EMORY UNIVERSITY

Beyond Pixels: Semantic Understanding of Remote Sensing Imagery with AI Captioning and Foundation Models

Xiao Huang

Department of Environmental Sciences

Emory University

xiao.huang2@emory.edu

ISDE Seminar Series

July 10, 2025

Malik (2023)

Source : https://www.officetimeline.com/blog/artificial-intelligence-ai-and-chatgpt-history-and-timelines

History of Artificial Intelligence (AI) timeline

The Turing Test 1950

The term "artificial intelligence" is coined 1956

The term "machine learning" is coined 1959

The term "AI winter" is coined 1984

CNNs applicability demonstrated 1989

Deep Blue defeats Garry Kasparov 1997

A Neural Probabilistic Language Model 2000

Apple releases Siri 2011

Turing Natural Language Generation 2019

ChatGPT 2022

1950 | 1957 | 1964 | 1971 | 1978 | 1985 | 1992 | 1999 | 2006 | 2013 | 2020

1950-1979

- SNARC 1951
- ELIZA and Shakey 1966
- Samuel Checkers-Playing Program 1952
- Artificial Intelligence: A General Survey 1973
- The Logic Theorist 1956
- The General Problem Solver 1957
- The first AI winter begins 1974

1980-2009

- Symbolics Lisp machines commercialized 1980
- Fifth Generation Computer Systems 1982
- Large-Scale Deep Unsupervised Learning 2009
- Bayesian networks causal analysis 1985
- A Statistical Approach to Language Translation 1988

2010-2014

- Watson wins on Jeopardy! 2011
- Googles's deep neural network project 2012
- China's Tianhe-2 fastest system 2013
- Facebook introduces DeepFace 2014

2015-2019

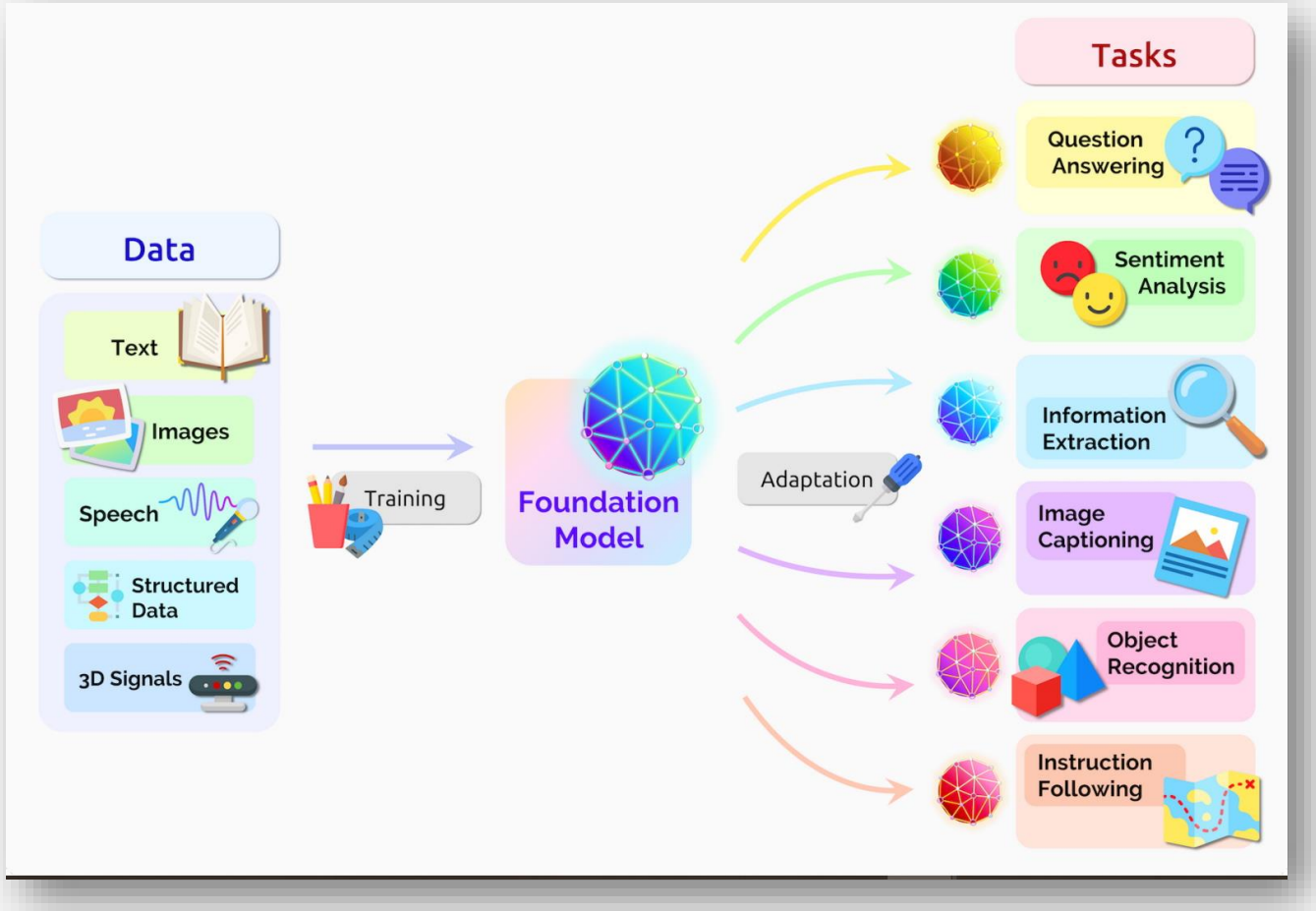
- AlphaGo wins at Go 2016
- Sophia: the first robot with citizenship 2017
- Google introduces the Transformer 2017
- Cimon, GPT, Lovot 2018

2020-2025

- Curial GPT-3 LLM 2020
- Dall-E 2021
- GPT-4 2023

What are foundation models?

The Stanford Institute for Human-Centered Artificial Intelligence's (HAI) for Research on Foundation Models (CRFM) coined the term "foundation model" in August 2021 to mean "any model that **is trained on broad data** (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to **a wide range of downstream tasks**



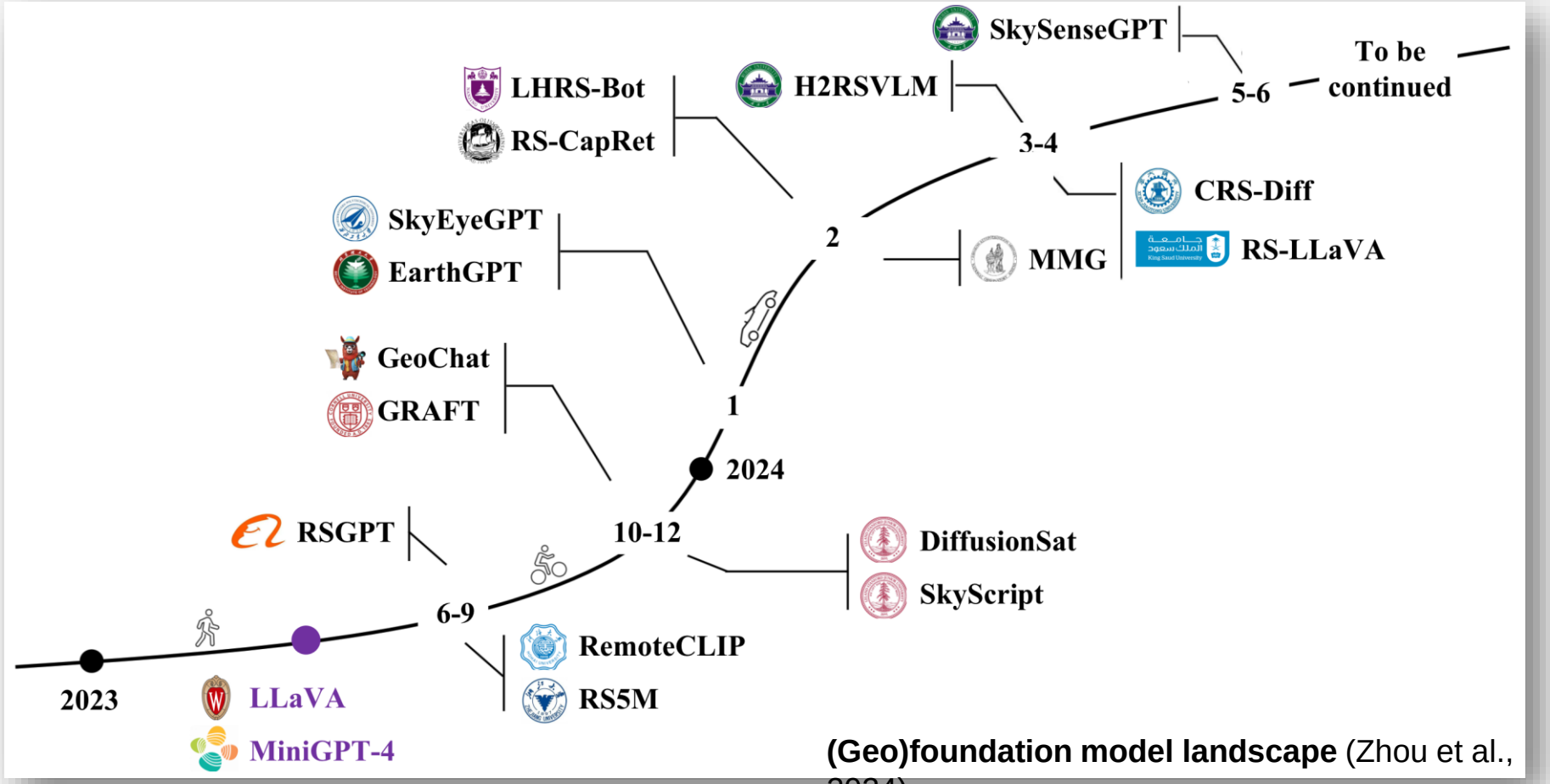
ChatGP T

Source: https://arxiv.org/pdf/2108.07258.pdf

What are Geo-foundation models?

