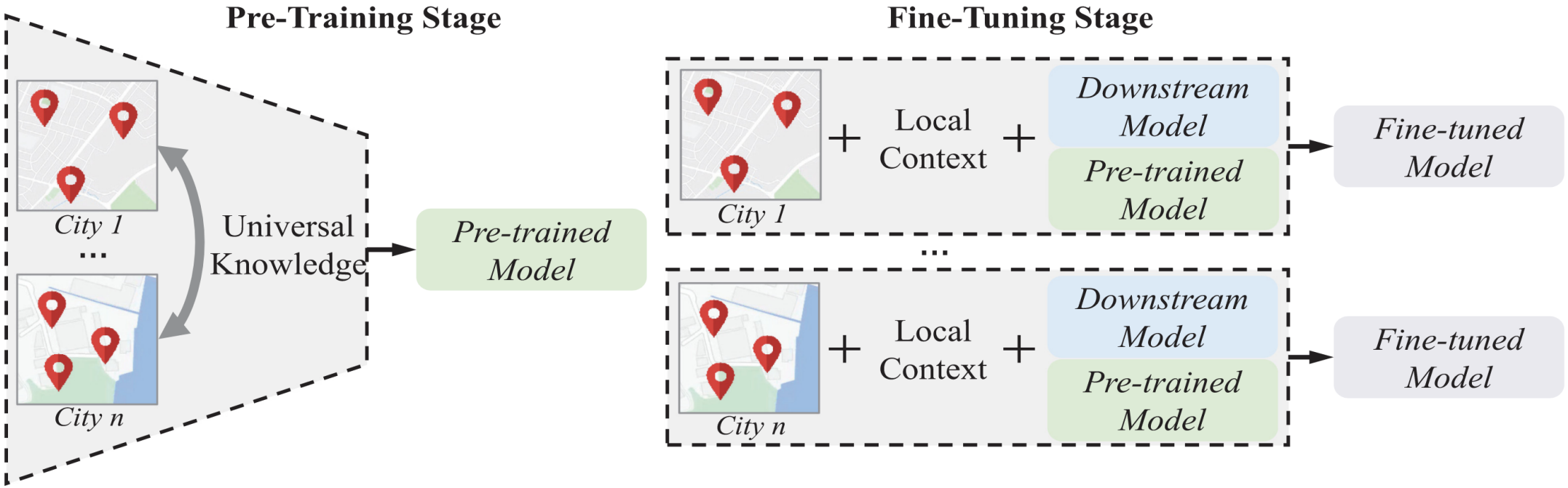


(c) Washington

Vinayak Gupta, et, all. Doing more with less: Overcoming data scarcity for POI recommendation via cross-region transfer. ACM TIST, 2022.

Model Fine-tuning - CATUS

- The need to address the local specificity of POI transitions.
- Fine-tuning in data-scarce regions involves minimal adjustments to the model parameters learned from the source region, with adapting these pre-trained parameters to better suit the target region's unique user-location dynamics.



Sun, Ke, et al. Pre-Training Across Different Cities for Next POI Recommendation. ACM Transactions on the Web, 2023.

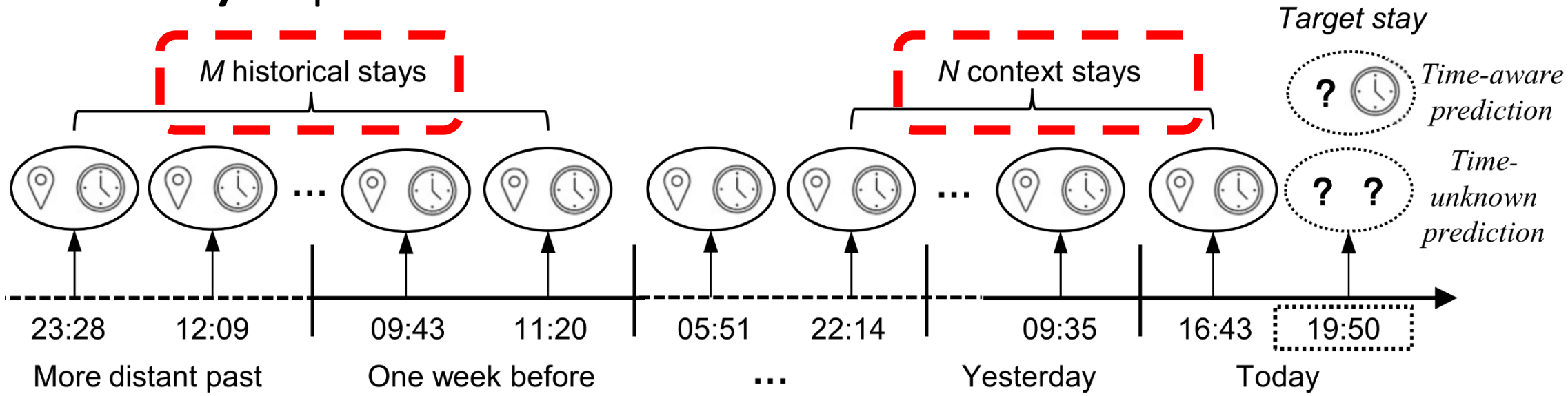
Trajectory-based UFMs

- **Trajectory-based Pre-training**
  - Road network trajectory
  - Free space trajectory
- **Trajectory-based Adaptation**
  - Model fine-tuning
- **Cross-modal Adaptation**
  - *Prompt engineering*

Prompt Engineering – LLM-Mob

- Data formatting process: transform mobility data into historical stays and context stays.

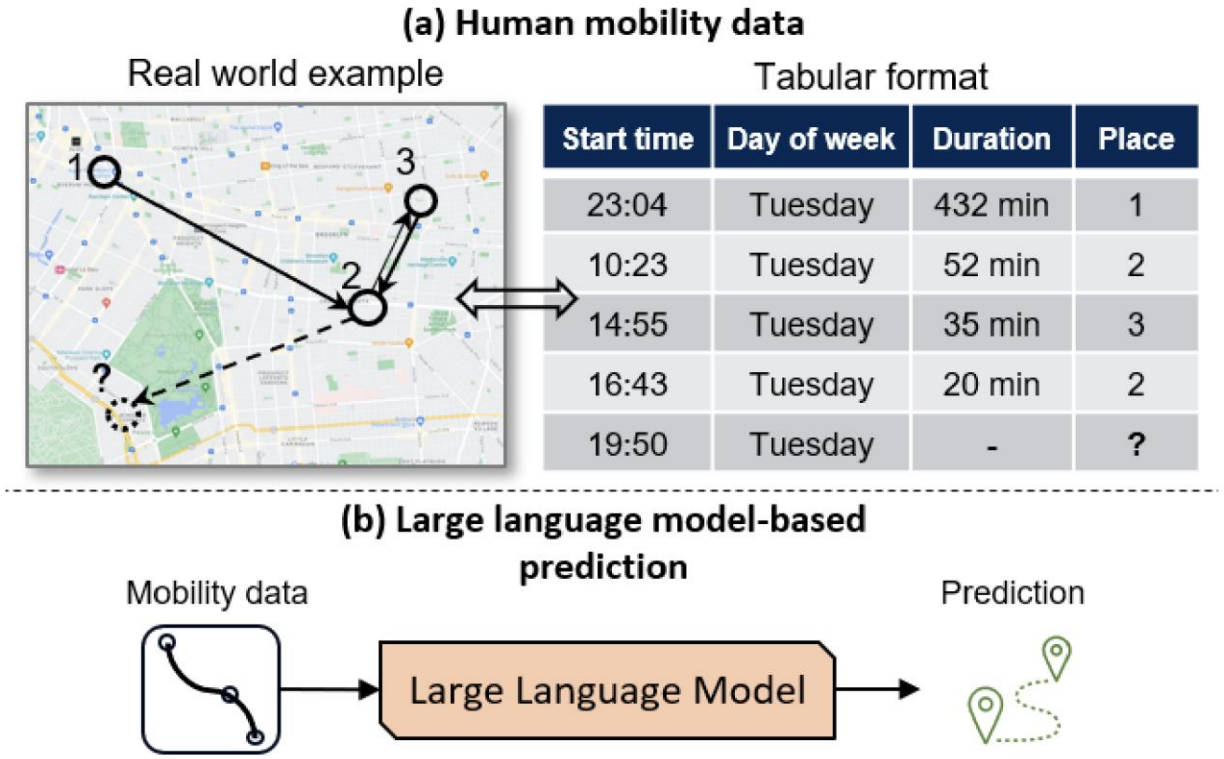
- **Historical stays** reflect long-term movement patterns (e.g., regular visits to work on weekdays)
- **Context stays** represent more immediate and recent movements.

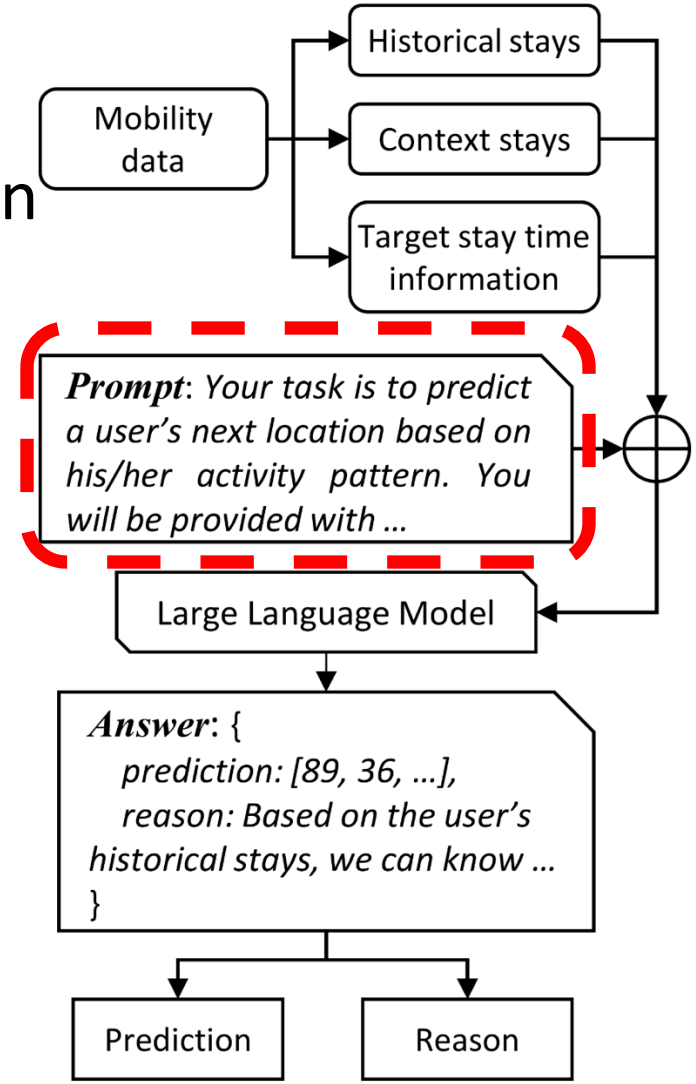


Wang, Xinglei, et al. Where would i go next? large language models as human mobility predictors. arXiv 2023.

Prompt Engineering – LLM-Mob

- Prompts include both the raw data and explicit instructions that encourage the model to "think" about both the long-term and short-term patterns in the data.

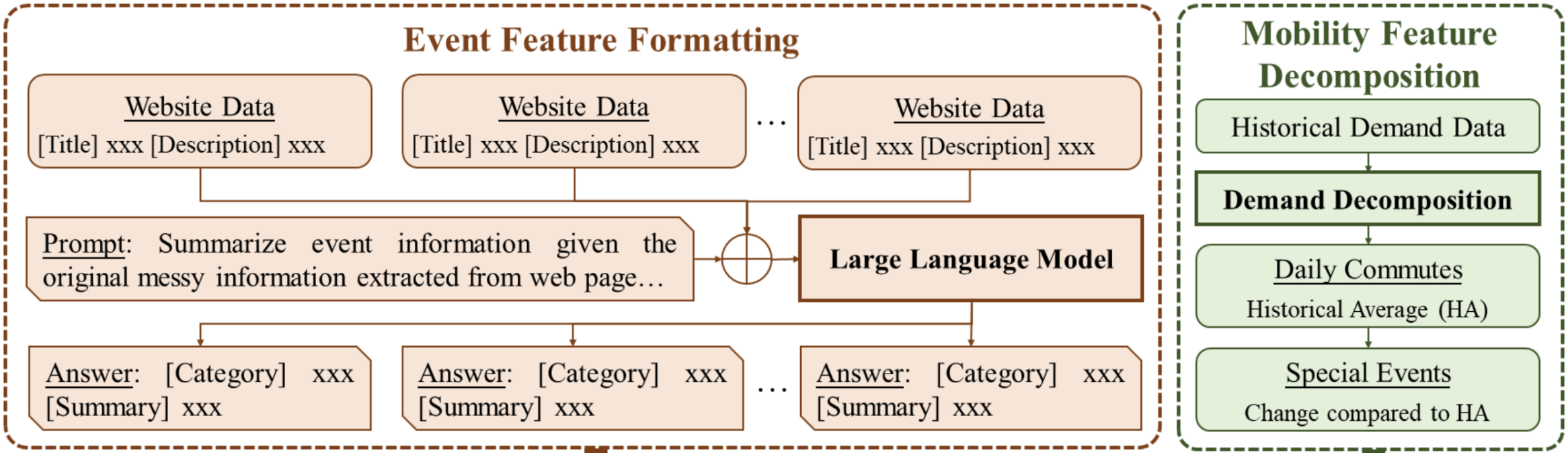




Wang, Xinglei, et al. Where would i go next? large language models as human mobility predictors. arXiv 2023.

Prompt Engineering – LLM-MPE

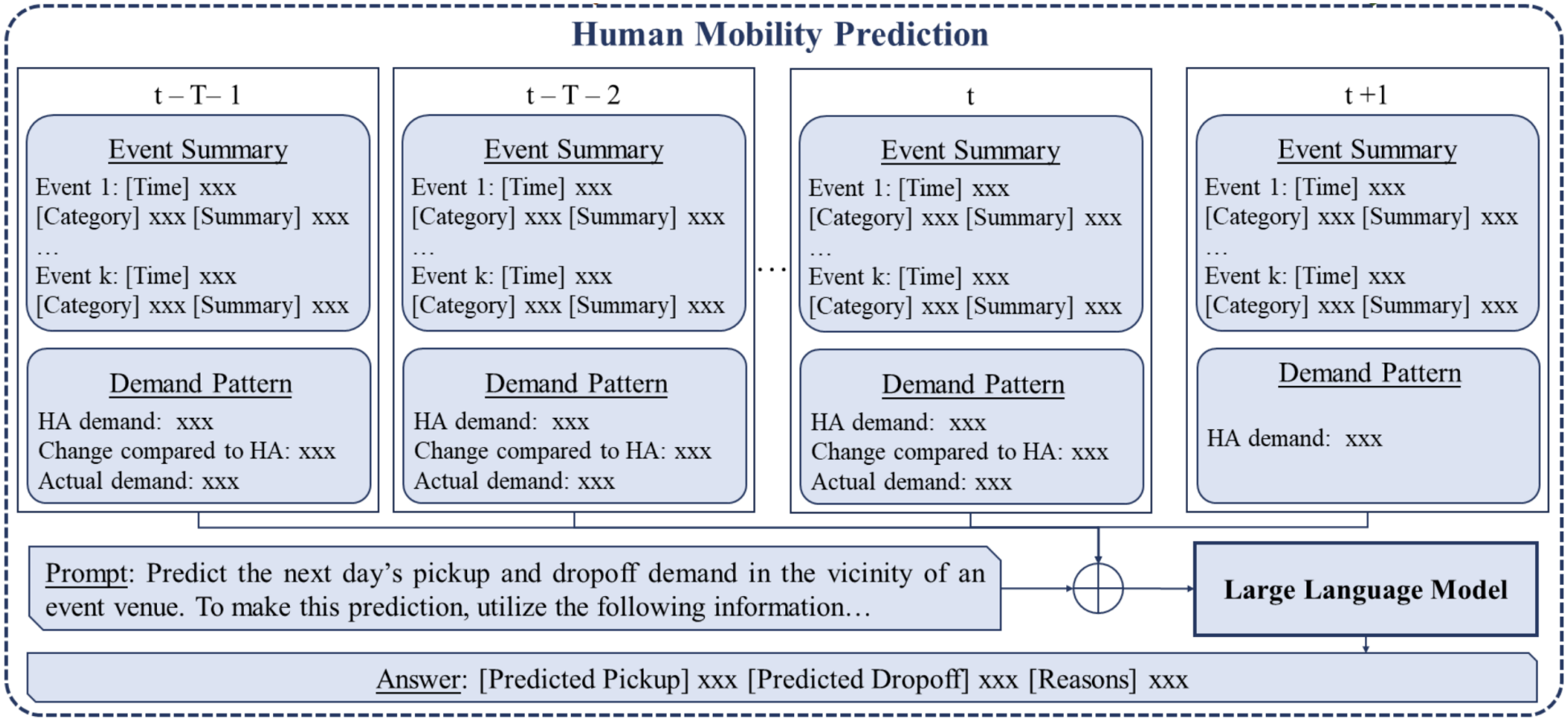
- Prompt LLM to process unstructured event descriptions from online sources and convert them into a standardized, concise format that can be easily understood by the LLM.
- Separates human mobility patterns into regular (commuting-related) and irregular (event-induced) components.



Liang, Yuebing, et al. Exploring large language models for human mobility prediction under public events. arXiv 2023.

Prompt Engineering – LLM-MPE

- Use carefully designed prompts and chain-of-thought prompting strategy to guide the LLM in predicting travel demand by incorporating both historical mobility data and the upcoming event features.



Liang, Yuebing, et al. Exploring large language models for human mobility prediction under public events. arXiv 2023.

Summary & Future Directions

- Summary
  - Existing unimodal studies primarily focus on addressing specific trajectory-related tasks, which limits their effectiveness in handling more complex or diverse urban scenarios.
  - Current adaptation methods also struggle with generalization, making it difficult for them to perform well across a wide range of downstream tasks.

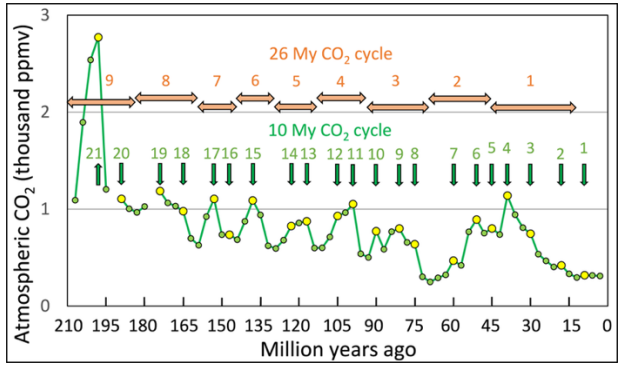
Summary & Future Directions

- Future Directions
  - Build foundation models by pre-training on extensive trajectory data to enhance generalizability, followed by fine-tuning across a wide range of trajectory-related tasks.
  - Develop methodologies for integrating trajectory data with other urban modalities. By effectively fusing trajectory data with these modalities, we can deepen the model's understanding of urban environments.

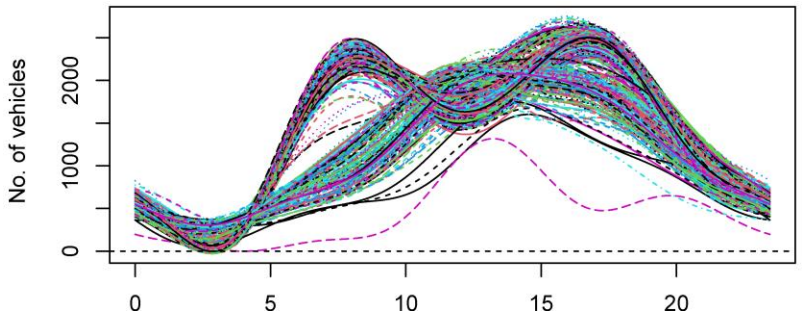
Time Series-based UFMs

Time Series Data

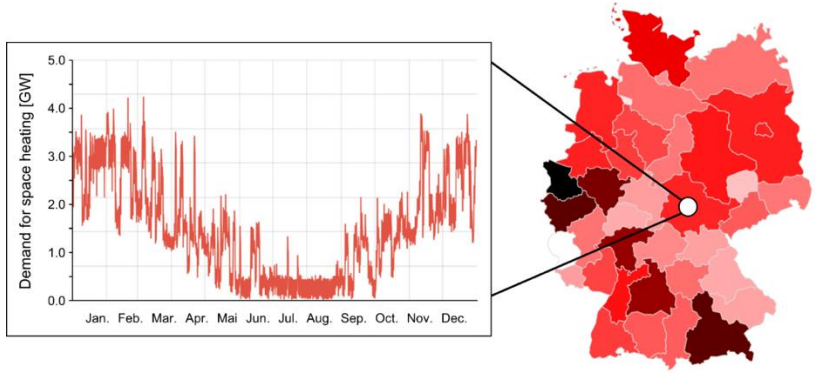
- Time series data in various urban scenarios.