Satellite imagery

Foundation models



(Geo)foundation model landscape (Zhou et al., 2024)

Geo-foundation models: foundation model trained specifically to understand satellite imagery

Questions to be addressed in this presentation

Study 1

- To what extent do translated textual descriptions **preserve the similarity** observed in the corresponding images?

Study 2

- How to design a **flexible SAM** framework for segmenting multimodal remote sensing data?

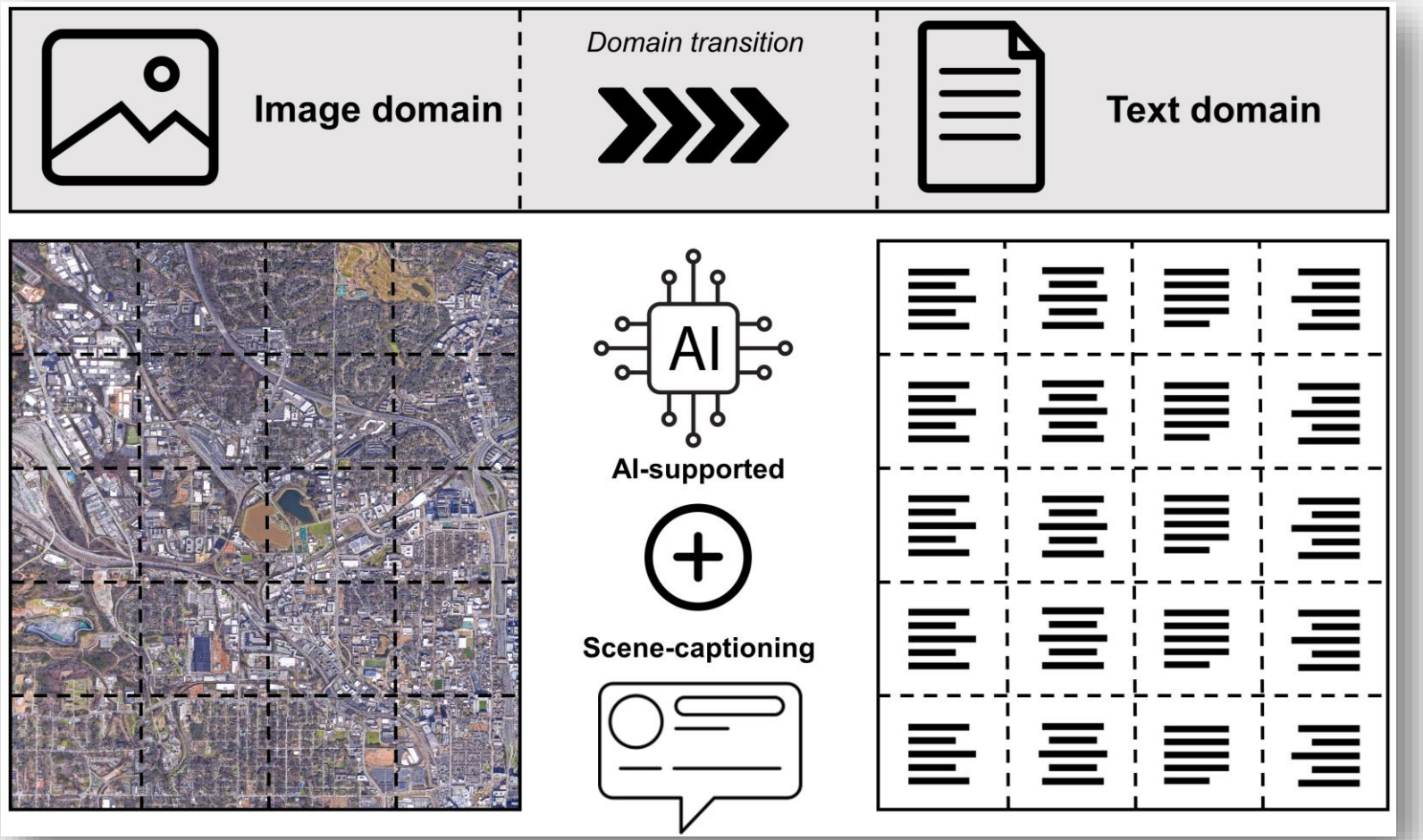
Study 3

- How to develop an **improved Visual-Language Model** for enhanced remote sensing image comprehension?

Study 4

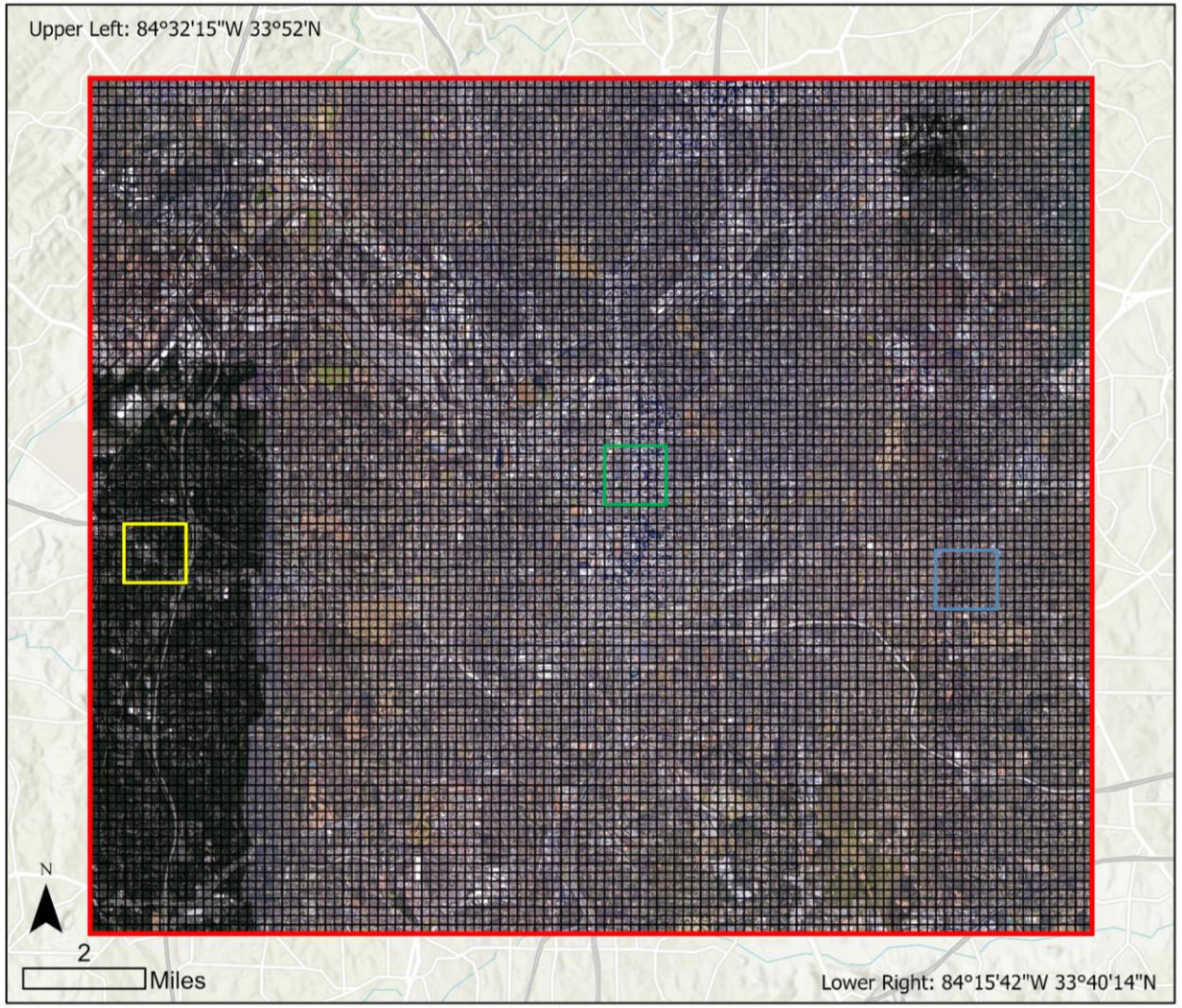
- How can **multimodal foundation models** be adapted into a **unified forecasting framework** to enhance predictive analytics of urban dynamics in complex environments?