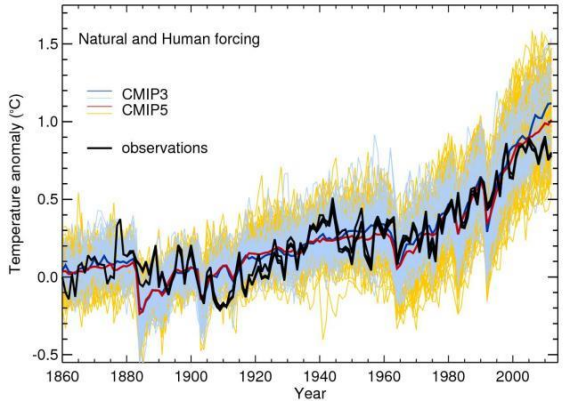
Air quality



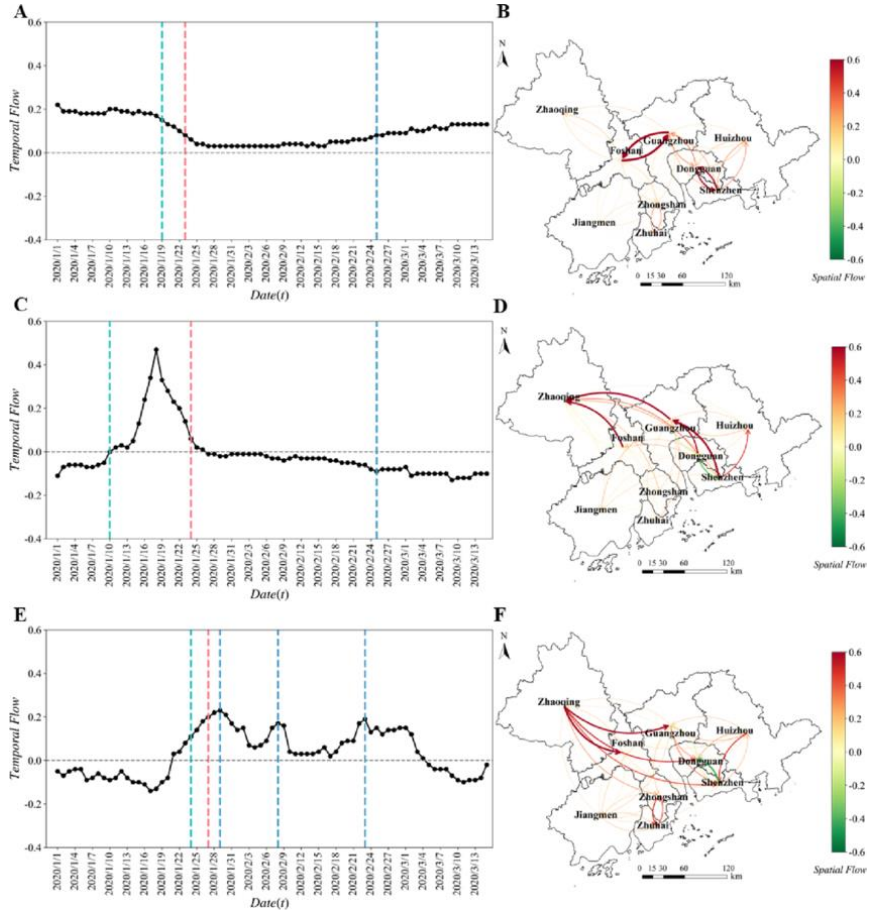
Traffic flow



Energy Demand



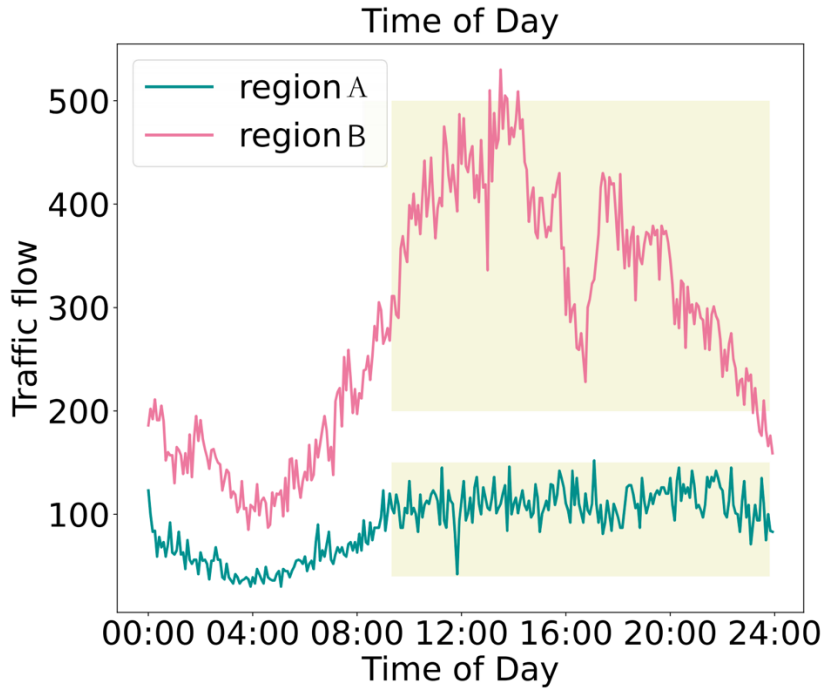
Weather



Human mobility

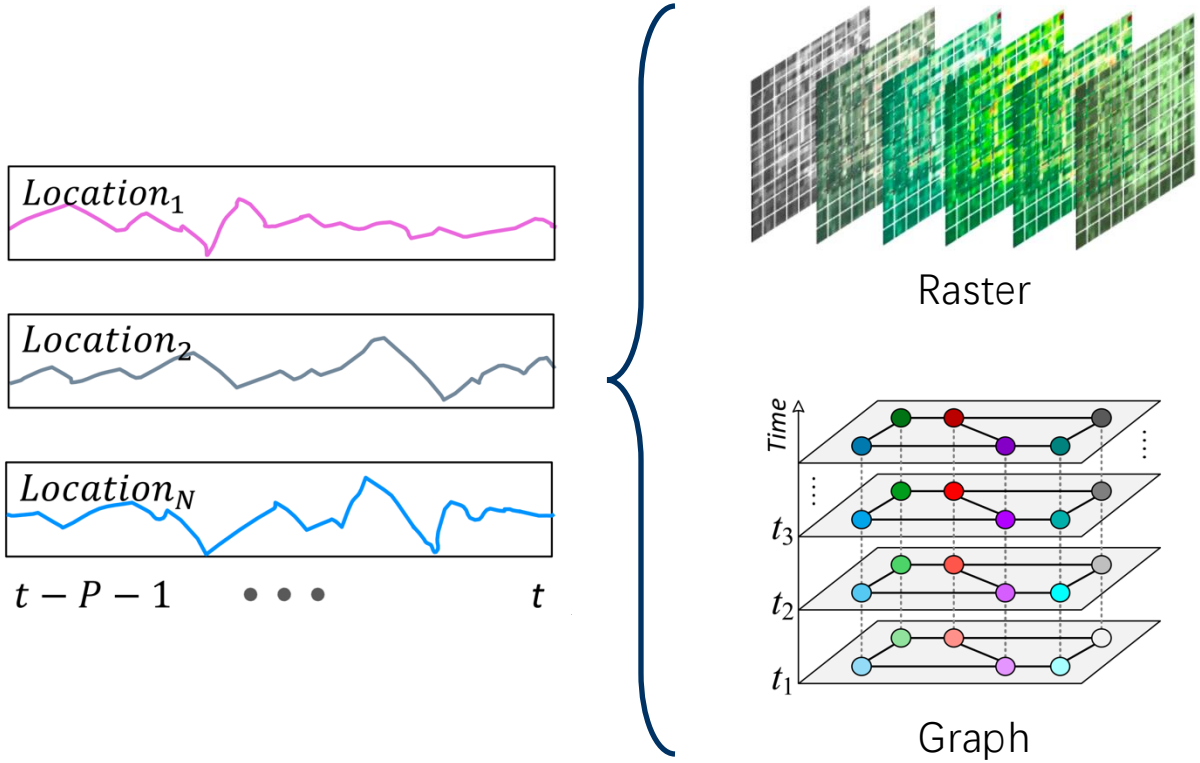
Time Series Data

- Time series is the collection of observable records in an ordered manner, representing the evolution of the variable states in the real world.



$X \in R^{T \times C}$ T - Temporal factor
C - Variable factor

Ordinary time series

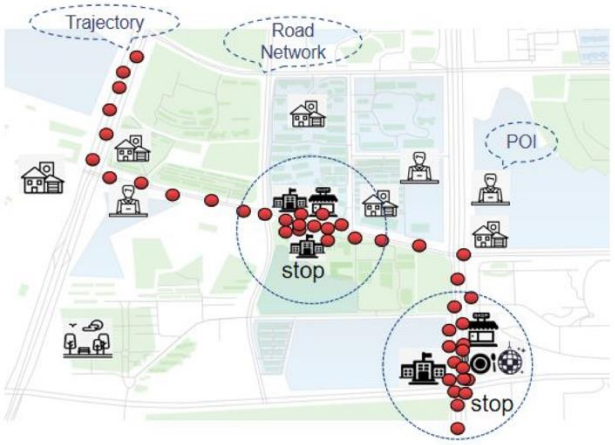


$X \in R^{T \times N \times C}$ N - Spatial factor

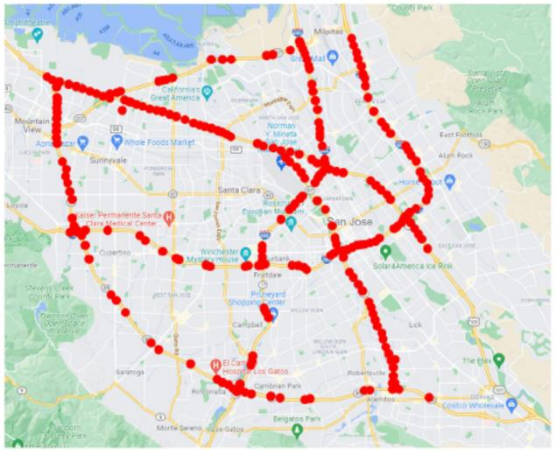
Spatial-correlated time series

Limitations of Trajectory-based UFMs

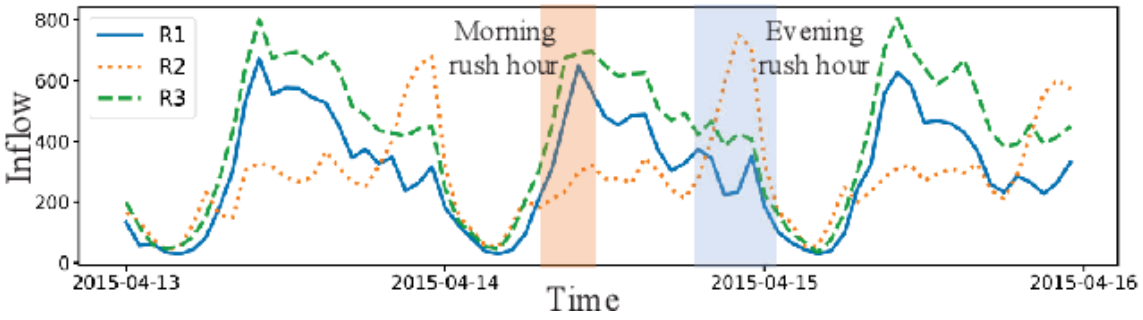
- Why are time series-based urban foundation models necessary?
  - Trajectory data capture the movements of objects through space over time.
  - Time series data represent the observations at specific locations over time.
  - The static spatial characteristic necessitates the foundation model tailored for time series data.



Trajectory movements



Geographical distribution of sensors



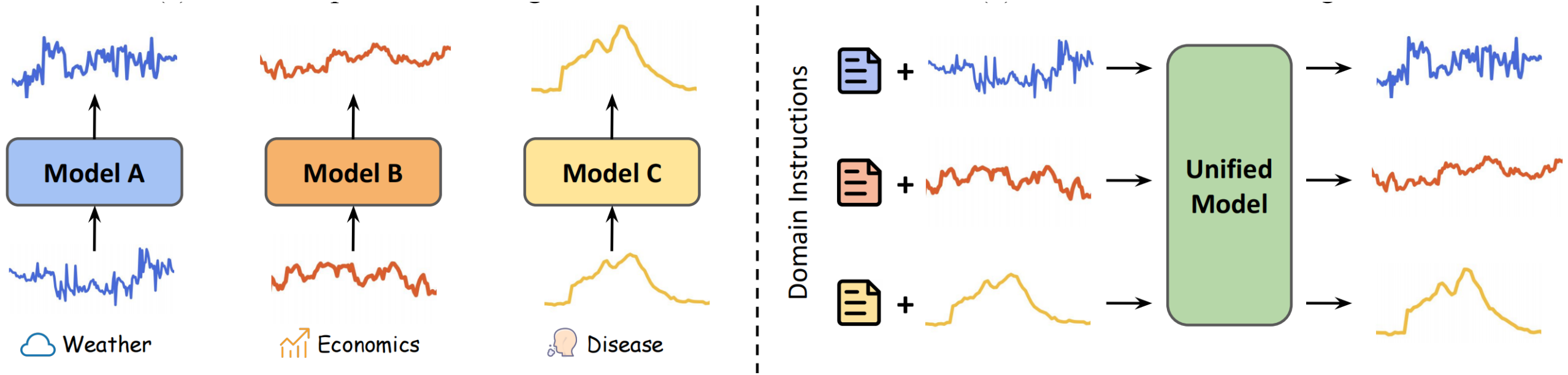
Time series data observed at specific sensors

Time Series-based UFMs

- **Time Series-based Pre-training**
  - *Ordinary time series*
  - *Spatial-correlated time series*
- **Time Series-based Adaptation**
  - Prompt tuning
- **Cross-modal Adaptation**
  - Prompt engineering
  - Model fine-tuning
  - Model reprogramming

Ordinary Time Series

- UniTime
  - Pre-training a unified model generalizing across various domains.



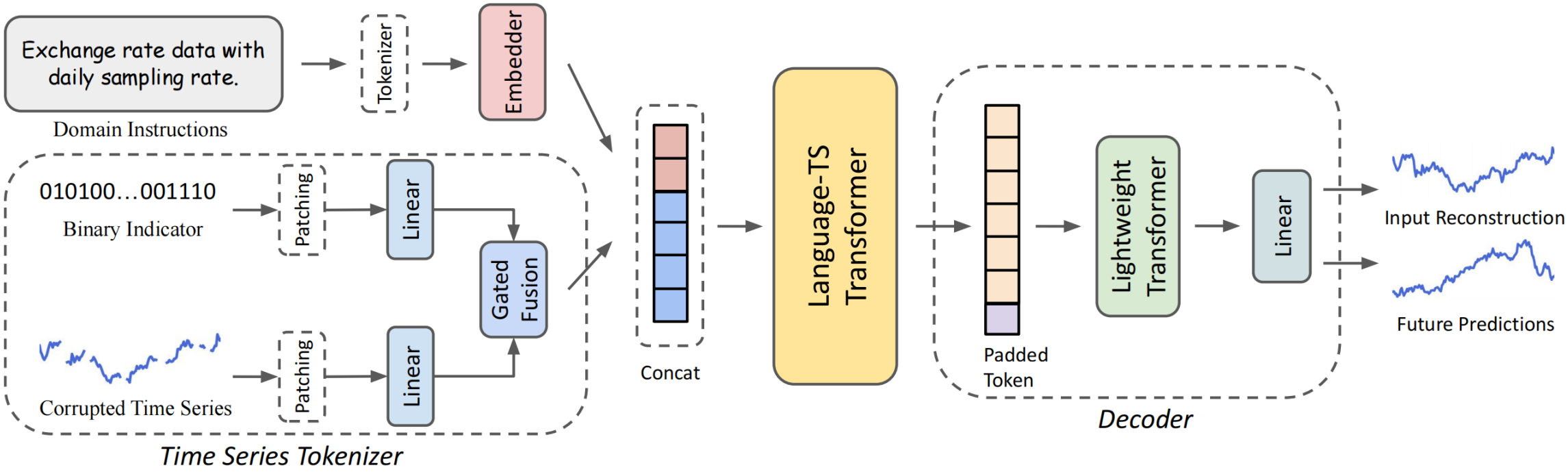
Domain-specific learning

Cross-domain learning

Liu, Xu, et al. Unitime: A language-empowered unified model for cross-domain time series forecasting. WWW,

Ordinary Time Series

- UniTime
  - Domain-specific instructions to solve varied domain contexts.
  - Binary masking to balance convergence rates of various domains.
  - Token padding to accommodate variations in token lengths.

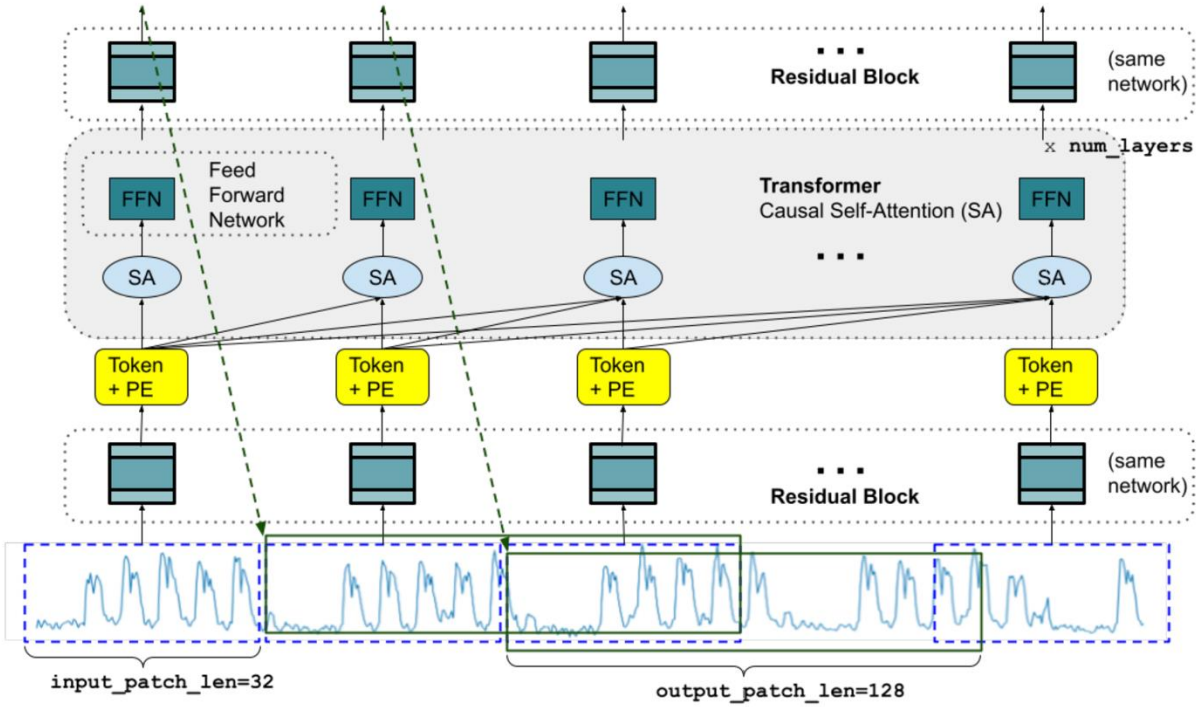


UniTime overview from the perspective of a univariate time series.

Liu, Xu, et al. Unitime: A language-empowered unified model for cross-domain time series forecasting. WWW, 2024.

Ordinary Time Series

- TimesFM
  - A large-scale time-series corpus, with both real-world and synthetic data.
  - A decoder-only attention architecture with input patching.
  - Close to state-of-the-art zero-shot forecasting accuracy.



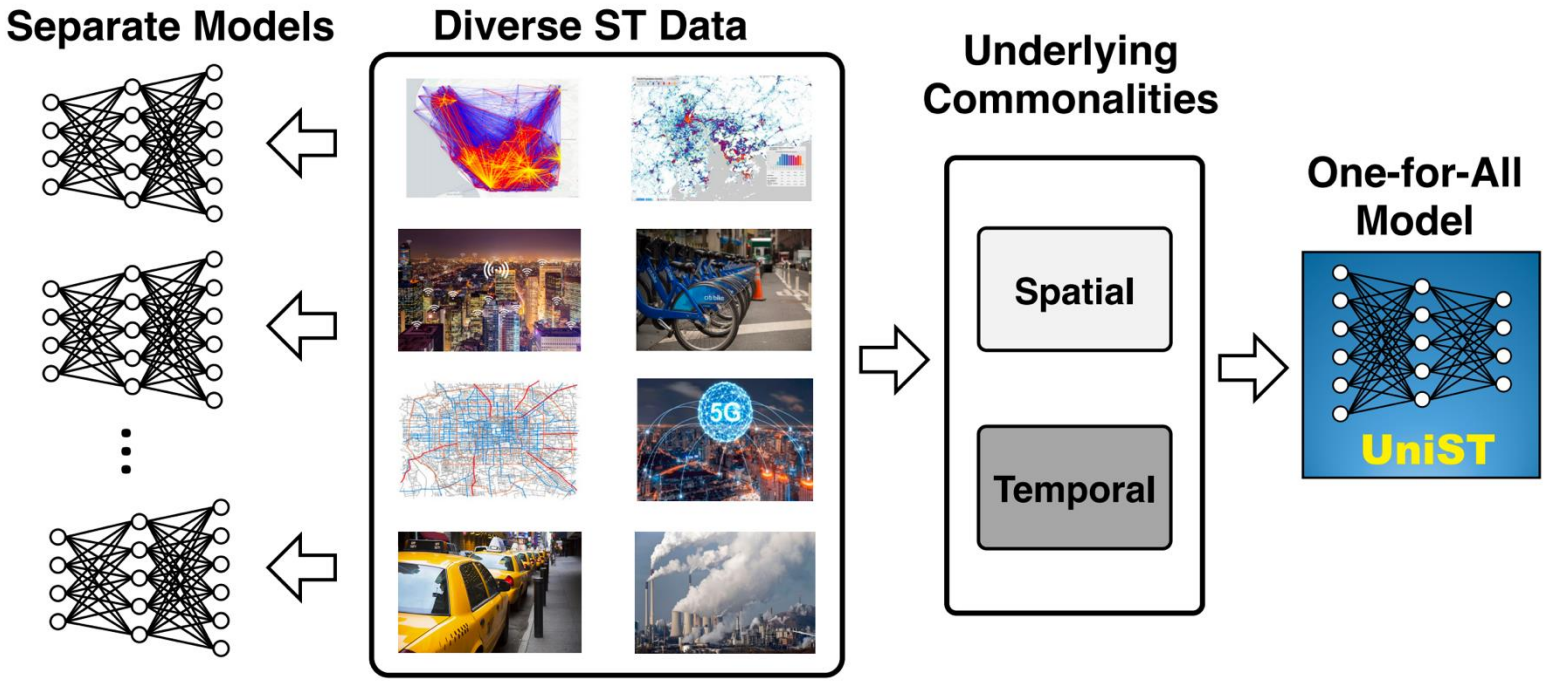
TimesFM model architecture during training.

- ✓ Longer output patches for better accuracy and efficiency.
- ✓ Patch masking to accommodate all possible context lengths.

Das, Abhimanyu, et al. A decoder-only foundation model for time-series forecasting. ICML, 2024.

Spatial-Correlated Time Series

- Two capabilities of a universal spatio-temporal model.
  - Leverage abundant and rich data from urban scenarios for training.
  - Robust generalization across diverse spatio-temporal scenarios.



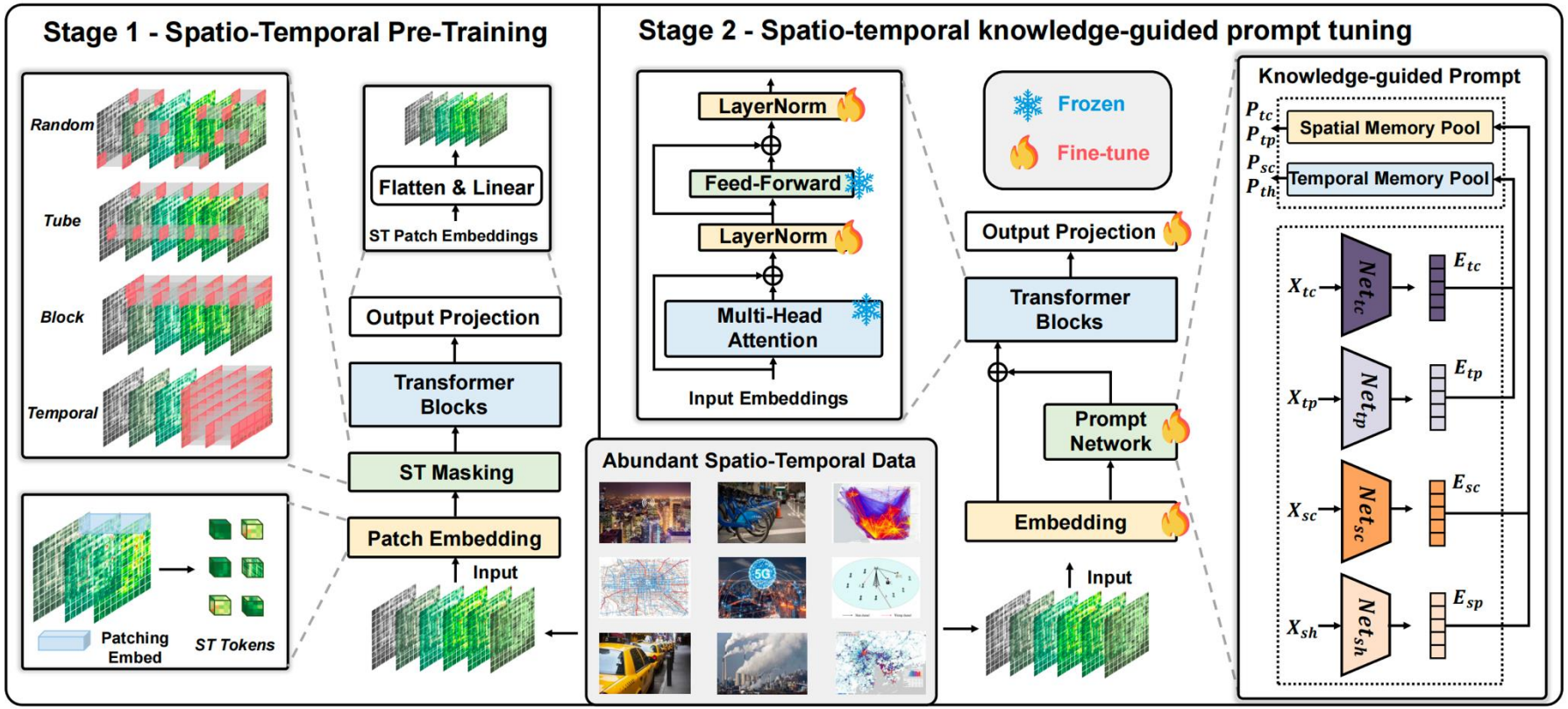
- Challenges
  - ✓ Diverse data formats.
  - ✓ Varied data distributions.

From separate models to a one-for-all universal model for urban spatio-temporal prediction.

Yuan, Yuan, et al. UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. KDD, 2024.

Spatial-Correlated Time Series

- UniST - Pre-training
  - Patch and transform 3D data into 1D sequential data.
  - Distinct masking strategies from both spatial and temporal perspectives.



UniST overview.

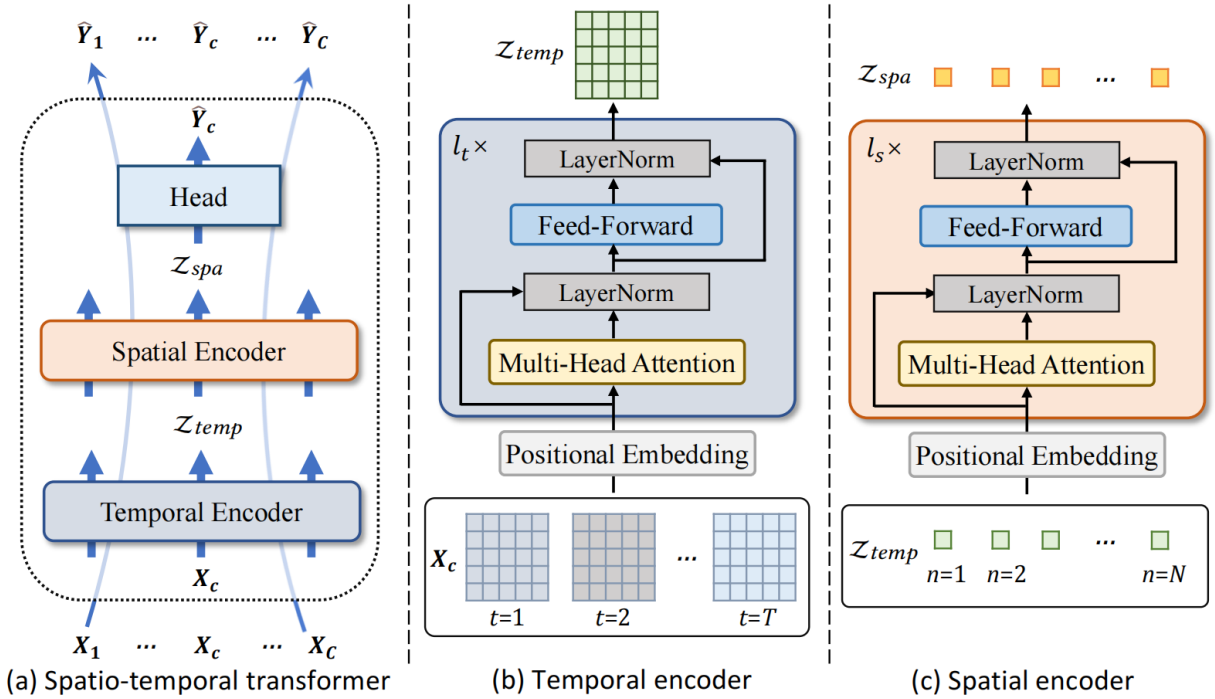
Yuan, Yuan, et al. UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. KDD, 2024.

Time Series-based UFMs

- **Time Series-based Pre-training**
  - Ordinary time series
  - Spatial-correlated time series
- **Time Series-based Adaptation**
  - *Prompt tuning*
- **Cross-modal Adaptation**
  - Prompt engineering
  - Model fine-tuning
  - Model reprogramming

Prompt Tuning

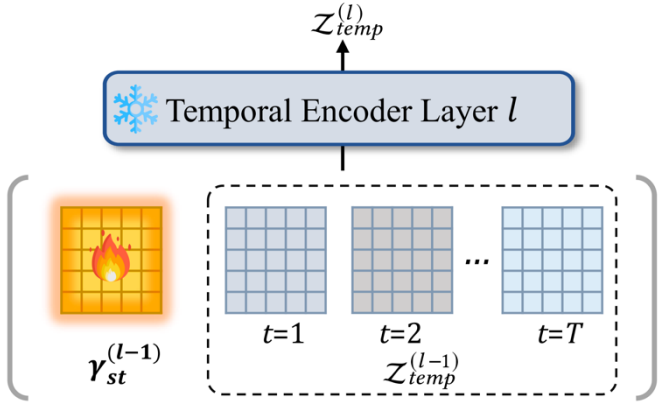
- PromptST
  - Pre-train a light-weight spatio-temporal transformer.
  - Freeze the pre-trained model and fine-tune the inserted prompts.



(a) Spatio-temporal transformer (b) Temporal encoder (c) Spatial encoder

Pre-training phase.

✓ Add learnable prompts to the intermediate temporal features along the temporal dimension.

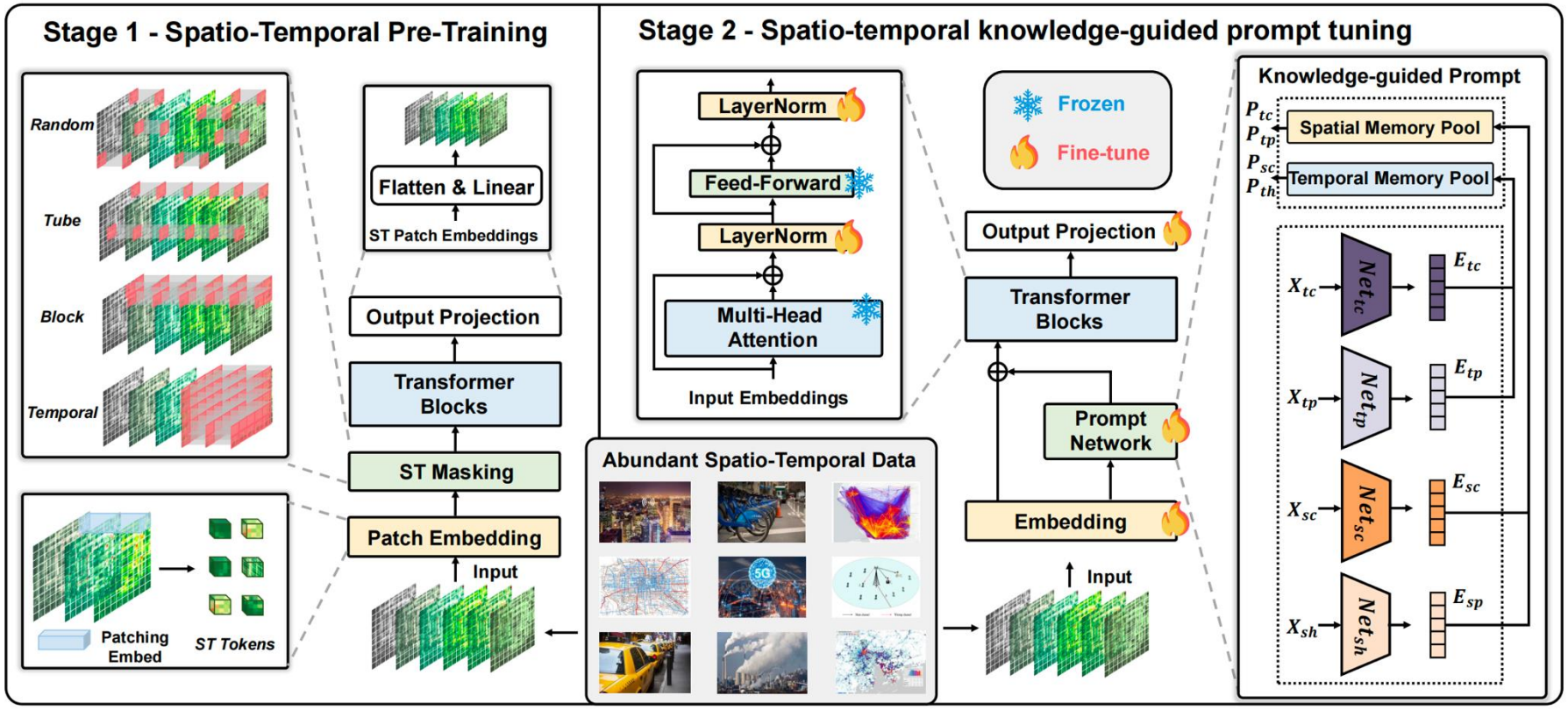


Prompt tuning phase.

Zhang, Zijian, et al. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. CIKM, 2023.

Prompt Tuning

- UniST - Adaptation (Prompt tuning)
  - Learn a spatial and a temporal memory pool with key-value structured parameters.
  - Learn query representations with four aspects of spatio-temporal domain knowledge.
  - Extract knowledge-guided prompts by querying the prompt memory pools.



UniST overview.

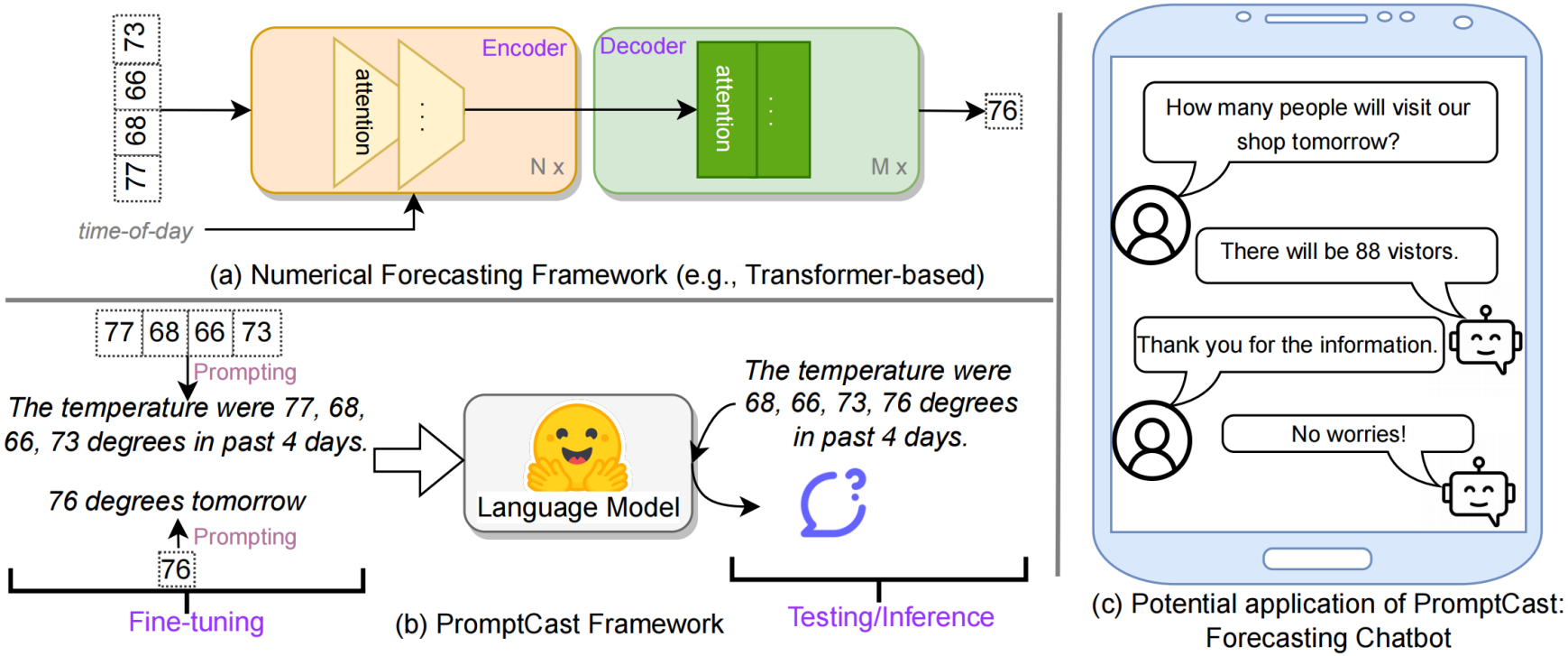
Yuan, Yuan, et al. UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. KDD, 2024.

Time Series-based UFMs

- **Time Series-based Pre-training**
  - Ordinary time series
  - Spatial-correlated time series
- **Time Series-based Adaptation**
  - Prompt tuning
- **Cross-modal Adaptation**
  - *Prompt engineering*
  - *Model fine-tuning*
  - *Model reprogramming*

Prompt Engineering

- PromptCast
  - The first prompt-based time series forecasting paradigm, which is formulated as a question-answering task.



Xue, Hao, and Flora D. Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. TKDE, 2023.

Prompt Engineering

- PromptCast
  - The first prompt-based time series forecasting paradigm, which is formulated as a question-answering task.