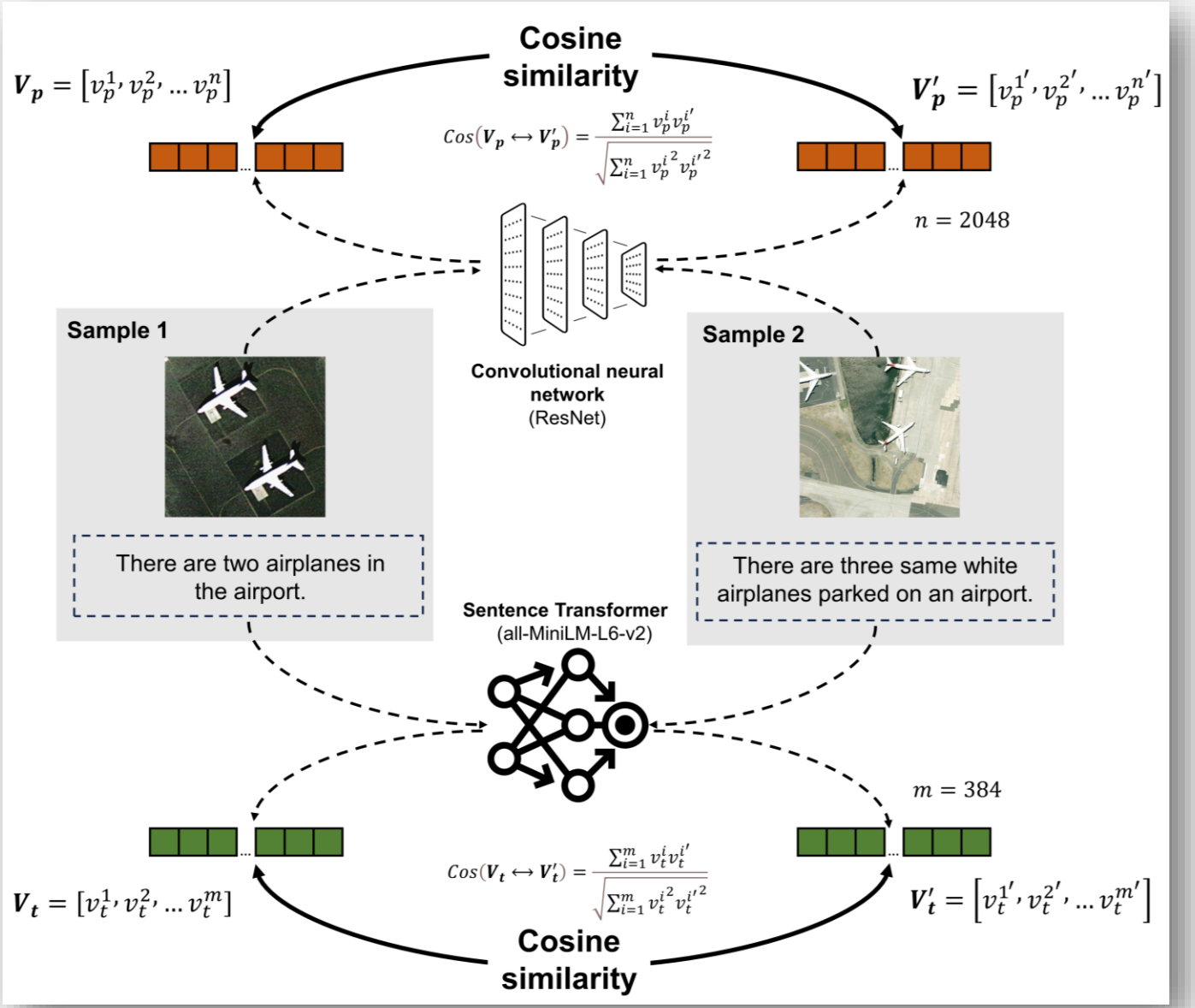
Understanding remote sensing imagery like reading a text document



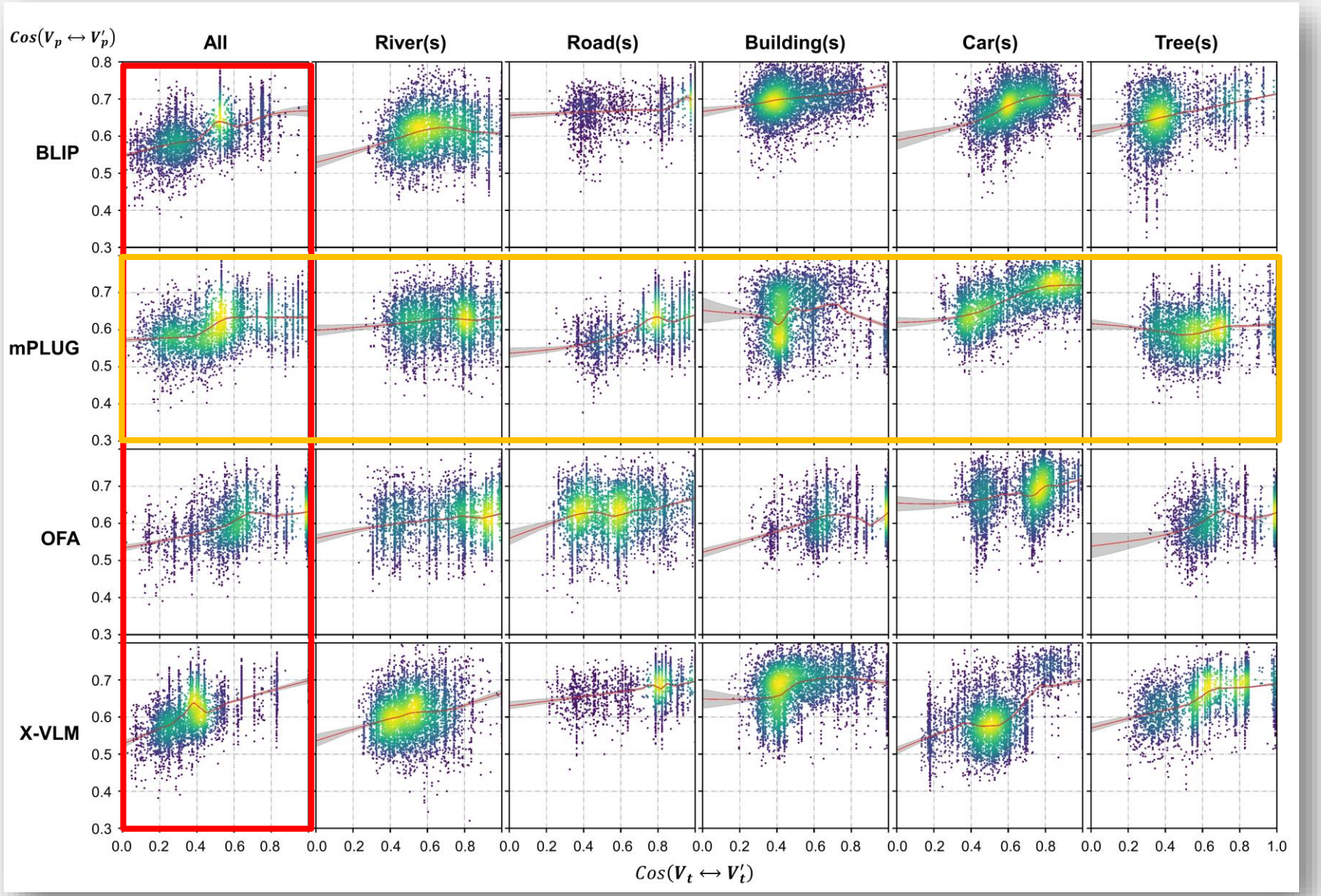
To what extent do translated textual descriptions **preserve the similarity** observed in the corresponding images?