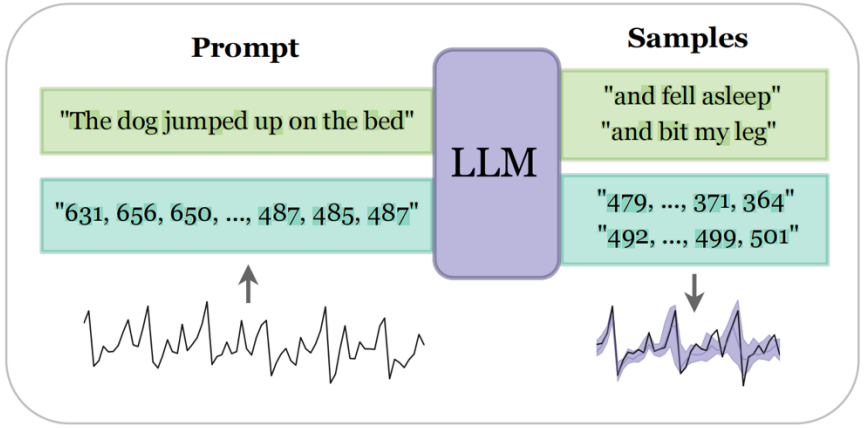
			Template	Example
CT	Input Prompt (Source)	Context	From $\{t_1\}$ to $\{t_{\text{obs}}\}$, the average temperature of region $\{U_m\}$ was $\{x^m_{t_1:t_{\text{obs}}}\}$ degree on each day.	From August 16, 2019, Friday to August 30, 2019, Friday, the average temperature of region 110 was 78, 81, 83, 84, 84, 82, 83, 78, 77, 77, 74, 77, 78, 73, 76 degree on each day.
		Question	What is the temperature going to be on $\{t_{\text{obs}+1}\}$?	What is the temperature going to be on August 31, 2019, Saturday?
	Output Prompt (Target)	Answer	The temperature will be $\{x^m_{t_{\text{obs}+1}}\}$ degree.	The temperature will be 78 degree.
ECL	Input Prompt (Source)	Context	From $\{t_1\}$ to $\{t_{\text{obs}}\}$, client $\{U_m\}$ consumed $\{x^m_{t_1:t_{\text{obs}}}\}$ kWh of electricity on each day.	From May 16, 2014, Friday to May 30, 2014, Friday, client 50 consumed 8975, 9158, 8786, 8205, 7693, 7419, 7595, 7596, 7936, 7646, 7808, 7736, 7913, 8074, 8329 kWh of electricity on each day.
		Question	What is the consumption going to be on $\{t_{\text{obs}+1}\}$?	What is the consumption going to be on May 31, 2014, Saturday?
	Output Prompt (Target)	Answer	This client will consume $\{x^m_{t_{\text{obs}+1}}\}$ kWh of electricity.	This client will consume 8337 kWh of electricity.
SG	Input Prompt (Source)	Context	From $\{t_1\}$ to $\{t_{\text{obs}}\}$, there were $\{x^m_{t_1:t_{\text{obs}}}\}$ people visiting POI $\{U_m\}$ on each day.	From May 23, 2021, Sunday to June 06, 2021, Sunday, there were 13, 17, 13, 20, 16, 16, 17, 17, 19, 20, 12, 12, 14, 12, 13 people visiting POI 324 on each day.
		Question	How many people will visit POI $\{U_m\}$ on $\{t_{\text{obs}+1}\}$?	How many people will visit POI 324 on June 07, 2021, Monday?
	Output Prompt (Target)	Answer	There will be $\{x^m_{t_{\text{obs}+1}}\}$ visitors.	There will be 15 visitors.

Templates and examples of prompts in urban scenarios.

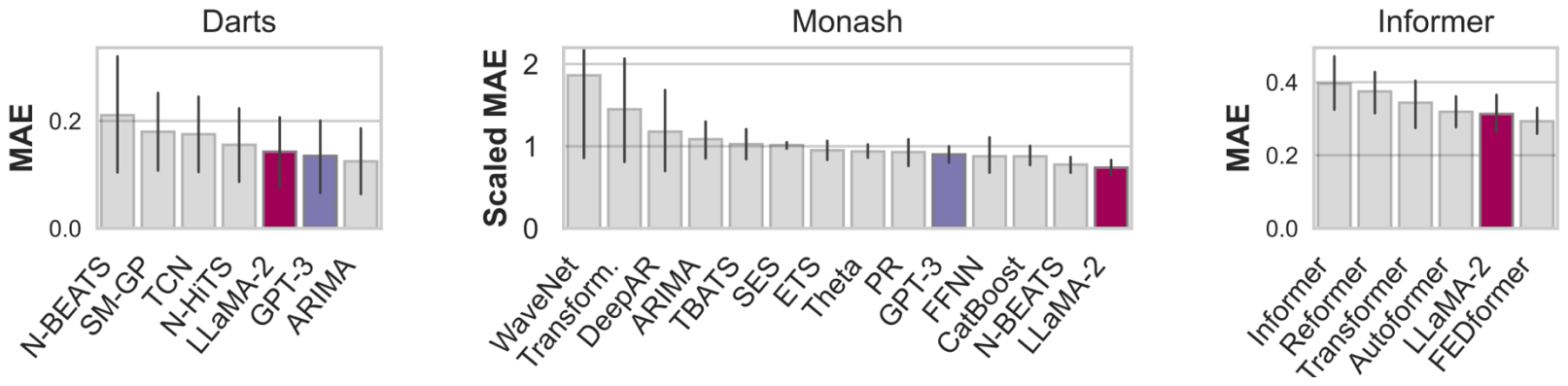
Xue, Hao, and Flora D. Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. TKDE, 2023.

Prompt Engineering

- LLMTime
  - LLMs can be used directly as zero-shot forecasters without any added text prompts, if the input time series data are carefully preprocessed.
  - Deterministic predictions from LLMTime are ranked best or second best on all the considered benchmarks while having no trainable parameters.



LLMTime forecasting paradigm.

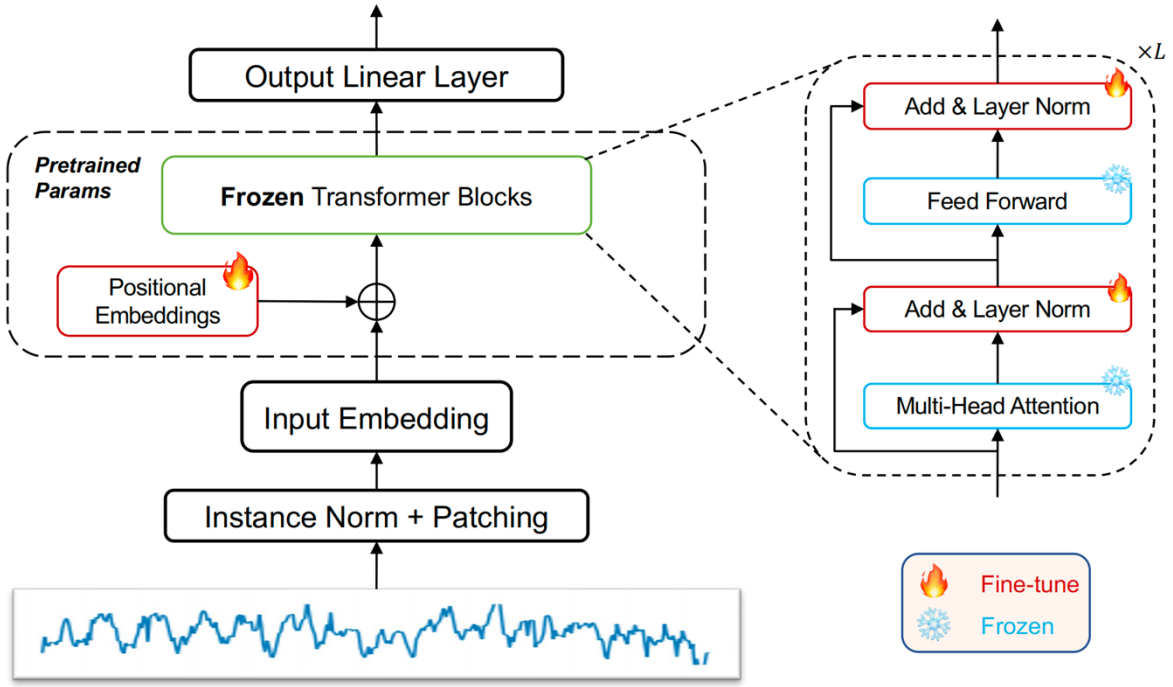


Forecasting performances of LLMTime with base model GPT-3 or LLaMA-2 70B.

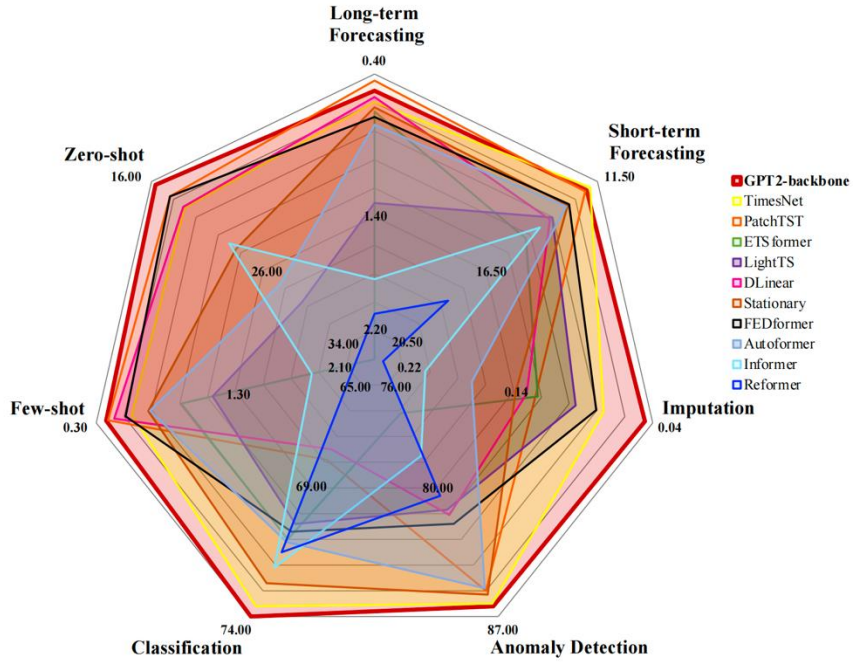
Gruver, Nate, et al. Large language models are zero-shot time series forecasters. NIPS, 2024.

Model Fine-tuning

- GPT2-backboned Frozen Pre-trained Transformer (FPT)
  - Leverage the pre-trained LLM for time series analysis.
  - Freeze self-attention and FFN layers and only fine-tune positional embedding and normalization layers.
  - Superior performance on various time series tasks.



Model architecture.

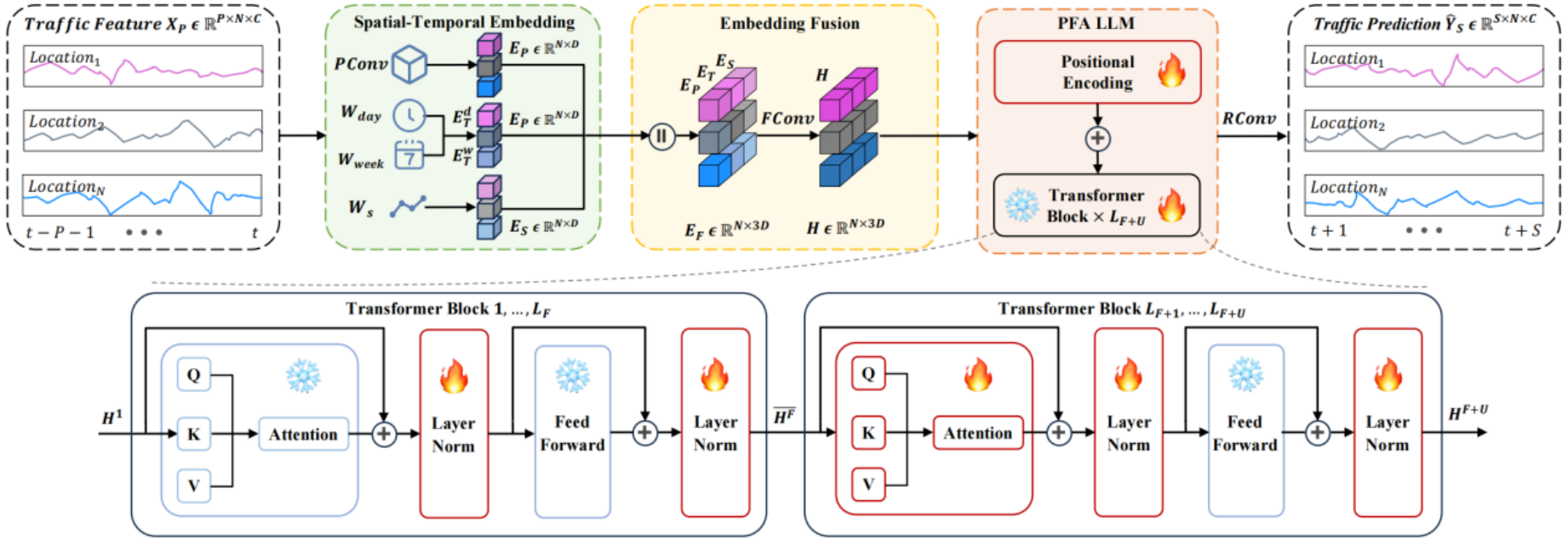


Model performance comparison on various tasks.

Zhou, Tian, et al. One fits all: Power general time series analysis by pretrained lm. NIPS, 2023.

Model Fine-tuning

- ST-LLM
  - Leverage the pre-trained LLM for traffic prediction.
  - Transform traffic features into location-specific embeddings.
  - Unfreeze the last several multi-head attention layers in the LLM.

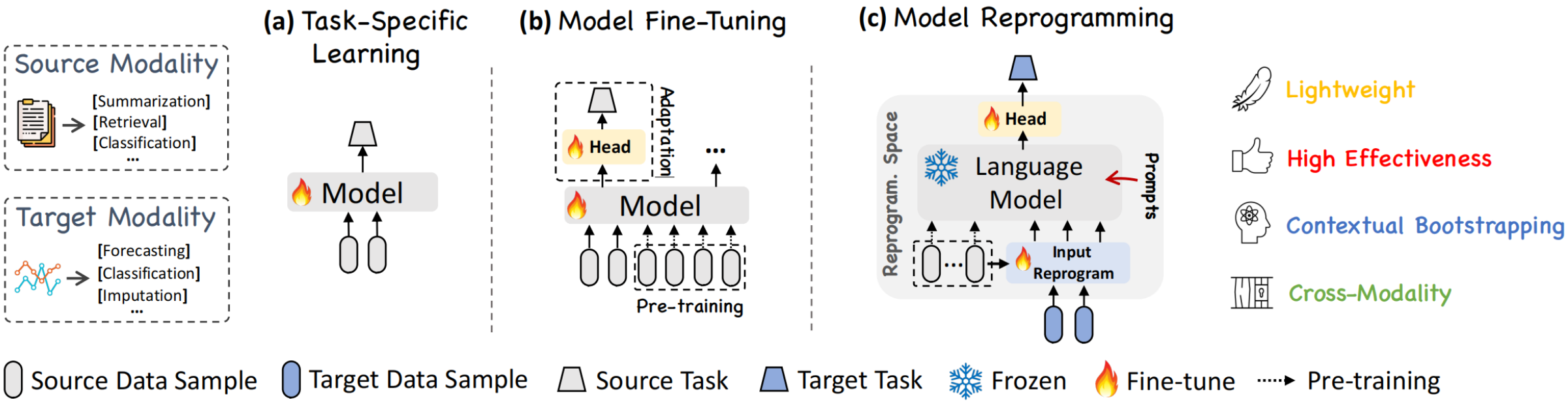


ST-LLM framework.

Liu, Chenxi, et al. Spatial-temporal large language model for traffic prediction. MDM, 2024.

Model Reprogramming

- Time-LLM
  - A reprogramming framework that aligns the modalities of time series and natural language.

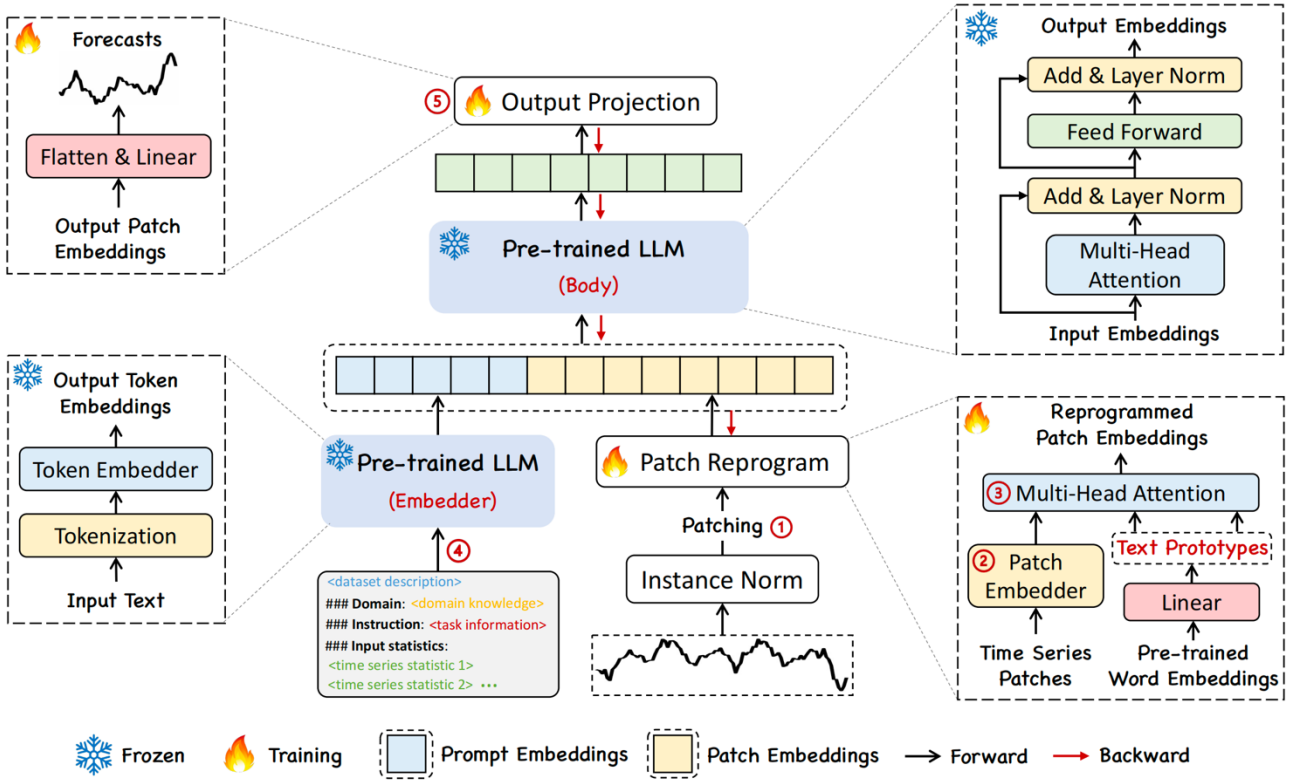


Comparison of three kinds of time series forecasting frameworks.

Jin, Ming, et al. Time-llm: Time series forecasting by reprogramming large language models. ICLR, 2024.

Model Reprogramming

- Time-LLM
  - Reprogram the time series patches into text prototype representations.
  - Add dataset description prompts as the prefix.



Time-LLM overview.

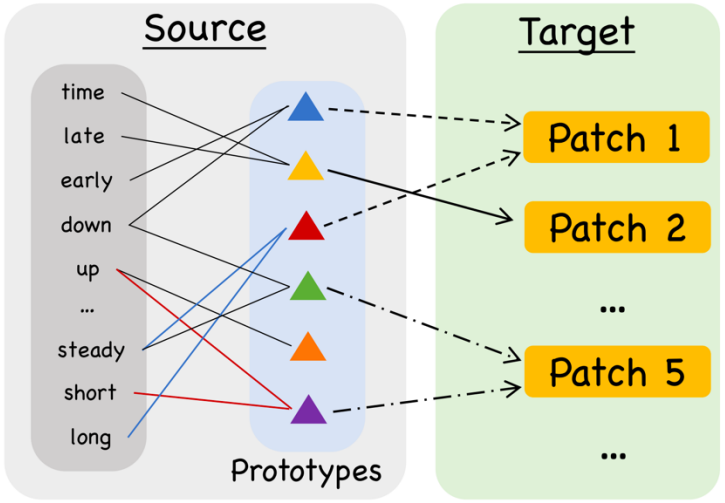


Illustration of Patch Programming.