Case study area




Cosine similarity
$V_p = [v_p^1, v_p^2, \dots v_p^n]$
$V'_p = [v_p^{1'}, v_p^{2'}, \dots v_p^{n'}]$
$Cos(V_p \leftrightarrow V'_p) = \frac{\sum_{i=1}^{n} v_p^i v_p^{i'}}{\sqrt{\sum_{i=1}^{n} v_p^{i\,2} v_p^{i'\,2}}}$
$n = 2048$
Convolutional neural network
(ResNet)
Sample 1
There are two airplanes in the airport.
Sample 2
There are three same white airplanes parked on an airport.
Sentence Transformer
(all-MiniLM-L6-v2)
$m = 384$
$V_t = [v_t^1, v_t^2, \dots v_t^m]$
$V'_t = [v_t^{1'}, v_t^{2'}, \dots v_t^{m'}]$
$Cos(V_t \leftrightarrow V'_t) = \frac{\sum_{i=1}^{m} v_t^i v_t^{i'}}{\sqrt{\sum_{i=1}^{m} v_t^{i\,2} v_t^{i'\,2}}}$
Cosine similarity

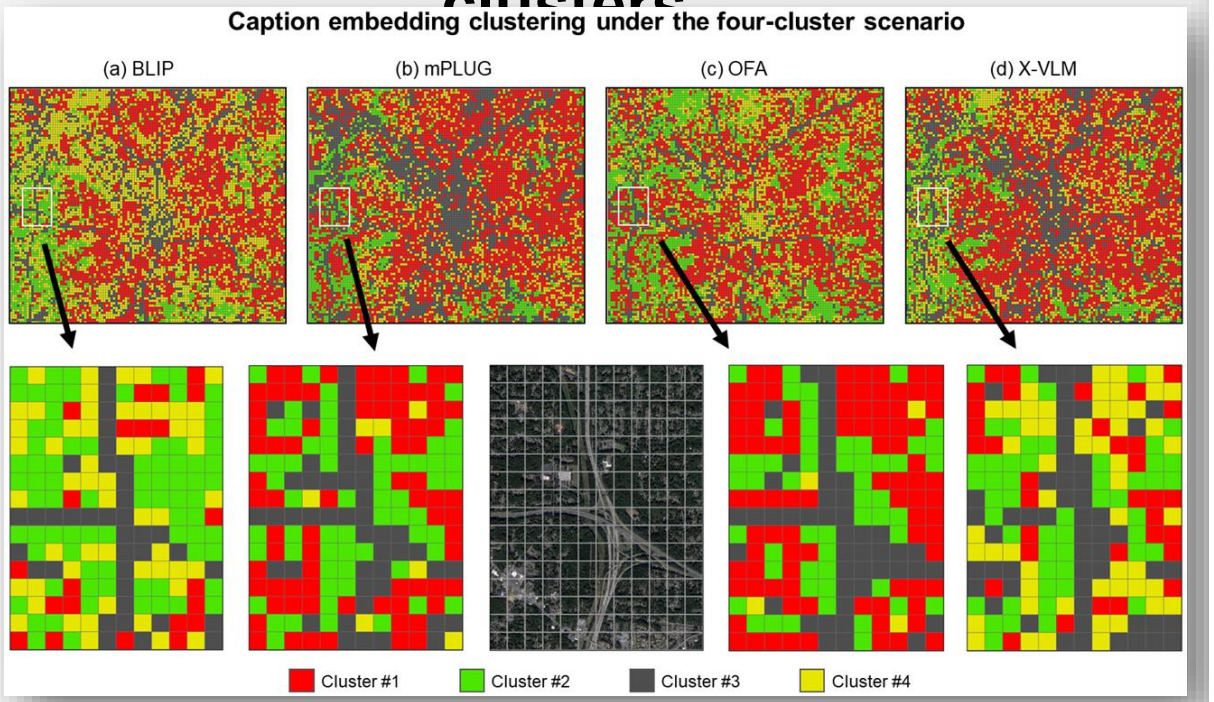
Results



The relationship between cosine similarity of caption embeddings, i.e., $Cos(\boldsymbol{V_t} \leftrightarrow \boldsymbol{V'_t})$ and cosine similarity of the corresponding image embeddings, i.e., $Cos(\boldsymbol{V_p} \leftrightarrow \boldsymbol{V'_p})$.

Results

Four clusters



Word cloud representations for the four investigated algorithms under the **four-cluster scenario**.

Results

- **Only a moderate cross-modal match overall.**
  Across the 11,270 Atlanta image patches, the correlation between image-space cosine similarity and caption-space cosine similarity never exceeds $r \approx 0.52$. Kendall's τ shows the same pattern (0.27 – 0.35). This indicates that the textual descriptions retain some—but far from all—of the visual similarity structure.
- **Model choice matters.**
  BLIP captions preserve the image similarity best, while mPLUG performs the worst; OFA and X-VLM sit in between. The gap ($\Delta r \approx 0.15$) confirms that preservation quality is model-dependent.

- **Preservation is object-specific.**
  When captions mention **cars**, visual–textual similarity strengthens markedly (e.g., mPLUG r = 0.563; X-VLM r = 0.517). In contrast, scenes dominated by **rivers** show almost no correlation (mPLUG r = 0.082; BLIP r = 0.097). Built features such as **roads** and **buildings** fall in between. Thus, the extent of preservation hinges on which urban element anchors the caption.

- **Similarity curves are non-linear and "wavy."**
  LOWESS plots reveal a fluctuating relationship: high-similarity image pairs are *sometimes* matched by high-similarity captions, but mid-range image similarities often scatter widely in caption space. This suggests that linguistic translation injects additional semantic nuance not present, or not weighted equally, in the raw pixels.

International Journal of Applied Earth Observation and Geoinformation 131 (2024) 103939

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Understanding remote sensing imagery like reading a text document: What can remote sensing image captioning offer?

Xiao Huang [a,*], Kaixuan Lu [b], Siqin Wang [c], Junyu Lu [d], Xiao Li [e], Ruiqian Zhang [f]

[a] Department of Environmental Sciences, Emory University, Atlanta, GA 30322, USA
[b] Piesat Information Technology Co., Ltd., Beijing 100195, China
[c] Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089, USA
[d] School of Community Resources and Development, Arizona State University, Phoenix, AZ 85004, USA
[e] Transport Studies Unit, University of Oxford, Oxford OX1 3QY, UK
[f] Institute of Photogrammetry and Remote Sensing at Chinese Academy of Surveying and Mapping, Beijing 100036, China

ARTICLE INFO

Keywords:
Remote sensing
Domain transfer
Image caption

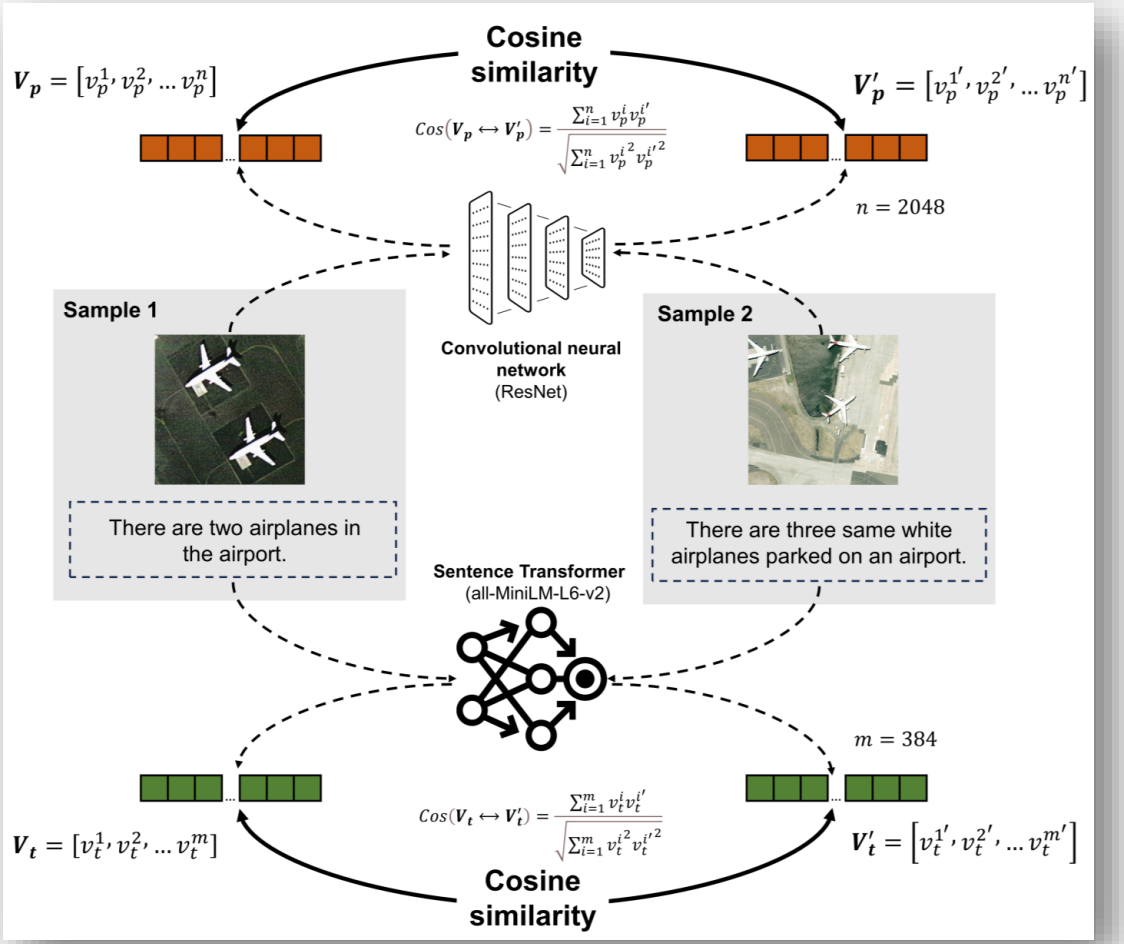
ABSTRACT

Remote sensing imagery offers intricate and nuanced data, emphasizing the need for a profound understanding of the relationships among varied geographical elements and events. In this study, we explore the transitions from the image domain to the text domain by employing four state-of-the-art image captioning algorithms, i.e., BLIP, mPLUG, OFA, and X-VLM. Specifically, we investigate (1) the stability of these image captioning algorithms for remote sensing image captioning, (2) the preservation of similarity between images and their corresponding captions, and (3) the characteristics of their caption embedding spaces. The results suggest a moderate consistency across generated captions from different image captioning models, with observable variations contingent upon the urban entities presented. In addition, a dynamic relationship emerges between image space and the corresponding caption space, evidenced by their fluctuated correlation coefficient. Most importantly, patterns within the caption embedding space align with the observed land cover and land use in the image patches, reaffirming the potential of our pilot work as an impactful analytical approach in future remote sensing analytics. We advocate that integrating image captioning techniques with remote sensing imagery paves the way for an innovative data extraction and interpretation approach with diverse applications.