The Electricity Transformer Temperature (ETT) indicates the electric power long-term deployment. Each data point consists of the target oil temperature and 6 power load features ... Below is the information about the input time series:

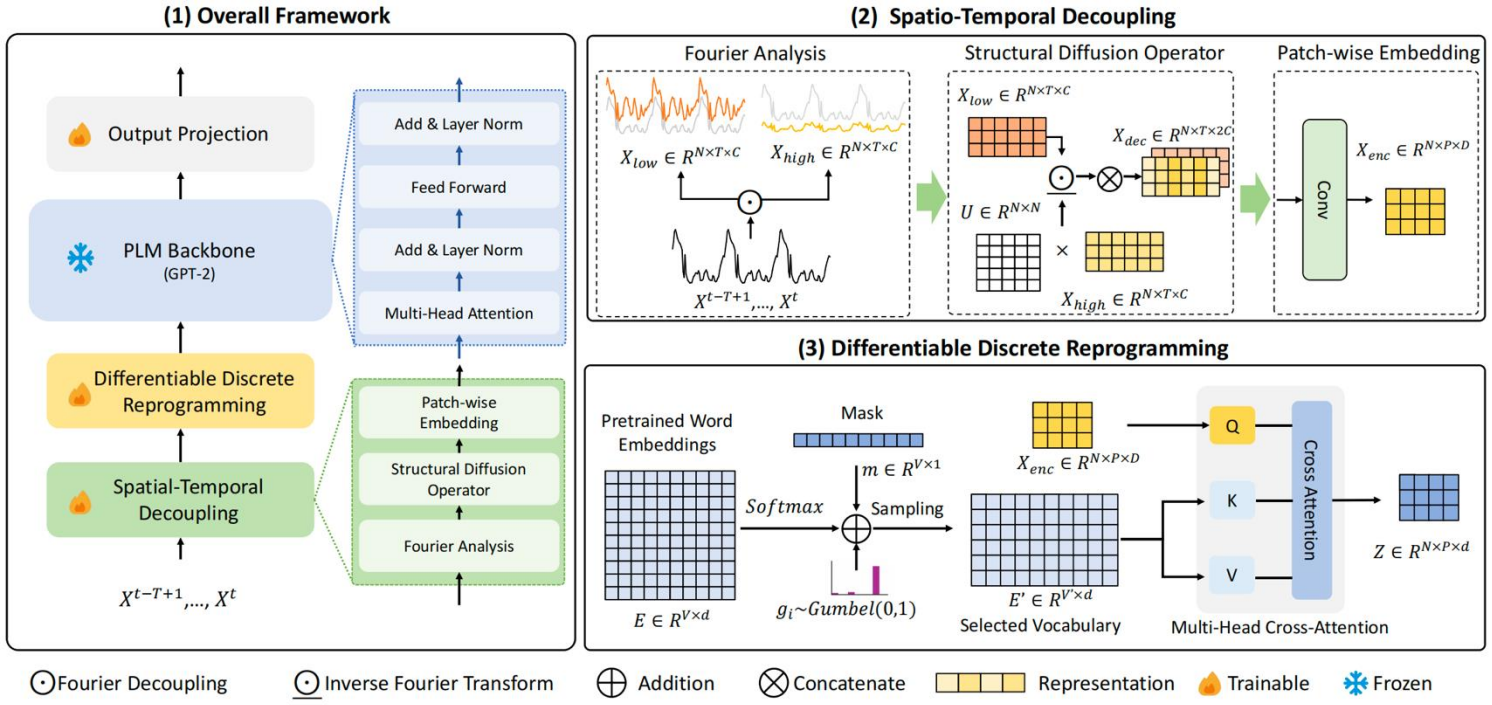
[BEGIN DATA]
***
[Domain]: We usually observe that electricity consumption peaks at noon, with a significant increase in transformer load
***
[Instruction]: Predict the next <H> steps given the previous <T> steps information attached
***
[Statistics]: The input has a minimum of <min_val>, a maximum of <max_val>, and a median of <median_val>. The overall trend is <upward or downward>. The top five lags are <lag_val>.
[END DATA]

Prompt example.

Jin, Ming, et al. Time-llm: Time series forecasting by reprogramming large language models. ICLR, 2024.

Model Reprogramming

- RePST: a reprogramming framework for spatio-temporal data.
  - Decouple the spatio-temporal data in the frequency space to capture intricate spatio-temporal dependencies.
  - Discretely sample word embeddings from the whole vocabulary to avoid semantic mixing and enhance the expressivity of text prototypes for spatio-temporal data.

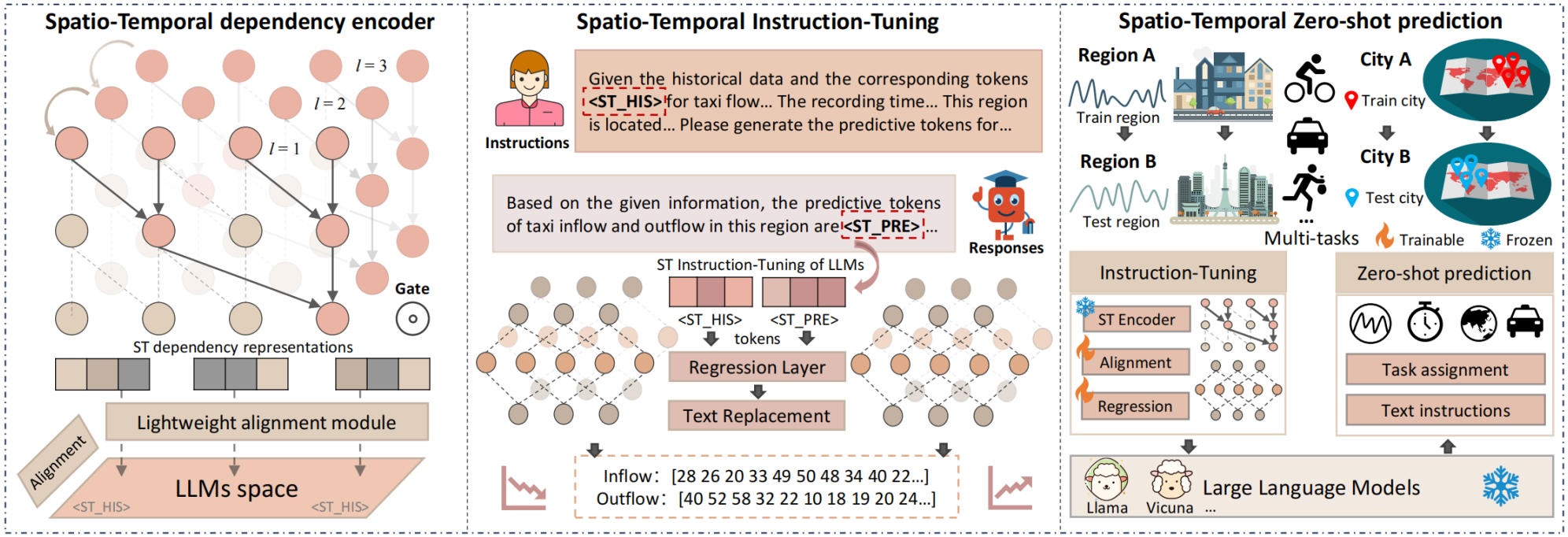


RePST framework.

Wang, Hao, et al. Empowering Pre-Trained Language Models for Spatio-Temporal Forecasting via Decoupling Enhanced Discrete Reprogramming. arxiv, 2024.

Model Reprogramming

- UrbanGPT: A spatio-temporal instruction tuning paradigm with the knowledge space of LLMs.
  - Spatio-temporal dependency encoding to capture spatio-temporal dynamics.
  - Spatio-temporal instruction tuning with input alignment and regression mapping.
  - Integrate multi-granularity time information and spatial details into prompt instructions.



The overall architecture of UrbanGPT.

Li, Zhonghang, et al. Urbangpt: Spatio-temporal large language models. KDD, 2024.

Summary & Future Directions

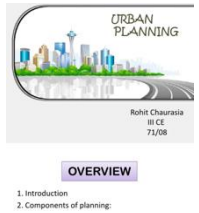
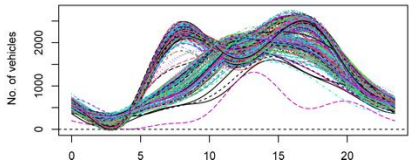
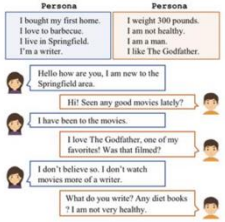
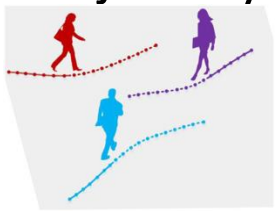
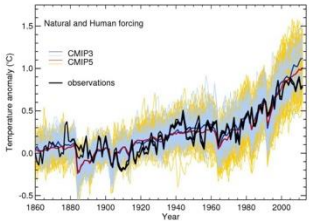
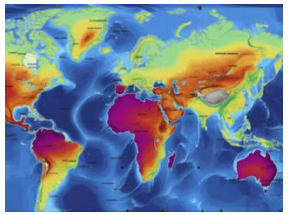
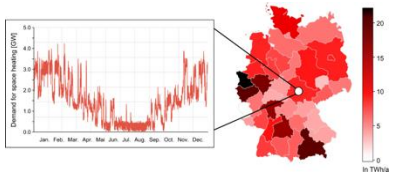
- Summary
  - Existing unimodal approaches have put some preliminary efforts into building a universal time series foundation model. However, the foundation model tailored for urban scenarios is situated in the nascent state.
  - Existing cross-modal approaches hinge on how to make LLMs comprehend time series data, hence unleashing the reasoning ability of LLMs. However, most of these works are limited to a very high-level understanding of LLM for time series.

- Future Directions
  - Build a universal urban time series foundation model that can accommodate diverse urban data formats, varied urban domain distributions, and distinct urban tasks.
  - Theoretically analyze how time series data influence LLM comprehension and prediction, and how LLMs make predictions when applied to urban time series data.
  - Integrate time series with other urban modalities to enrich the comprehension abilities of urban foundation models.

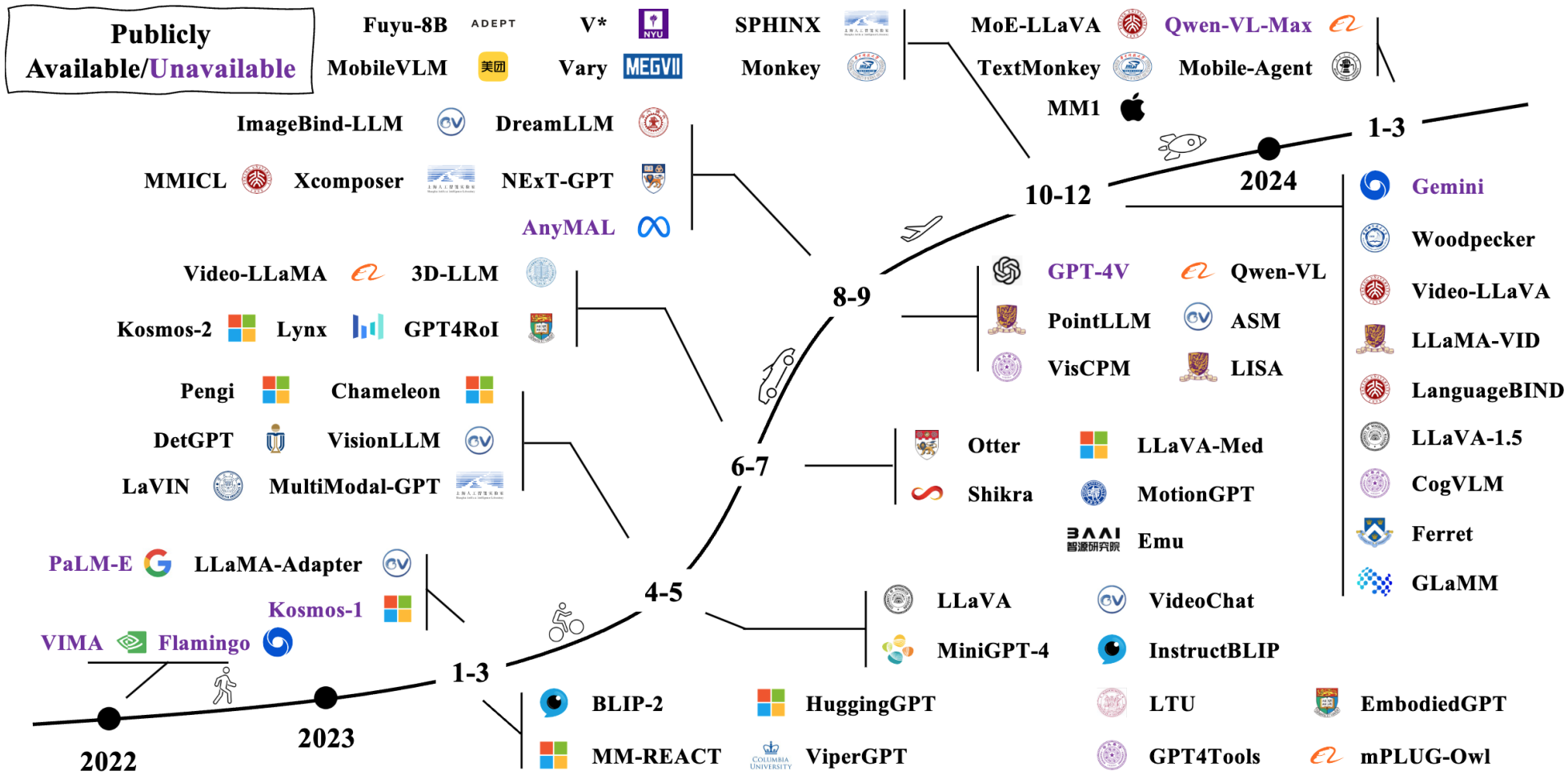
Multimodal UFMs

Multimodal Urban Data

- Multimodal understanding empowers UFMs with stronger versatility, generalization, and adaptation capabilities across broader urban tasks and domains.

Text	Vision	Trajectory	Time Series	Etc.
 Documents	 Street view	 Road network trajectory	 Traffic flow	 Road networks
 Dialogs	 Remote	 Pedestrian trajectory	 Weather observation	 POIs
 Geo-text	 Meteorological raster data	 Check in	 Energy demand	 Regions

Multimodal Foundation Models



- Lack of sufficient urban domain knowledge and spatio-temporal reasoning ability.
- Uncompetitive to process broader data modalities in urban domains.

Yin, Shukang, et al. A survey on multimodal large language models. arXiv, 2023.

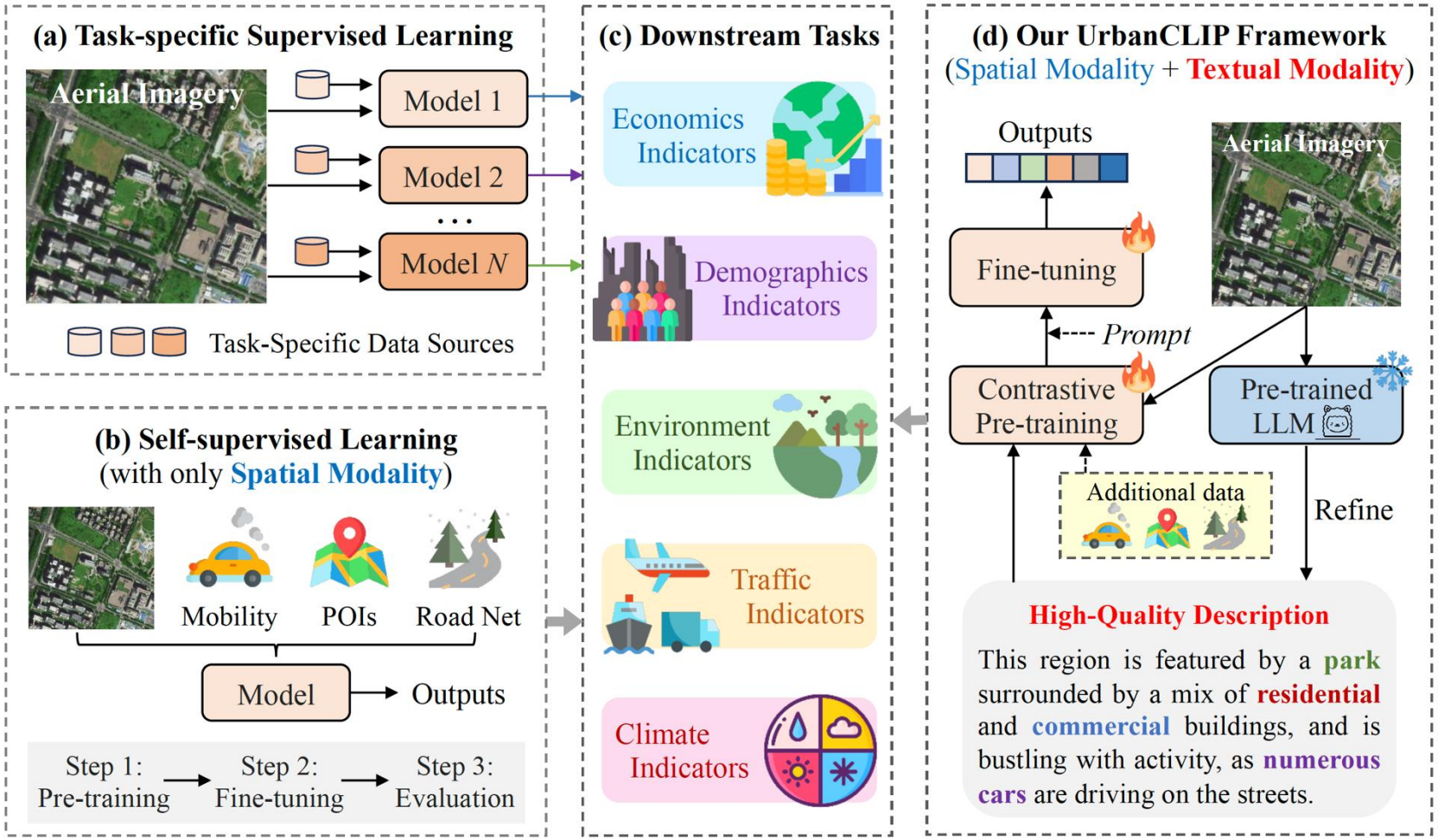
Multimodal UFMs

- *Multimodal Pre-training*
  - *Multimodal urban data*

- **Multimodal Adaptation**
  - Prompt engineering
  - Model fine-tuning

UrbanCLIP: Multimodal Urban Region Profiling

- UrbanCLIP introduces text description to enhance satellite vison for Urban Region Profiling.

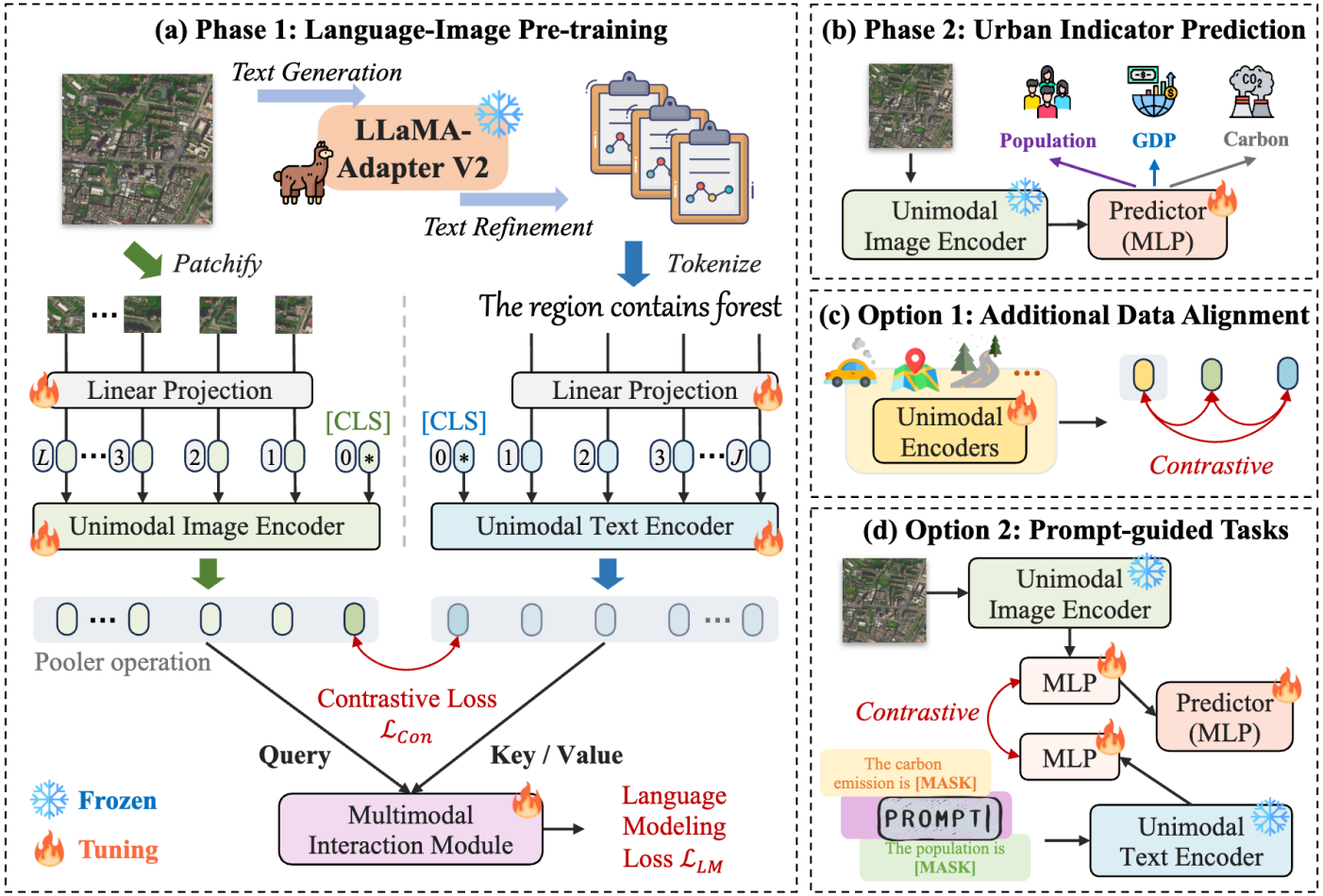


UrbanCLIP is a framework applies LLMs to enhance multimodal urban region profiling.