1. Introduction

Remote sensing imagery, as a sophisticated and versatile resource, plays a critical role across a multitude of sectors due to its capacity to generate expansive and real-time data from a remote perspective. This technology finds its application in numerous fields, such as agriculture (Khanal et al., 2020), meteorology (Tomlinson et al., 2011), environmental conservation (Turner et al., 2003), and urban planning (Netzband et al., 2007), among others. The comprehension of remote sensing imagery has witnessed substantial progression via the inception of cutting-edge techniques and methodologies. The advancement in Geographic Information Systems (GIS) has facilitated superior geospatial analysis and visual representation of remote sensing data. Further, the adoption of Machine Learning (ML) and Artificial Intelligence (AI) has been escalated to augment the interpretation and scrutinization of imagery data (Zhu et al., 2017). Complex algorithms, specifically Convolutional Neural Networks (CNNs), are now deployed to discern patterns and features in images, thereby bolstering object recognition and classification (Song et al., 2019; Ren et al., 2018). The ceaseless evolution of these methodologies is instrumental in deriving more insightful data from remote sensing imagery, catalyzing advancements across diverse domains.

As widely recognized in the field, remote sensing imagery presents intricate and diversified data, requiring extensive comprehension of the interactions among various geographical aspects and phenomena (Dronova, 2015). Such complexity often surpasses the capacity of traditional image processing methods and even transcends the capabilities of advanced image understanding mechanisms. Moreover, nowadays a huge volume of satellite images is increasingly generated every day, especially with more and more commercial companies launching



Revisiting SAM-based semantic segmentation models

Notable efforts for RS land use and land cover segmentation:

- **RSPrompter** *(Chen et al., 2024)*: uses an anchor-and-query-based prompt generator and multi-scale feature enhancer to extract features from the SAM encoder, producing distinct segmentation results
- **CWSAM** *(Pu et al., 2024)*: fine-tunes the SAM encoder with lightweight adapters for SAR imagery, adding a classwise mask decoder with a class prediction head for precise pixel-level land cover classification

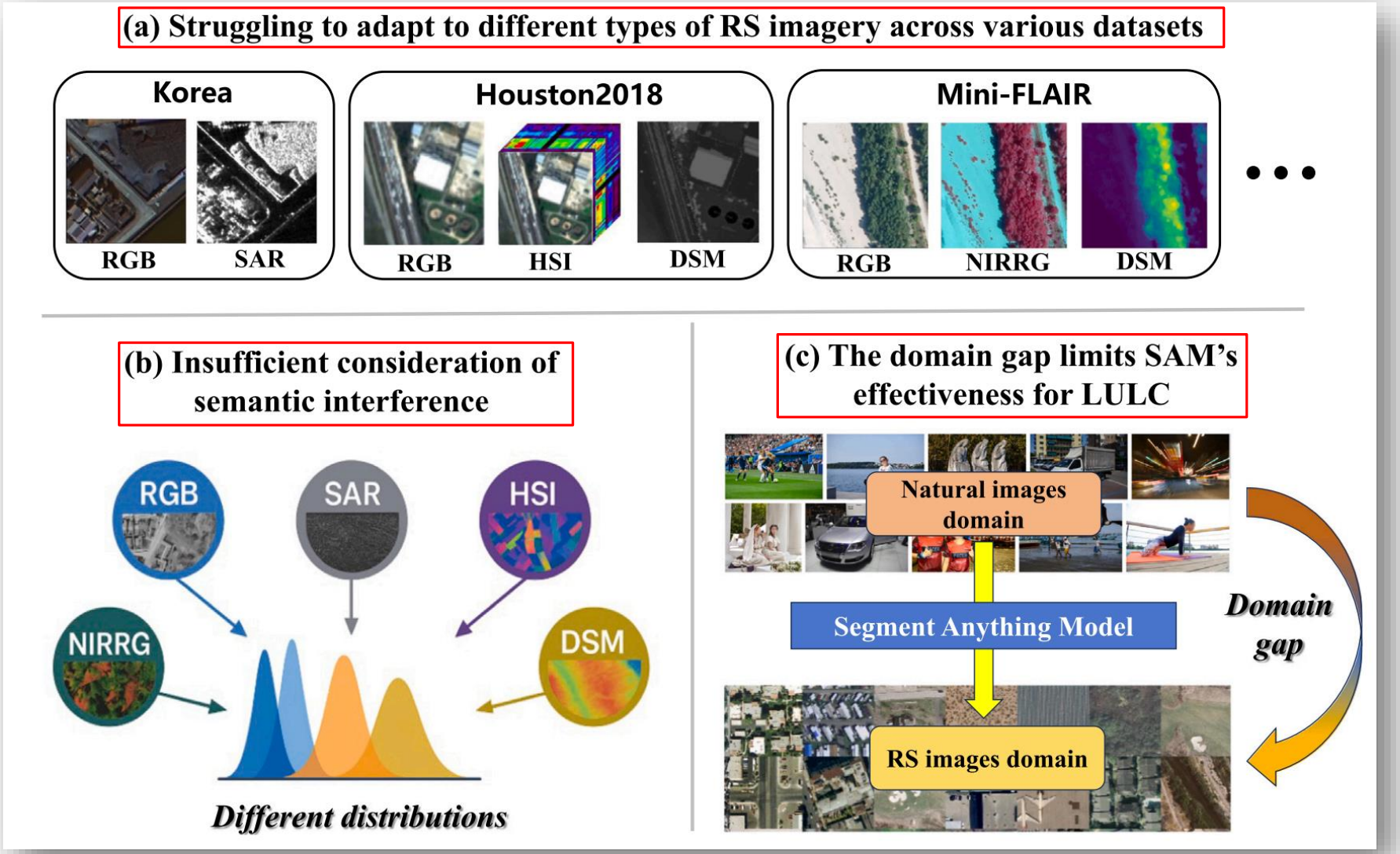
Notable multimodal efforts:

- **RingMo-SAM** *(Yan et al., 2023)*: adapts SAM for multimodal RS data, using an embedding feature prompt encoder and multi-box prompts.

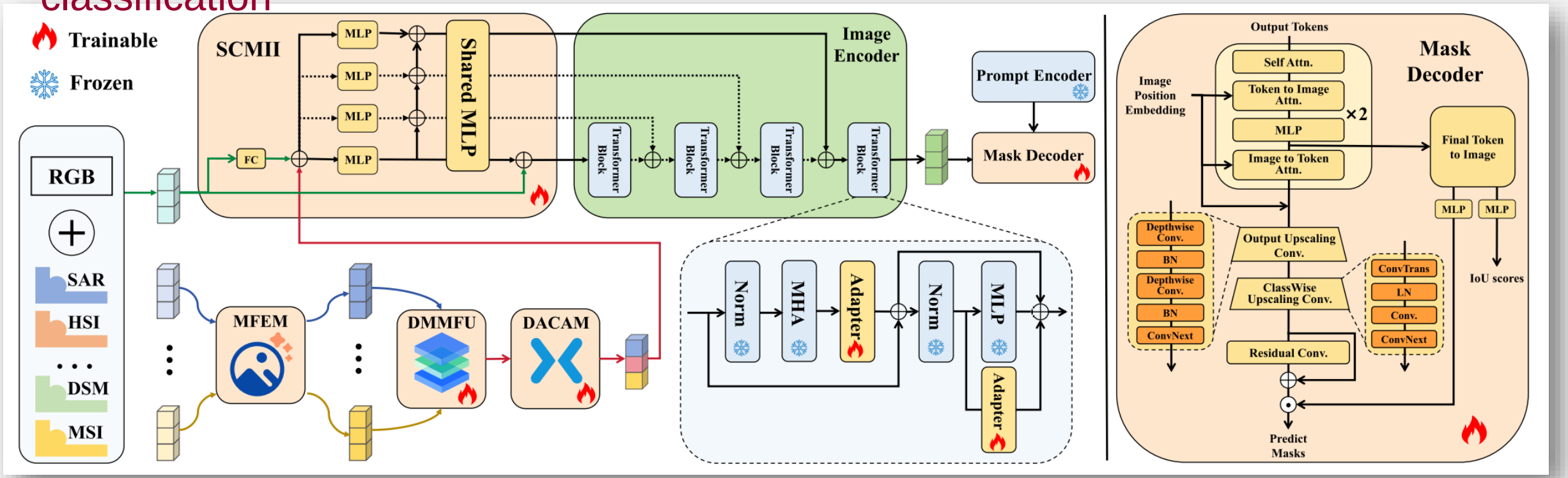
- **SAM-MCD** *(Ding et al., 2024)***:** leverages SAM's zero-shot transfer to generate segmentation maps for optical imagery and uses OpenStreetMap (OSM) data with a connected component labeling algorithm to identify LULC changes

SAM-based multimodal models still struggle with **adaptability** and **accuracy** issues due to information interference between modalities

Challenges in applying SAM-based multimodal semantic segmentation models to high-precision LULC classification



FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification



Key modules: 1) Multimodal Feature Enhancement Module (MFEM)

2) Dynamic Multimodal Feature Fusion Unit (DMMFU)

3) Dynamic Attention and Context Aggregation Mixer (DACAM)

4) Semantic Cross-Modal Integration Module (SCMII)

FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification

Key modules: 1) **Multimodal Feature Enhancement Module (MFEM)**

Cleans and amplifies each input modality with **domain-specific filters** (e.g., SAR speckle suppression, HSI PCA), producing noise-reduced, standardized feature maps that give the downstream network a solid, modality-aware starting point

2) **Dynamic Multimodal Feature Fusion Unit (DMMFU)**

Uses **lightweight multi-head attention** to learn per-scene confidence scores for every auxiliary modality, then **fuses the most informative channels with RGB,** so complementary signals are emphasized and irrelevant noise is suppressed.

3) **Dynamic Attention and Context Aggregation Mixer (DACAM)**

Refines the fused tensor through **multi-scale convolutional kernels and context-aware attention**, jointly modeling local details and broad spatial relationships to deliver semantically richer, scale-robust features.

4) **Semantic Cross-Modal Integration Module (SCMII)**

Applies **a stack of shared MLP layers** to **align heterogeneous embeddings**, blending auxiliary and RGB streams into a single, semantically consistent representation that the SAM encoder can ingest without architectural changes.

FlexiSAM : PyTorch

FlexiSAM* : LuoJiaNET*(Zhang et al., 2023)*

Table 1

Quantitative comparison of multiple SOTA semantic segmentation models on the Korea dataset, including RGB and SAR modalities. Models are evaluated on land cover categories: building, road, farmland, water, and greenery. 'N/A' indicates models not adaptable to the modalities in this dataset. FlexiSAM* indicates accuracies obtained using the LuoJiaNET framework, while all other methods are based on PyTorch. For consistent comparison in PyTorch, bold values indicate the highest accuracy, and underlined values represent the second-highest.

Method	Modal	Category Accuracy	mIoU (%)	mF1 (%)	OA (%)
		building, road, farmland, water, greenery			
UNetFormer (Wang et al., 2022c)	RGB	95.43, 78.57, 92.11, 99.16, 90.57, 81.61	80.87	88.69	93.10
SwinT-V2 (Liu et al., 2022a)	RGB	95.03, 84.20, 91.81, 99.17, 90.80, 86.24	82.54	93.36	89.95
Segformer (Xie et al., 2021)	RGB	97.01, 86.58, 95.49, 99.19, 94.25, 89.90	87.56	93.08	95.70
SOLC (Li et al., 2022b)	RGB+SAR	94.70, 89.15, 93.90, 99.15, 93.72, 89.48	85.61	90.95	94.64
MFT (Roy et al., 2023)	HSI+DSM	N/A	N/A	N/A	N/A
FTransUNet (Ma et al., 2024b)	RGB+DSM	N/A	N/A	N/A	N/A
UisNet (Fan et al., 2022)	RGB+SAR	95.26, 87.02, 96.64, 99.45, 96.90, 63.55	82.49	89.60	93.31
MSSeg (Wang et al., 2024a)	RGB+SAR	97.90, 88.01, 96.63, 99.40, 96.56, 92.37	90.26	94.67	94.88
CWSAM (Pu et al., 2024)	SAR	95.82, 74.73, 93.68, 98.87, 93.61, 86.51	82.09	89.33	94.10
RSAM-Seg (Zhang et al., 2024e)	RGB	96.46, 83.30, 95.87, 99.47, 95.72, 90.42	86.95	92.62	95.66
SAM-RS (Ma et al., 2024a)	RGB	97.08, 88.36, 96.33, 99.34, 93.42, 89.62	87.92	93.28	96.01
SEFM (Shi et al., 2023)	RGB+SAR	97.65, 86.44, 96.83, 99.49, 96.56, 93.05	89.97	94.49	96.78
CMAA (Shi et al., 2023)	RGB+SAR	97.30, 87.45, 96.94, 99.49, 96.33, 91.78	89.47	94.19	96.67
FlexiSAM (Ours)	RGB+SAR	**98.13, 91.13, 97.42, 99.55, 97.36, 93.73**	**91.84**	**96.08**	**97.53**
FlexiSAM* (Ours)	RGB+SAR	98.40, 91.80, 97.80, 99.65, 97.70, 94.20	92.25	96.45	97.90

Table 2

Quantitative comparison of multiple SOTA semantic segmentation models on the Houston2018 dataset, including RGB, HSI, and DSM modalities. Models are evaluated on land cover categories: Healthy Grass (HG), Stressed Grass (SG), Artificial Turf (AT), Evergreen Trees (ET), Deciduous Trees (DT), Bare Earth (BE), Water (W), Residential Buildings (RB), Non-Residential Buildings (NR), Roads (R), Sidewalks (S), Crosswalks (CW), Major Thoroughfares (MT), Highways (H), Railways (RW), Paved Parking Lots (PP), Unpaved Parking Lots (UP), Cars (C), Trains (T), and Stadium Seats (SS). 'N/A' indicates models not adaptable to the modalities in this dataset. FlexiSAM* indicates accuracies obtained using the LuoJiaNET framework, while all other methods are based on PyTorch. For consistent comparison in PyTorch, bold values indicate the highest accuracy, and underlined values represent the second-highest.

Method	Modal	Category Accuracy	mIoU (%)	mF1 (%)	OA (%)
		HG, SG, AT, ET, DT, BE, W, RB, NR, R, S, CW, MT, H, RW, PP, UP, C, T, SS			
UNetFormer (Wang et al., 2022c)	RGB	82.65, 87.28, 98.03, 93.08, 90.93, 98.28, 99.25, 95.58, 94.96, 87.88, 85.72, 62.87, 88.74, 91.10, 90.73, 91.07, 90.85, 89.44, 94.77, 96.56	75.45	85.15	92.09
SwinT-V2 (Liu et al., 2022a)	RGB	75.09, 89.50, 99.45, 93.96, 86.50, 97.51, 84.32, 98.50, 97.25, 87.06, 83.06, 48.44, 92.57, 97.54, 86.94, 81.97, 94.96, 73.88, 91.24, 99.42	77.61	85.58	92.92
Segformer (Xie et al., 2021)	RGB	79.56, 90.29, 95.34, 95.18, 88.44, 92.23, 96.47, 97.04, 97.31, 86.06, 82.53, 51.95, 90.55, 95.81, 91.50, 94.03, 91.47, 92.80, 92.70, 98.94	78.73	86.58	93.44
SOLC (Li et al., 2022b)	RGB+SAR	N/A	N/A	N/A	N/A
MFT (Roy et al., 2023)	HSI+DSM	83.23, 91.32, 95.97, 95.97, 92.45, 97.49, 96.57, 96.78, 97.55, 89.83, 88.67, 64.94, 91.88, 95.64, 94.96, 94.32, 93.03, 93.65, 95.40, 97.32	82.43	89.50	94.69
FTransUNet (Ma et al., 2024b)	RGB+DSM	N/A	N/A	N/A	N/A
UisNet (Fan et al., 2022)	RGB+HSI+DSM	88.93, 95.11, 95.44, 97.78, 96.37, 99.28, 99.41, 99.37, 99.10, 93.46, 92.70, 68.16, 96.06, 98.78, 98.62, 99.24, 97.71, 99.02, 99.34, 98.86	85.7	90.05	97.34
MSSeg (Wang et al., 2024a)	RGB+HSI+DSM	85.45, 93.89, 99.82, 97.22, 95.42, 99.74, 100.00, 99.10, 98.91, 92.67, 91.97, 64.62, 94.29, 97.55, 99.34, 98.67, 99.68, 98.37, 99.42, 98.75	86.91	91.42	96.76
CWSAM (Pu et al., 2024)	SAR	N/A	N/A	N/A	N/A
RSAM-Seg (Zhang et al., 2024e)	RGB	79.52, 90.37, 93.63, 95.36, 87.66, 91.67, 96.16, 96.85, 97.29, 86.01, 82.75, 51.70, 90.56, 95.67, 92.65, 93.96, 90.50, 92.76, 94.16, 98.41	78.71	86.55	93.43
SAM-RS (Ma et al., 2024a)	RGB	84.41, 92.73, 96.68, 96.52, 93.04, 95.06, 98.54, 98.42, 98.04, 90.92, 89.33, 63.83, 92.91, 96.10, 94.46, 95.27, 94.75, 93.49, 95.74, 98.40	83.71	90.02	95.39
SEFM (Shi et al., 2023)	RGB+HSI+DSM	80.64, 92.20, **99.94**, 96.82, 92.15, 99.53, 92.08, 99.55, 98.63, 91.85, 90.24, 58.25, 95.04, 98.62, 94.60, 94.08, 99.45, 91.31, 96.55, 99.72	86.87	91.80	95.99
CMAA (Shi et al., 2023)	RGB+HSI+DSM	**90.08**, **96.27**, 99.68, **99.32**, 98.18, **99.99**, **100.00**, 99.76, 99.62, 97.17, **97.43**, 80.14, 97.33, 99.15, 99.69, 99.66, **99.98**, 99.25, 99.67, 99.00	88.46	92.29	98.61
FlexiSAM (Ours)	RGB+HSI+DSM	89.68, 96.18, 99.47, **99.32**, **98.69**, 99.83, 99.63, **99.85**, **99.64**, **98.00**, 97.34, **80.85**, **97.79**, **99.47**, **99.75**, **99.77**, 99.72, **99.64**, **99.80**, **99.91**	**89.23**	**93.94**	**98.73**
FlexiSAM* (Ours)	RGB+HSI+DSM	89.70, 96.25, 99.50, 99.35, 98.75, 99.80, 99.70, 99.85, 99.65, 97.95, 97.45, 80.75, 97.85, 99.48, 99.70, 99.78, 99.73, 99.68, 99.81, 99.92	89.40	94.10	98.85