Yan, Yibo, et al. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. WWW, 2024.

UrbanCLIP: Multimodal Urban Region Profiling

- UrbanCLIP leverages LLMs to generate textual descriptions for satellite images and employs contrastive learning as supervision for urban visual representation learning.



Framework of UrbanCLIP.

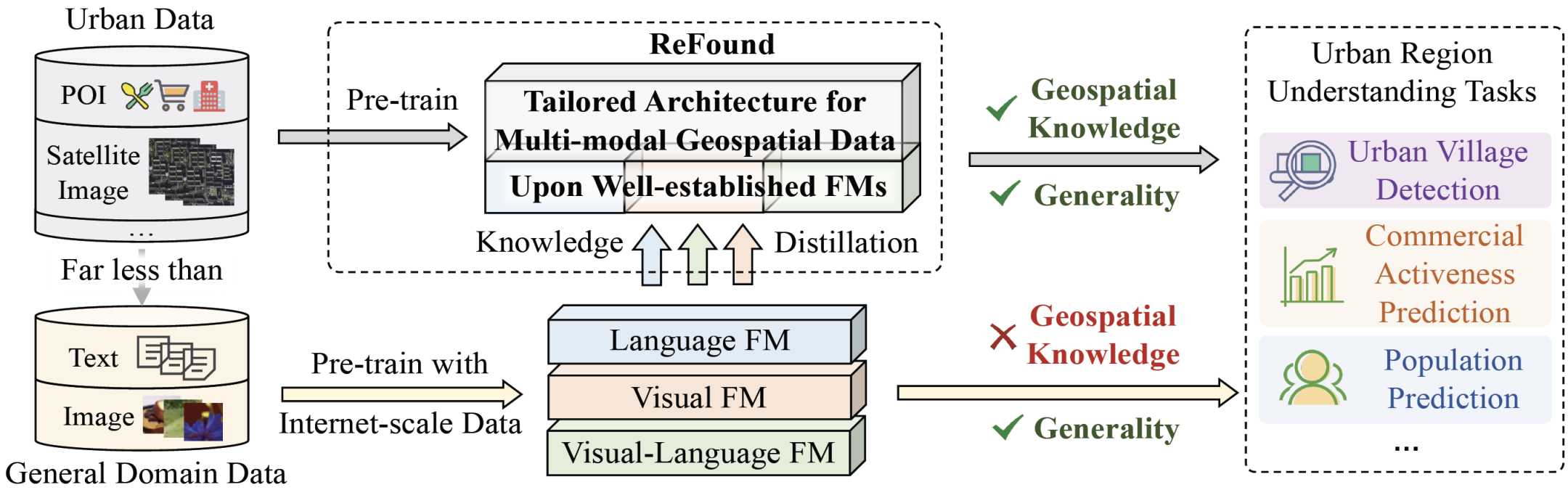
- **Phase 1**:
  - LLM to generate textual descriptions from satellite imagery.
  - Process image-text pairs with unimodal encoders.
  - Align the representations by contrastive learning.

- **Phase 2**:
  - Fine-tune a MLP Predictor for urban indicator prediction based on the frozen image encoder.

Yan, Yibo, et al. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. WWW, 2024.

ReFound: Multimodal Urban Region Understanding

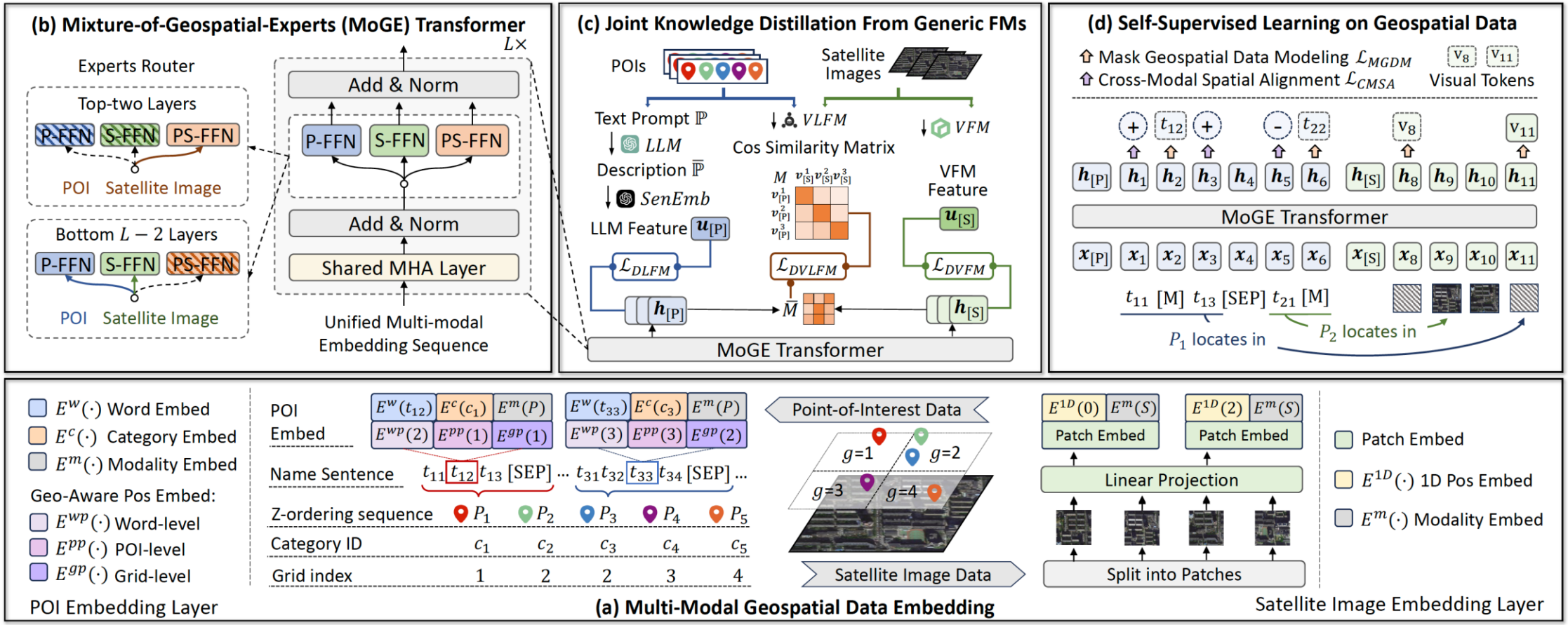
- ReFound learns in-domain knowledge from multi-modal geospatial data while distills general knowledge from established FMs, empowering model with both general and domain knowledge.



ReFound is a special foundation model with multi-modal urban data for urban region understanding, involving satellite image, text, and POIs data.

Xiao, Congxi, et al. ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations. SIGKDD, 2024.

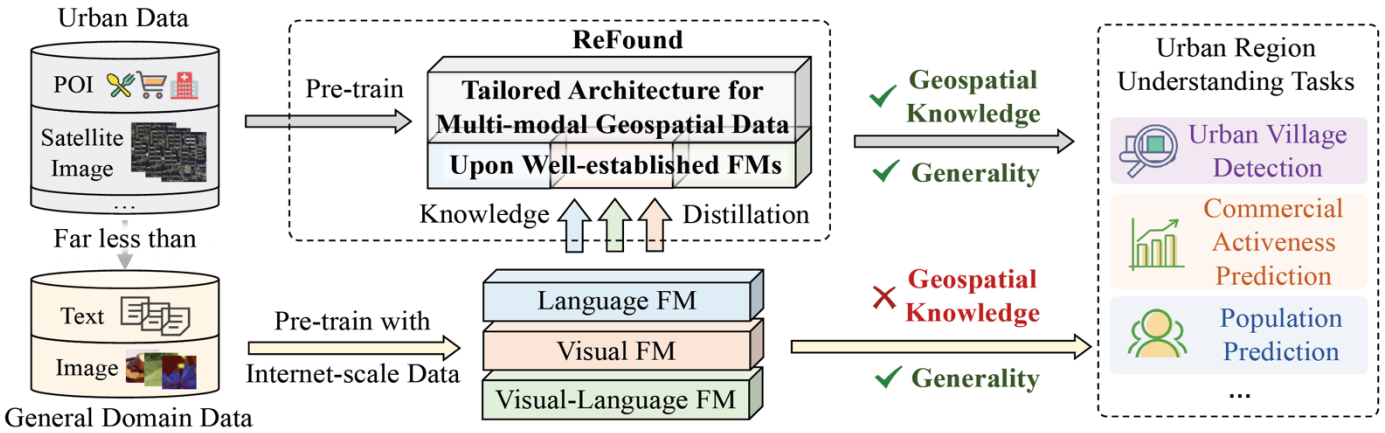
ReFound: Multimodal Urban Region Understanding



The architecture and pre-training framework of ReFound.

Xiao, Congxi, et al. ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations. SIGKDD, 2024.

ReFound: Multimodal Urban Region Understanding



Advantages of ReFound compared to general foundation models.

Advantages over general FMs:

- Less training data needed.
- Has both general and geospatial knowledge.

		Urban Village Detection		Commercial Activeness Prediction			Population Prediction		
Usage	Methods	AUC ↑	F1-score ↑	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑
Fine-tuning	BERT	0.73 ± 0.01	0.40 ± 0.09	17.31 ± 0.34	8.64 ± 0.24	0.44 ± 0.02	361.60 ± 2.11	266.99 ± 2.92	0.60 ± 0.00
	ViT	0.71 ± 0.01	0.39 ± 0.01	21.77 ± 0.19	10.95 ± 0.39	0.12 ± 0.02	338.23 ± 2.94	246.92 ± 2.90	0.65 ± 0.01
	CN-CLIP	0.74 ± 0.01	0.41 ± 0.03	18.39 ± 0.30	8.70 ± 0.11	0.37 ± 0.02	303.61 ± 5.35	220.79 ± 4.87	0.72 ± 0.01
	CN-CLIP-I	0.73 ± 0.02	0.38 ± 0.03	22.22 ± 0.19	11.63 ± 0.53	0.08 ± 0.02	337.68 ± 12.01	244.92 ± 8.20	0.65 ± 0.03
	SpaBERT	0.65 ± 0.02	0.31 ± 0.02	19.45 ± 0.35	10.26 ± 0.32	0.30 ± 0.03	389.93 ± 4.24	296.28 ± 1.45	0.53 ± 0.01
	GFM	0.76 ± 0.01	0.44 ± 0.03	21.43 ± 0.31	11.38 ± 0.45	0.15 ± 0.02	325.36 ± 4.81	237.47 ± 4.66	0.67 ± 0.01
	ReFound	**0.82 ± 0.02**	**0.44 ± 0.03**	**14.85 ± 0.16**	**7.57 ± 0.15**	**0.59 ± 0.01**	**286.10 ± 4.37**	**203.42 ± 3.39**	**0.75 ± 0.01**
Feature-based Prediction	HGI	0.57 ± 0.00	0.28 ± 0.01	20.18 ± 0.01	11.52 ± 0.03	0.24 ± 0.00	347.47 ± 2.09	263.69 ± 1.88	0.63 ± 0.00
	MMGR	0.70 ± 0.00	0.37 ± 0.02	21.86 ± 0.06	12.22 ± 0.22	0.11 ± 0.00	370.79 ± 0.38	279.34 ± 0.92	0.58 ± 0.00
	PG-SimCLR	0.68 ± 0.01	0.35 ± 0.03	21.70 ± 0.07	11.61 ± 0.21	0.13 ± 0.01	403.02 ± 0.99	303.82 ± 0.90	0.50 ± 0.00
	ReFound	**0.77 ± 0.00**	**0.44 ± 0.01**	**17.28 ± 0.20**	**9.96 ± 0.23**	**0.45 ± 0.01**	**308.45 ± 1.21**	**224.97 ± 0.87**	**0.71 ± 0.00**

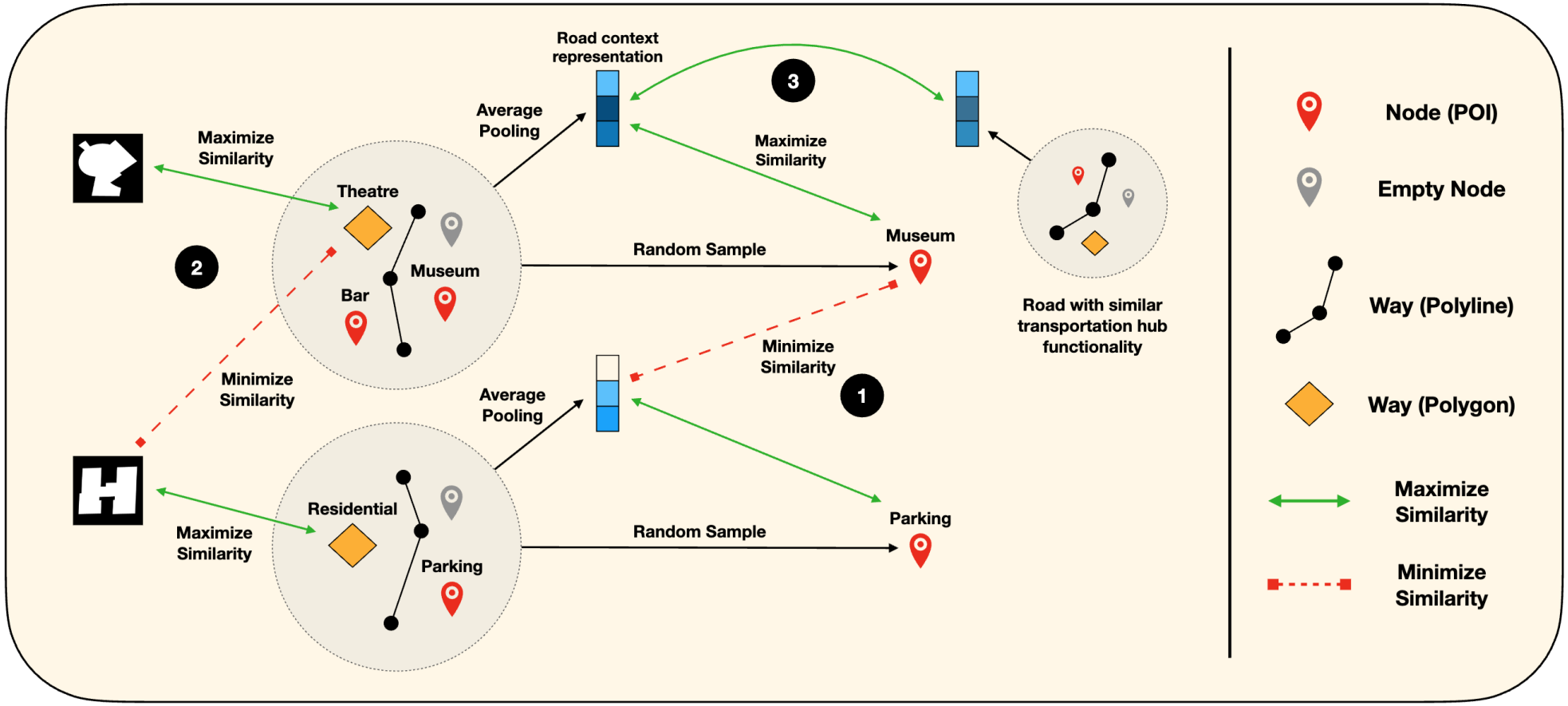
Performance comparison in three downstream tasks.

- Adapted to various downstream urban tasks.
- Achieve around 16% improvement compared to general FM.
- ReFound performs well without tuning the pre-trained parameters.

Xiao, Congxi, et al. ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations. SIGKDD, 2024.

CityFM: Learning Entity Representations from OSM

- CityFM designs three contrastive objectives using nodes, ways, and relations from OpenStreetMap to learn multimodal representations of geospatial entities.

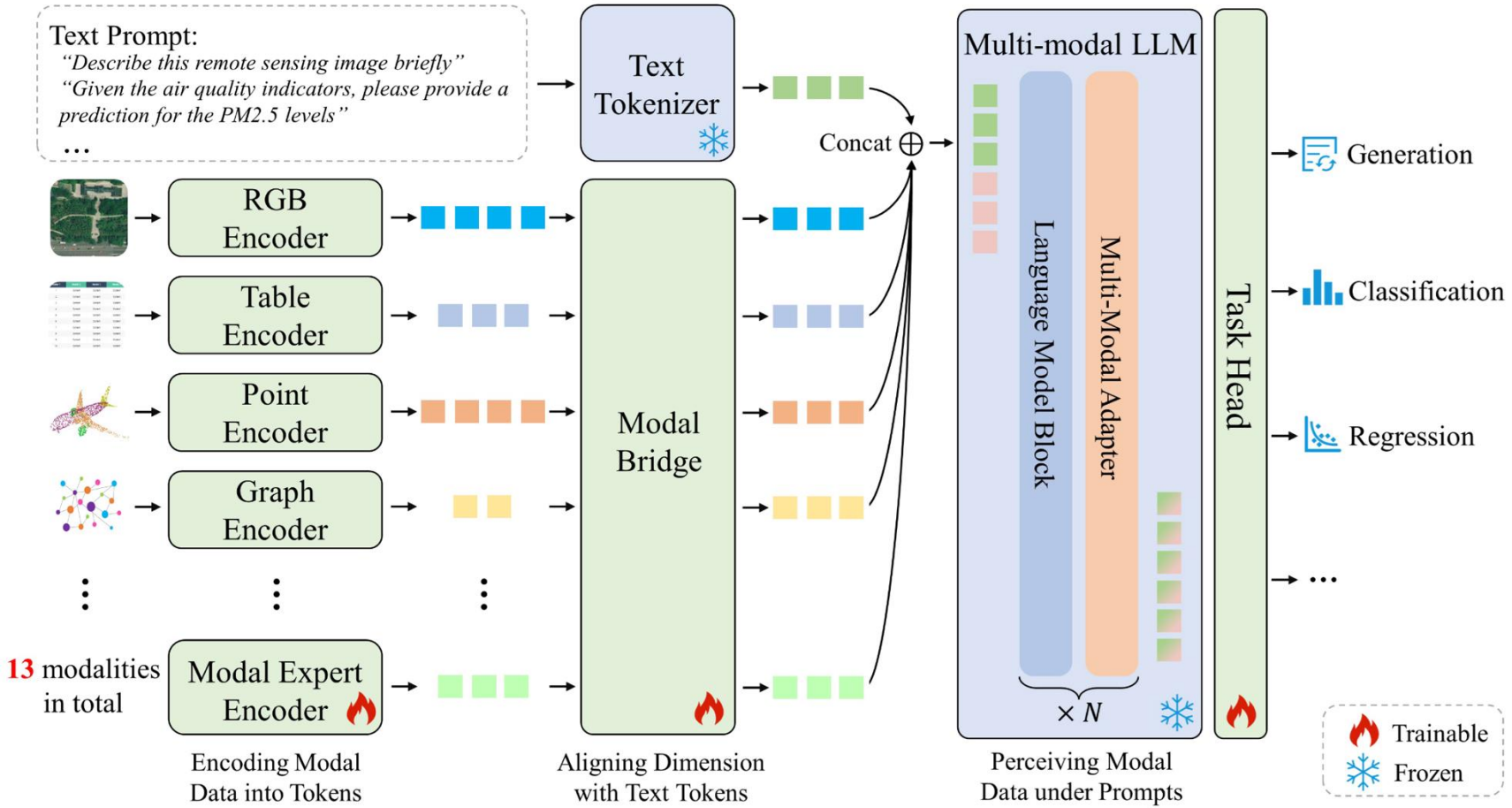


CityFM pre-trains foundation models by using spatial, visual, and textual data from OpenStreetMap in a self-supervised manner, learning multimodal representations of geospatial entities, such as node, polyline, polygon.