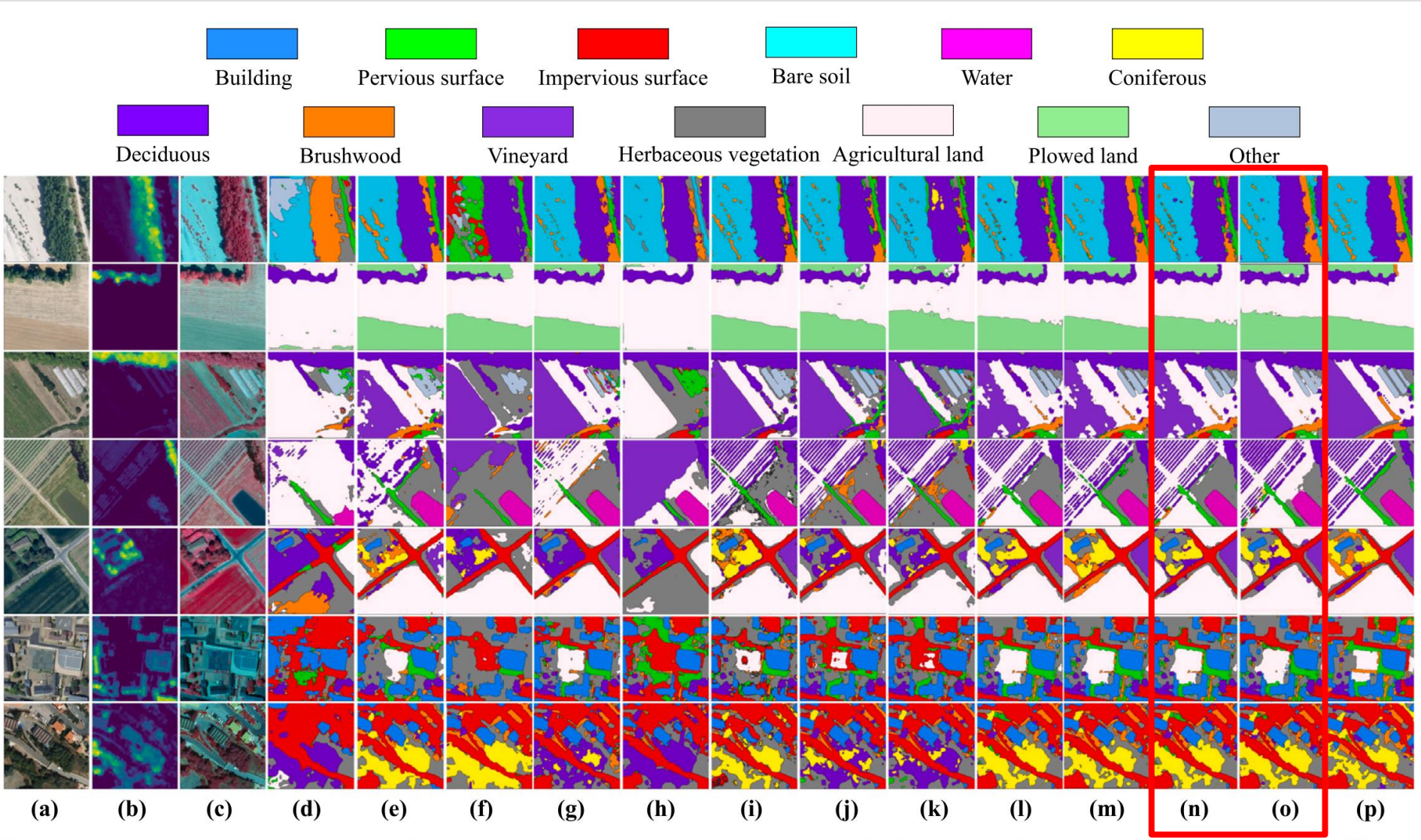


Fig. 8. Visualization results on the Korea dataset for multiple SOTA semantic segmentation models: (a) RGB modality, (b) SAR modality, (c) UNetFormer (Wang et al., 2022c), (d) SwinT-V2 (Liu et al., 2022a), (e) Segformer (Xie et al., 2021), (f) SOLC (Li et al., 2022b), (g) UisNet (Fan et al., 2022), (h) MSSeg (Wang et al., 2024a), (i) CWSAM (Pu et al., 2024), (j) RSAM-Seg (Zhang et al., 2024e), (k) SAM-RS (Ma et al., 2024a), (l) SEFM (Shi et al., 2023), (m) CMAA (Shi et al., 2023), (n) FlexiSAM (ours), (o) FlexiSAM* (ours), and (p) Ground truth.

ISPRS Journal of Photogrammetry and Remote Sensing 227 (2025) 594–612

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification using high-resolution multimodal remote sensing imagery

Zhan Zhang [a,1], Daoyu Shu [b,1], Cunyi Liao [c], Chengzhi Liu [d], Yuanxin Zhao [b], Ru Wang [e], Xiao Huang [f], Mi Zhang [b,g,*], Jianya Gong [b,**]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[c] Department of Electronics and Information Engineering, College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China
[d] School of Computer Science, Wuhan University, Wuhan 430072, China
[e] School of Urban Design, Wuhan University, Wuhan 430072, China
[f] Department of Environmental Sciences, Emory University, Atlanta, GA 30322, USA
[g] Hubei Luojia Laboratory, Wuhan 430079, China

ARTICLE INFO

Keywords:
Land use and land cover (LULC) classification
Multimodal remote sensing (RS) imagery
Semantic segmentation model
Segment anything model (SAM)