Balsebre, Pasquale, et al. City Foundation Models for Learning General Purpose Representations from OpenStreetMap. arXiv, 2023.

AllSpark: Multimodal Spatio-Temporal Model

- Align all data modalities to the language modality, then process them using LLMs.



AllSpark Architecture. (1) Extract multimodal data into token sequences using specific modal encoders; (2) Align dimensions of modal tokens with text prompt tokens via a modal bridge; (3) Interpret the aligned tokens with a multimodal LLM; (4) Apply to downstream tasks with task-specific heads.

Shao, Run, et al. AllSpark: A Multimodal Spatio-Temporal General Intelligence Model with Thirteen Modalities. arXiv, 2024.

Multimodal UFMs

Multimodal Pre-training

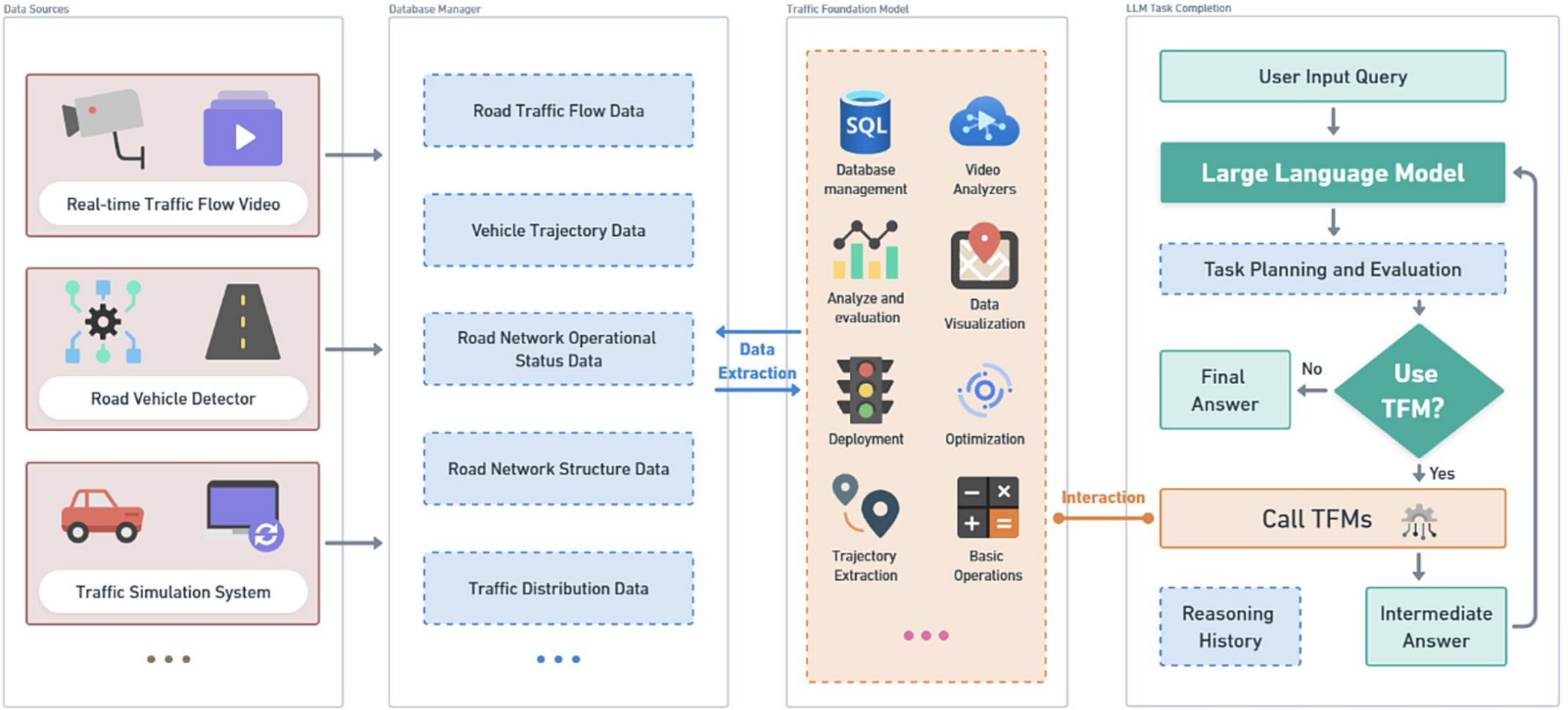
- Multimodal urban data

Multimodal Adaptation

- *Prompt engineering*
- *Model fine-tuning*

TrafficGPT: Empower LLMs with Traffic Analytical Tools

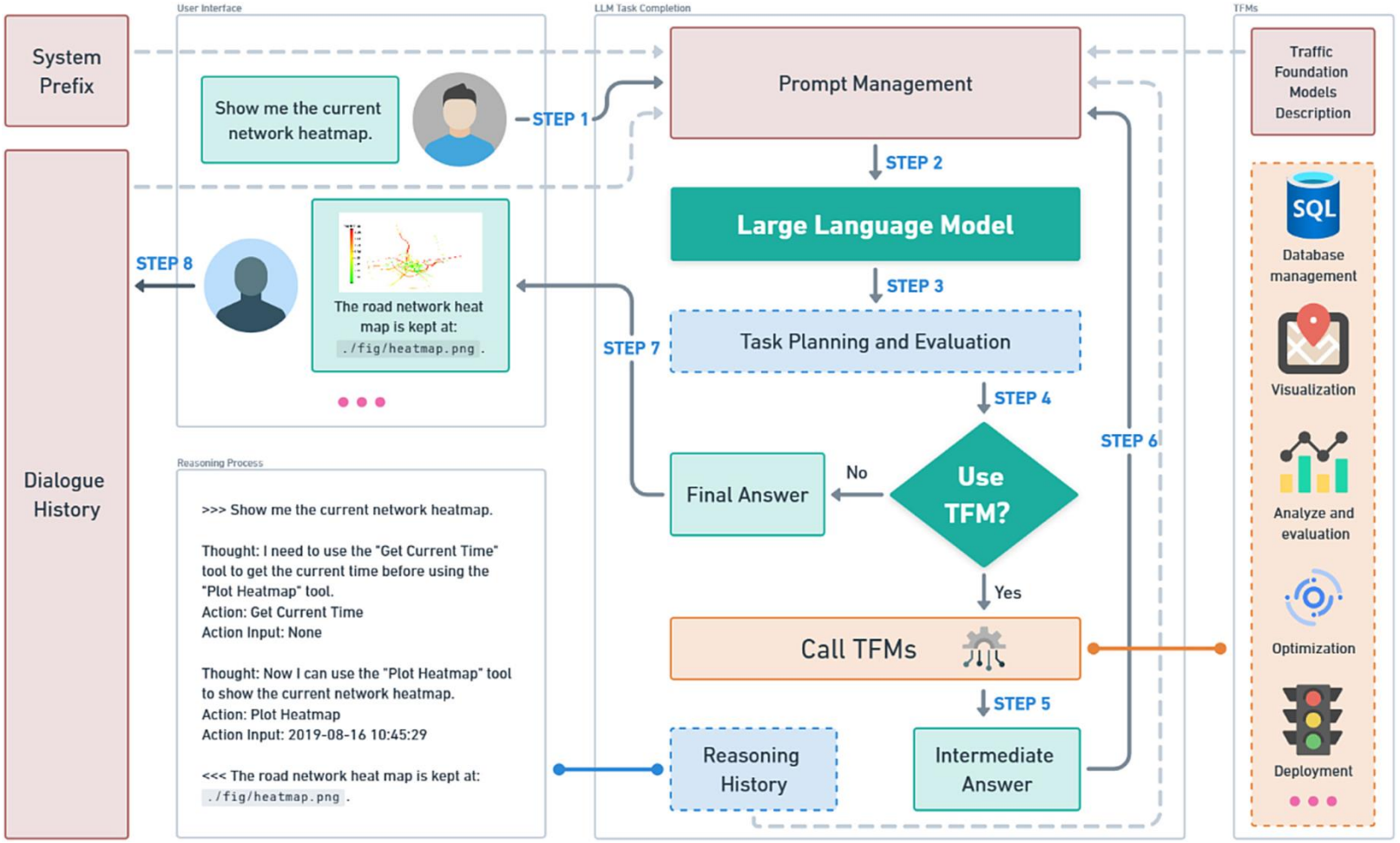
- TrafficGPT empowers urban traffic management by integrating LLMs and external tools.



TrafficGPT extracts and processes multi-modal traffic data via traffic analytical tools. A LLM is used to identifies user needs and orchestrates task execution by interacting with the traffic analytical tools.

Zhang, Siyao, et al. Trafficgpt: Viewing, processing and interacting with traffic foundation models. Transport Policy, 2024.

TrafficGPT: Empower LLMs with Traffic Analytical Tools

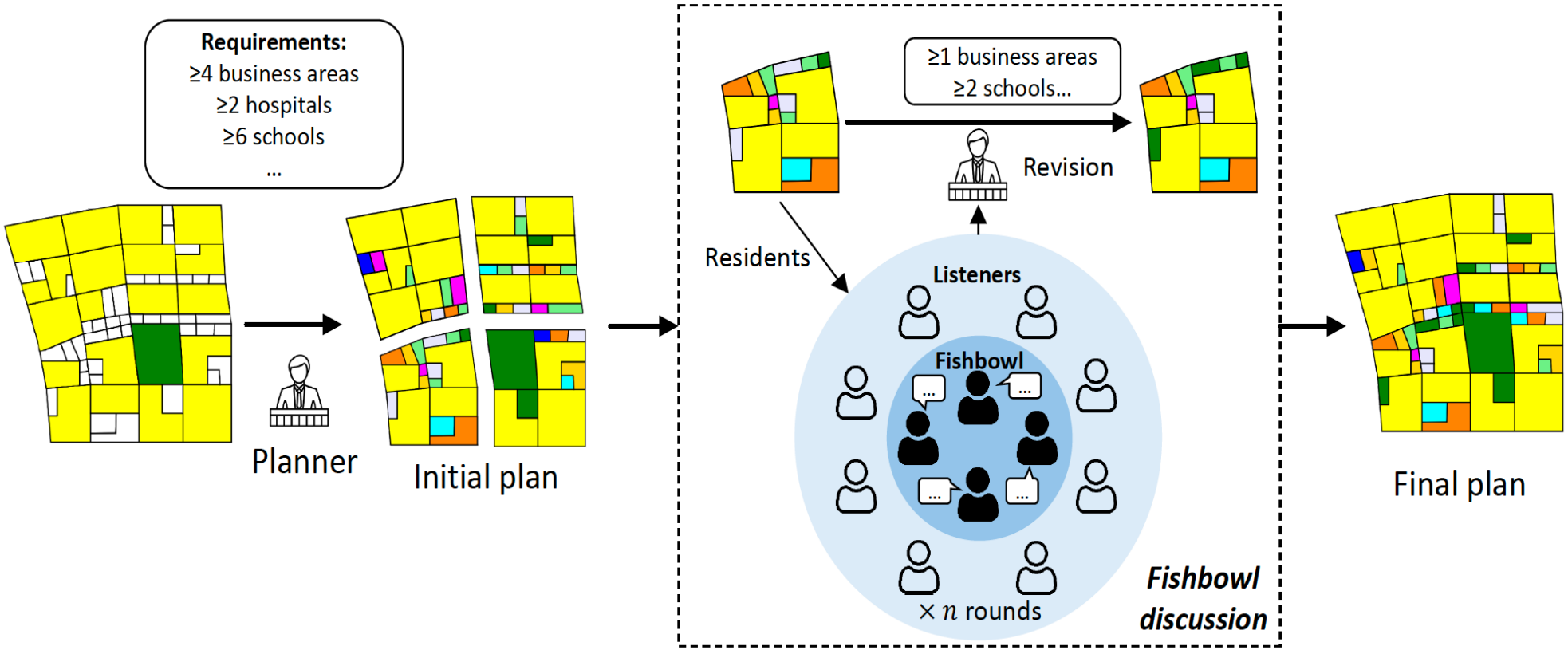


Overview of TrafficGPT. (1) User Interface section shows how users interact with TrafficGPT; (2) Reasoning Process section shows how TrafficGPT deconstructs the task and completes the sub-tasks step by step; (3) LLM Task Completion section shows how TrafficGPT form prompts and iteratively invokes tools to provide answers.

Zhang, Siyao, et al. Trafficgpt: Viewing, processing and interacting with traffic foundation models. Transport Policy, 2024.

LLMs for Participatory Urban Planning

- Leverage LLMs to simulate planners and residents agents, achieving participatory urban planning through multi-agent collaboration.

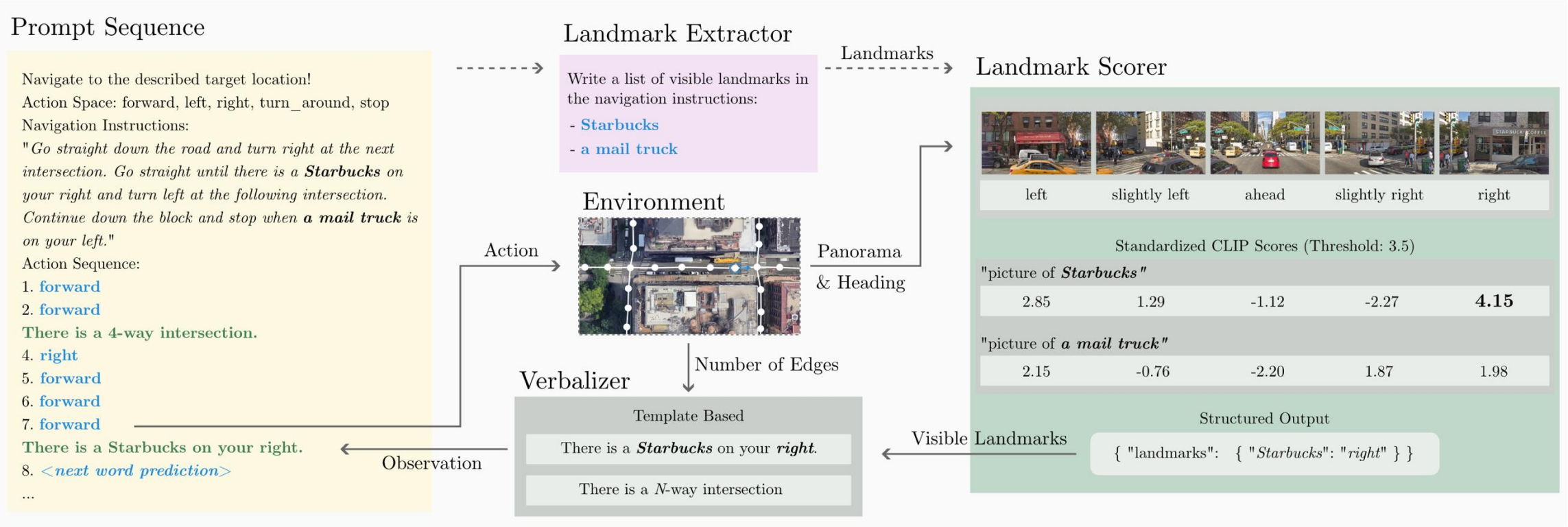


Framework overview: LLM agents are assigned roles as a planner and residents using carefully designed prompts. (1) Planner generates an initial land-use plan based on expert knowledge and quantity requirements; (2) Resident agents provide feedback through fishbowl discussions; (3) Planner revises the plan based on this feedback to balance the diverse needs of the community.

Zhou Z, Lin Y, Jin D, et al. Large Language Model for Participatory Urban Planning. arXiv, 2024.

VELMA: Vision-Language Navigation in Street View

- VELMA verbalizes the visual observations from the environment and then feed them into a LLM to predict navigation actions.



VELMA is an embodied LLM-based agent for vision and language navigation in the street view. (1) Extract key landmarks from textual instructions using LLM; (2) Employ CLIP to calculate similarity between the landmark's textual description and the visual content of panoramic image; (4) Transform the agent's observations into natural language text; (4) LLM predicts the appropriate action.

Schumann, Raphael, et al. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. AAAI, 2024.

Takeaways

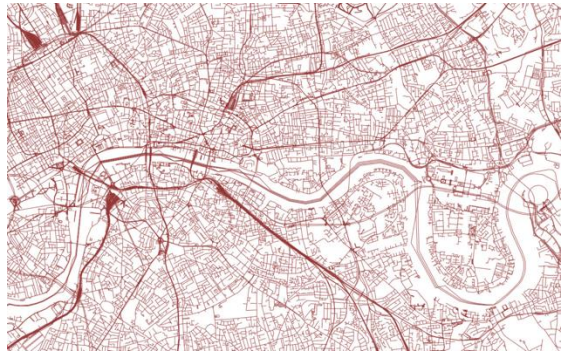
- Summary on existing studies
  - Mainly focus on language and vision modalities.
  - Leverage established FMs to enhance in-domain FMs.
  - Prompting established FMs (LLMs) with tool invocation or agent simulation to address urban tasks.

- Opportunities for future works
  - Integrating a broader range of urban data types and building more versatile UFMs applying to broader urban domains.
  - Fully harnessing the power of established FMs.
  - Effective coordination for various UFMs and urban analytical tools.
  - Spatio-temporal reasoning of UFMs.
  - Privacy and security issues in UFMs.

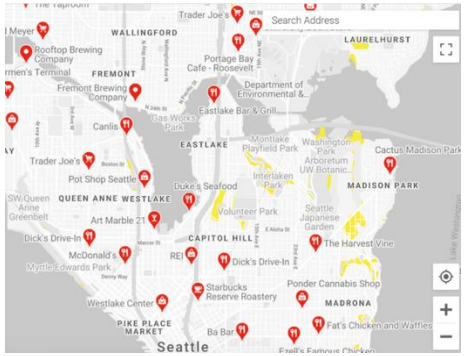
Other UFMs

Other Urban Data Types

Geo-vector data



Road networks



Point-of-Interests

Tabular data

Demographic Categories	Frequency	Valid Percentage	U.S. National Census Data (2012), %
Gender			
Female	150	49.5	51.1
Male	153	50.5	48.9
Age			
18–24	24	9.5	11.2
25–34	53	20.9	13.4
35–44	36	14.2	12.9
45–54	67	26.5	14.2
55–64	61	24.1	12.3
65+	12	4.7	13.5
Not specified	50	—	—
Ethnicity			
Euro-American/Caucasian	241	79.5	63.0
African American	26	8.6	13.1
Hispanic/Latino(a)	17	5.6	16.9
Asian American	11	3.6	5.1
Other	8	2.6	1.9
Marital status			
Married	168	55.4	56.4
Never married	70	23.1	26.9
Divorced/separated/widowed	68	21.5	16.7

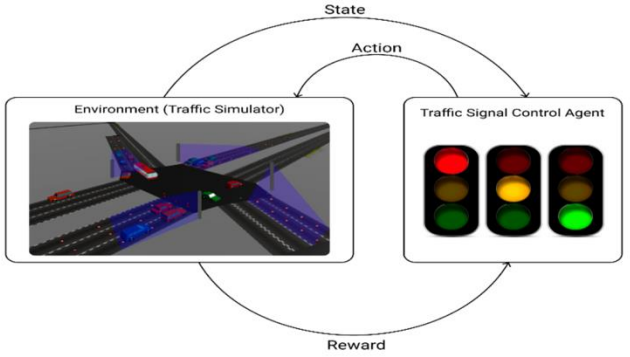
Demographic data

Country	Base Rate	GDP	GDP	CPI	CPI	PPI	PPI	Jobless	Jobless
	Current	YoY change	Period	YoY change	Period	YoY change	Period	Rate	Period
Eurozone	0.25%	-0.30%	Q3	0.80%	Dec			12.10%	Nov
France	0.25%	0.20%	Q3	0.80%	Dec			10.90%	Q3
Germany	0.25%	1.10%	Q3	1.40%	Dec	-0.50%	Dec	6.90%	Dec
Ireland	0.25%	1.70%	Q3	0.20%	Dec			12.40%	Dec
Italy	0.25%	-1.80%	Q3	0.70%	Dec	-1.80%	Nov	12.70%	Nov
Switzerland	0.25%	1.90%	Q3	0.10%	Dec	-0.40%	Dec	3.50%	Dec
UK	0.50%	1.50%	Q3	2.00%	Dec	1.00%	Dec	7.10%	Nov
Canada	1.00%	2.70%	Q3	0.90%	Nov			7.20%	Dec
US	0.25%	4.10%	Q3	1.50%	Dec	1.20%	Dec	6.70%	Dec
China	6.00%	7.70%	Q4	2.50%	Dec	-1.40%	Dec	4.00%	Q4
Japan	0.10%	1.10%	Q3	1.50%	Nov			4.00%	Nov

Economic data

Polygonal Region

Decision sequence



• • •

Traffic control trajectory

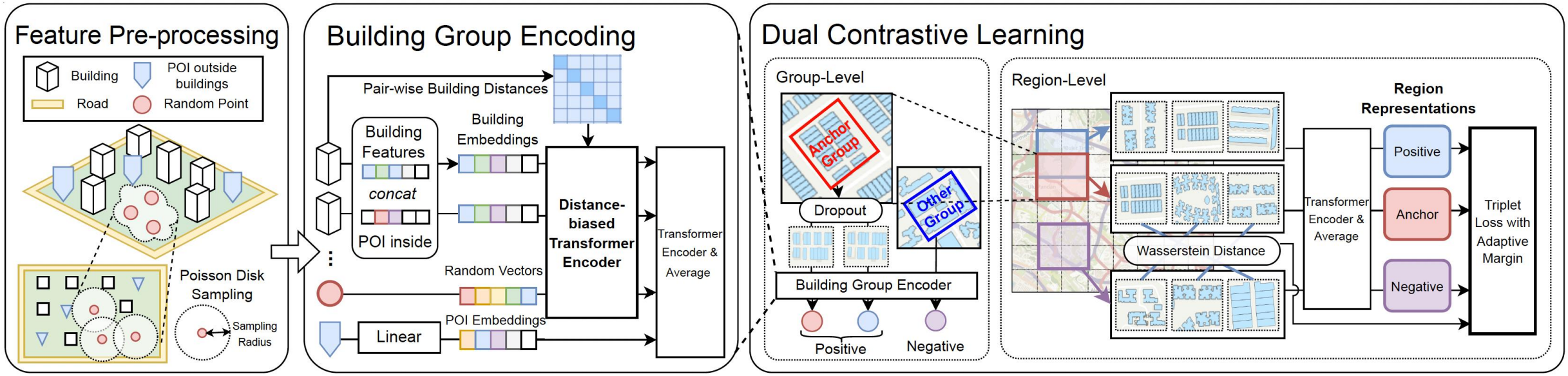
Other UFMs

- *Pre-Training*

- Adaptation

RegionDCL: Learning Region Embeddings from OSM

- RegionDCL Learns region embeddings from buildings and POIs via group-level and region-level contrastive learning with a distance loss between building groups.

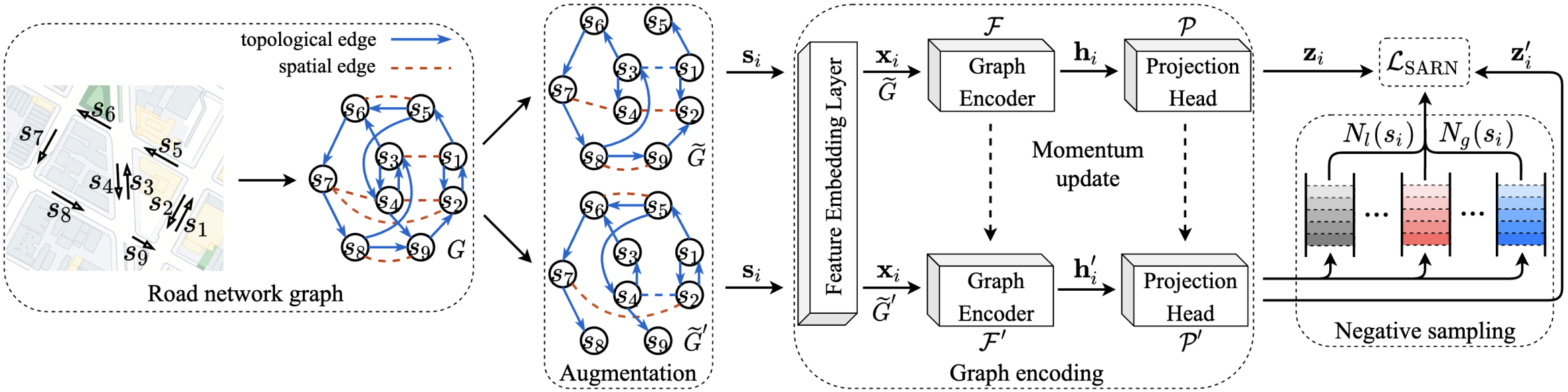


An overview of RegionDCL. (1) Feature Pre-Processing: process individual buildings and POIs into compact embeddings; (2) Building Group Encoding: Encod building in groups via distance-biased transformer encoder to consider their spatial distributions and intercorrelations; (3) Dual Contrastive Learning: Learn region embeddings via dual contrastive learning with a distance loss at both building group.

Li, Yi, et al. Urban Region Representation Learning with OpenStreetMap Building Footprints. KDD 2023.

SARN: Structure-Aware Road Network Embedding

- SARN learns generic embeddings for road networks applied across various urban tasks.
- Transform road network embedding to a graph node embedding learning problem, employing graph contrastive learning to learn road segment embeddings.



The architecture of SARN. The road network graph augmented into two views by selectively masking the edges; Each view is fed into separate graph encoders to obtain the embeddings; SARN constructs local and global contrastive objectives by obtaining negative samples from road segments within the same local area and across different areas.

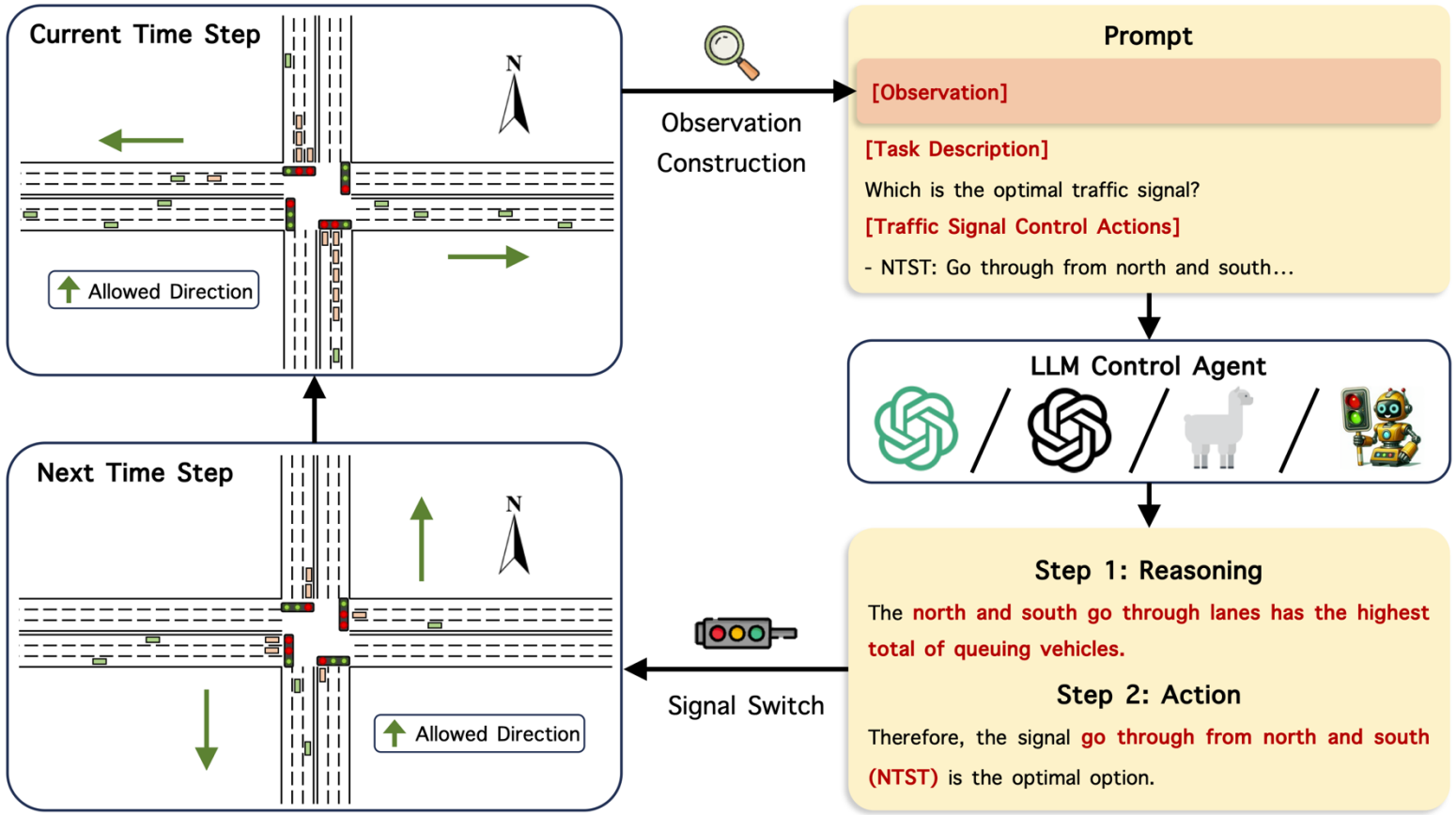
Chang, Yanchuan, et al. Spatial Structure-Aware Road Network Embedding via Graph Contrastive Learning. EDBT 2023.

Others UFMs

- Pre-Training
- *Adaptation*

LLMLight: LLM-based Traffic Control Agent

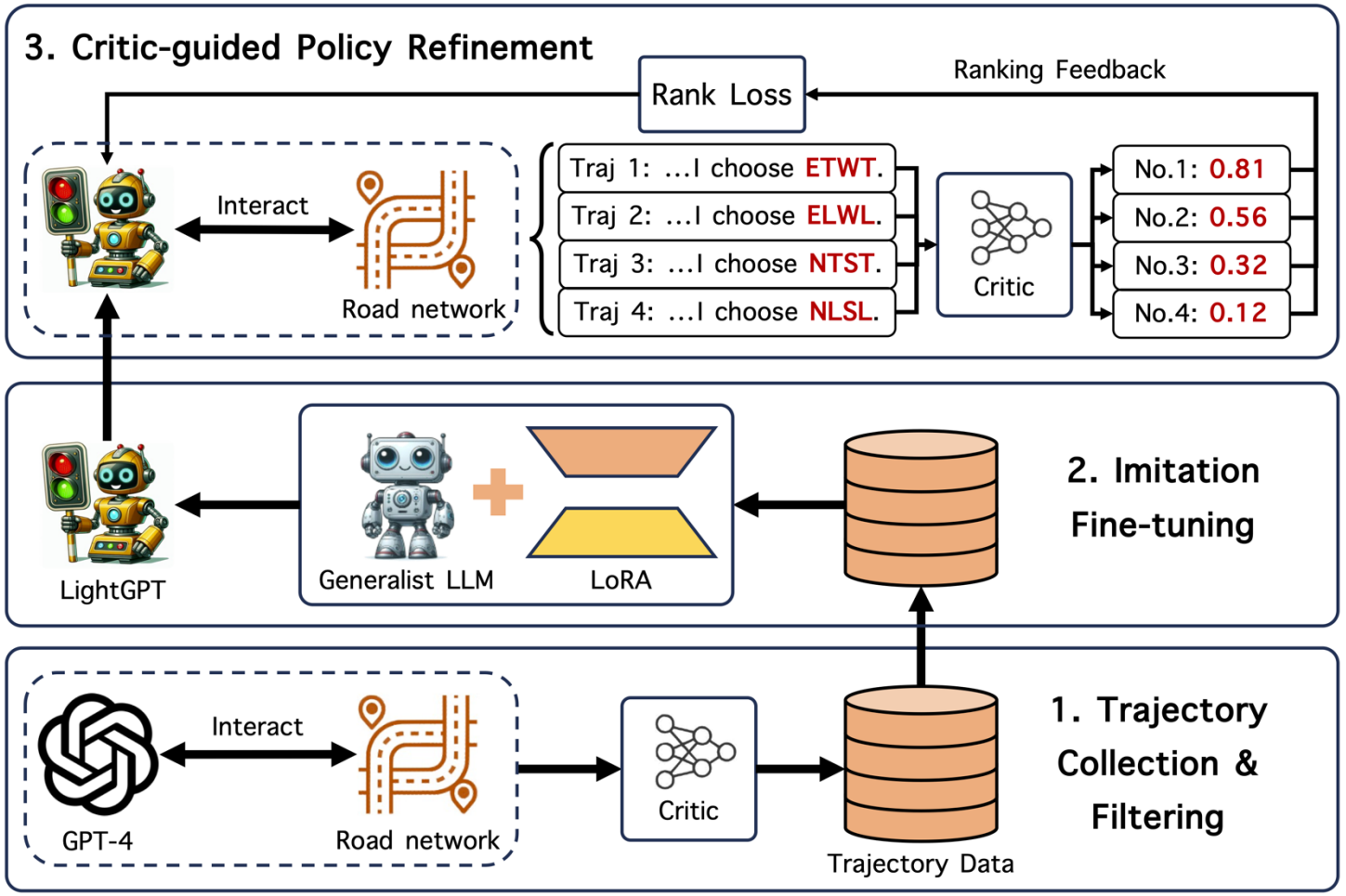
- Transform the real-time traffic observations of intersection into human-readable text.
- Leverage LLM's Chain-of-Thought (CoT) reasoning to determine the optimal traffic signal control strategy.



The workflow of LLMLight: a traffic signal control agent framework based on LLMs.

Lai, Siqi, et al. LLMLight: Large Language Models as Traffic Signal Control Agents. arXiv, 2023.

LLMLight: LLM-based Traffic Control Agent



The training procedure of LLMLight.

- **Trajectory collection and filtering**: Collect CoT reasoning trajectories from GPT-4 and select those aligning with long-term goals for quality data.
- **Imitation fine-tuning**: Train the backbone LLM using collected reasoning trajectories, learning underlying rationales from GPT-4.
- **Critic-guided policy refinement**: Enhance LLM decision-making by fine-tuning with feedback from a well-trained critic model.

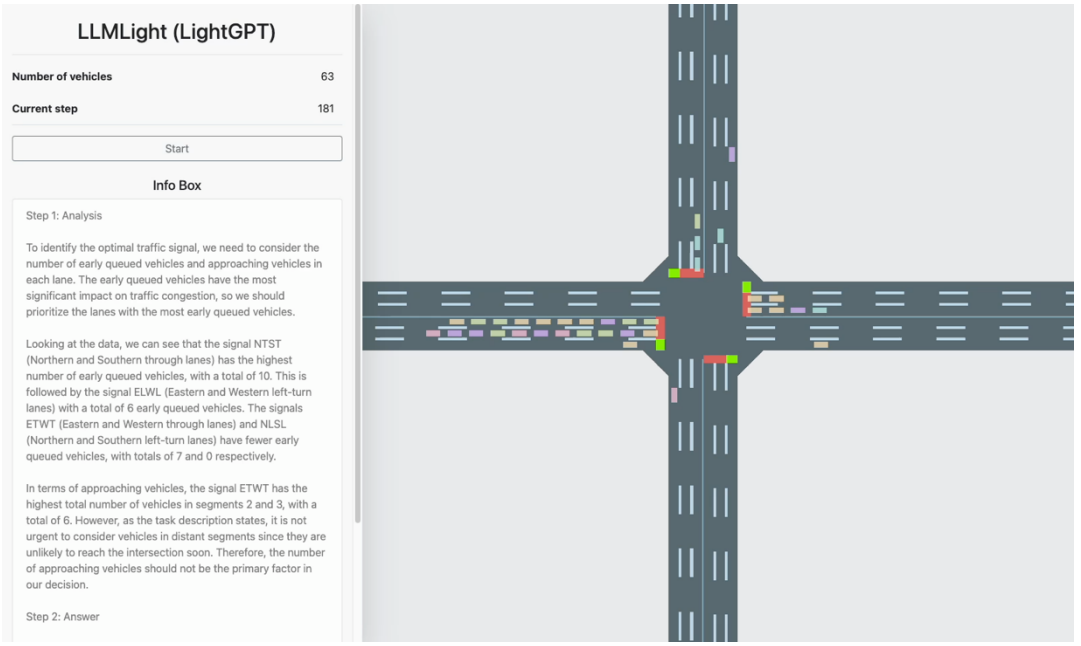
Lai, Siqi, et al. LLMLight: Large Language Models as Traffic Signal Control Agents. arXiv, 2023.

LLMLight: LLM-based Traffic Control Agent

- LLMLight outperforms existing SOTA algorithms on most datasets.
- Small road networks: travel time reduced by 3.5%, wait time by 16%.
- Large road networks/extreme congestion: travel time reduced by 18.5%, wait time by 58%.
- Advantages: more effective, generalizable, and interpretable.

Models	Jinan									Hangzhou					
	1			2			3			1			2		
	ATT	AQL	AWT	ATT	AQL	AWT	ATT	AQL	AWT	ATT	AQL	AWT	ATT	AQL	AWT
Transportation Methods															
Random	597.62	687.35	99.46	555.23	428.38	100.40	552.74	529.63	99.33	621.14	295.81	96.06	504.28	432.92	92.61
FixedTime	481.79	491.03	70.99	441.19	294.14	66.72	450.11	394.34	69.19	616.02	301.33	73.99	486.72	425.15	72.80
Maxpressure	281.58	170.71	44.53	273.20	106.58	**38.25**	265.75	133.90	40.20	325.33	68.99	49.60	347.74	215.53	70.58
RL Methods															
MPLight	307.82	215.93	97.88	304.51	142.25	90.91	291.79	171.70	89.93	345.60	84.70	81.97	358.56	237.17	100.16
AttendLight	291.29	186.25	61.73	280.94	115.52	52.46	273.02	144.05	55.93	322.94	66.96	55.19	358.81	239.05	72.88
PressLight	291.57	185.46	50.53	281.46	115.99	47.27	275.85	148.18	54.81	364.13	98.67	90.33	417.01	349.25	150.46
CoLight	279.60	168.53	58.87	274.77	108.28	54.14	266.39	135.08	53.33	322.85	66.94	61.82	342.90	212.09	99.74
Efficient-CoLight	277.11	163.60	43.41	269.24	102.98	39.74	262.25	129.72	**39.99**	311.96	58.20	**36.83**	333.27	189.65	**61.70**
Advanced-CoLight	274.67	160.85	49.30	268.25	102.12	41.11	260.66	127.83	43.54	**304.47**	**52.94**	41.75	**329.16**	**186.34**	76.59
LLMLight (with Generalist LLMs)															
Qwen2-72B	291.95	185.20	58.65	277.78	112.13	53.12	274.33	144.95	54.94	321.91	65.75	66.52	342.37	206.07	100.55
Llama2-70B	353.03	286.52	85.96	324.52	162.34	99.87	320.41	210.13	100.90	357.95	93.21	106.63	361.53	250.69	121.58
Llama3-70B	290.19	186.43	61.09	277.49	112.86	51.76	271.60	142.95	54.55	325.85	69.42	71.51	339.23	198.82	78.81
ChatGPT-3.5	536.79	952.93	155.62	524.81	393.21	179.34	501.36	479.50	154.19	463.04	181.95	191.87	418.75	336.47	130.56
GPT-4	275.26	160.93	46.61	271.34	105.22	47.55	264.70	132.53	46.16	318.71	62.84	58.09	335.81	193.32	66.02
LLMLight (with LightGPTs)															
LightGPT (Qwen2-0.5B)	296.71	193.93	46.80	292.99	129.93	80.63	290.08	169.56	91.20	328.11	70.90	84.49	351.21	233.16	102.39
LightGPT (Qwen2-7B)	275.92	163.34	48.56	270.41	104.40	44.93	263.10	130.94	45.75	313.37	59.03	50.65	335.30	198.47	72.93
LightGPT (Llama2-7B)	275.11	161.35	46.38	269.01	102.92	43.06	260.53	127.75	41.84	314.24	59.59	39.66	333.94	191.63	65.49
LightGPT (Llama3-8B)	275.10	161.92	48.25	268.81	102.54	42.74	262.29	130.32	44.96	311.72	58.40	47.60	333.85	192.06	69.75
LightGPT (Llama2-13B)	**274.03**	**159.39**	**43.24**	**266.94**	**100.46**	40.34	**260.17**	**127.08**	41.00	310.78	56.93	38.64	330.71	189.09	64.16

Overall performance comparison.



Demo of traffic control by LLMLight.

Lai, Siqi, et al. LLMLight: Large Language Models as Traffic Signal Control Agents. arXiv, 2023.