ABSTRACT

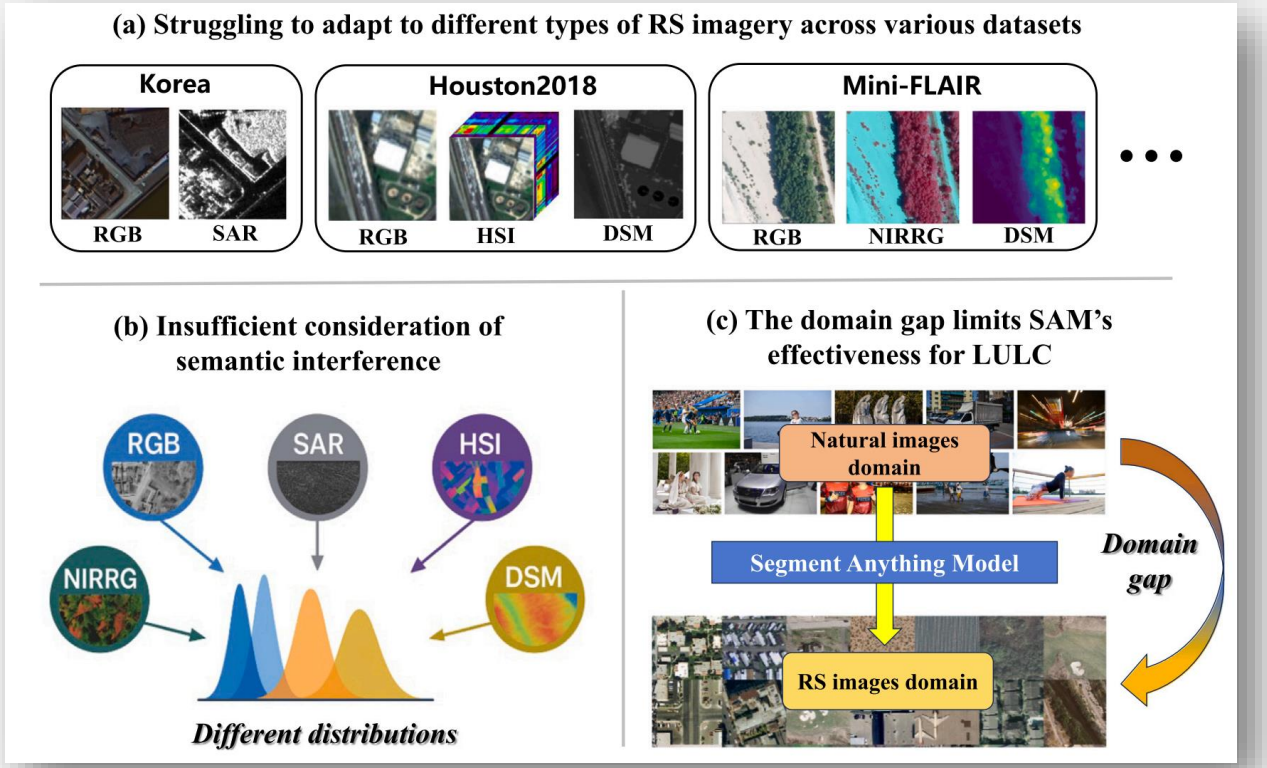
Fine-grained land use and land cover (LULC) classification using high-resolution remote sensing (RS) imagery is fundamental to scientific research. Recently, the Segment Anything Model (SAM) has emerged as a major advance in deep learning-based LULC classification due to its robust segmentation and generalization capabilities. However, existing SAM-based models predominantly rely on single-modal inputs (e.g., optical RGB or SAR), limiting their ability to fully capture the complex spatial and spectral characteristics of RS imagery. Although multimodal RS data can provide complementary information to enhance classification accuracy, integrating multiple modalities into SAM presents significant challenges, including modality adaptation, semantic interference, and domain gaps. Building on this, we propose FlexiSAM, a SAM-based multimodal semantic segmentation model designed to overcome these challenges. FlexiSAM uses RGB as the primary modality while seamlessly integrating auxiliary RS modalities through a modular pipeline. Key innovations include the Dynamic Multimodal Feature Fusion Unit (DMMFU) and Dynamic Attention and the Context Aggregation Mixer (DACAM) for robust cross-modal feature fusion and refinement, and the Semantic Cross-Modal Integration Module (SCMII) for mitigating modality-induced feature misalignments and ensuring coherent multimodal integration. These are then processed by the adapted SAM encoder, enhanced with a lightweight adapter tailored for RS data, and followed by a dedicated decoder that produces precise classification outputs. Extensive experiments on the Korea, Houston2018, and Mini-FLAIR datasets, conducted using LuoJiaNET for core evaluations and PyTorch for cross-method comparisons, demonstrate FlexiSAM's effectiveness and superiority, surpassing state-of-the-art models by at least 1.58% on Korea, 0.77% on Houston2018, and 1.14% in mIoU. Importantly, the LuoJiaNET framework delivers higher accuracy and efficiency compared to PyTorch. FlexiSAM also demonstrates strong adaptability and robustness across diverse RS modalities, establishing it as a versatile solution for fine-grained LULC classification.

1. Introduction

Fine-grained land use and land cover (LULC) classification, which systematically identifies and quantifies how Earth's surface is utilized and covered, is critical in fields such as climate change research, urban planning, and disaster management (Wang et al., 2023, 2022b; Li et al., 2024a; Vali et al., 2020). In intelligent remote sensing (RS) interpretation, fully supervised single-modal semantic segmentation models have excelled in LULC classification due to their robust end-to-end feature extraction and representation capabilities (Diakogiannis et al., 2020; Gao et al., 2021; He et al., 2022b). Notable models include CNN-based architectures like UNet (Ronneberger et al., 2015) and DeepLabV3Plus (Chen et al., 2018), ViT-based models such as

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.
** Corresponding author.
E-mail addresses: zhangzhanstep@whu.edu.cn (Z. Zhang), mizhang@whu.edu.cn (M. Zhang), gongjy@whu.edu.cn (J. Gong).
[1] These authors contributed equally to this work.

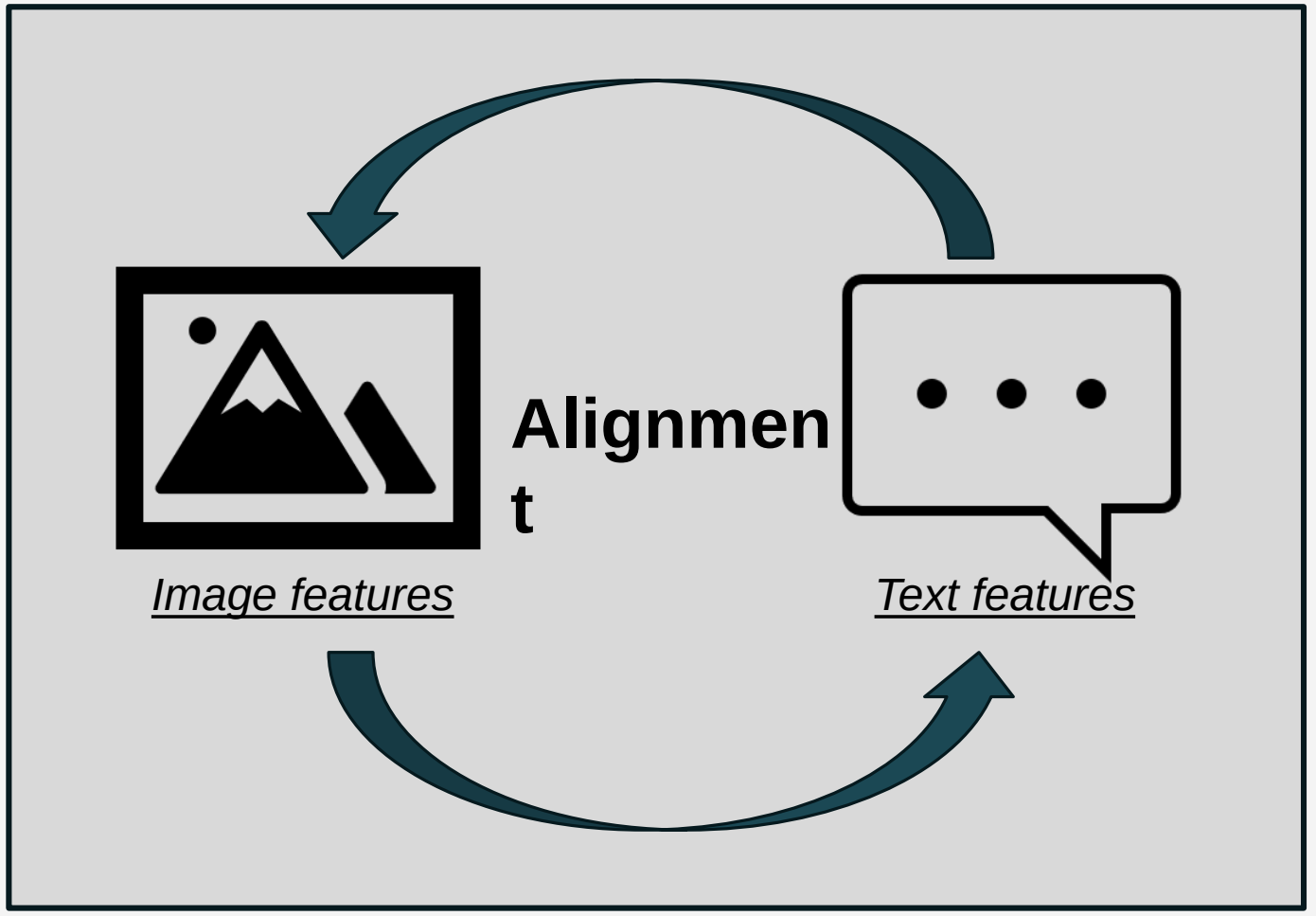
https://doi.org/10.1016/j.isprsjprs.2025.05.028
Received 17 November 2024; Received in revised form 23 May 2025; Accepted 23 May 2025
Available online 28 June 2025
0924-2716/© 2025 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

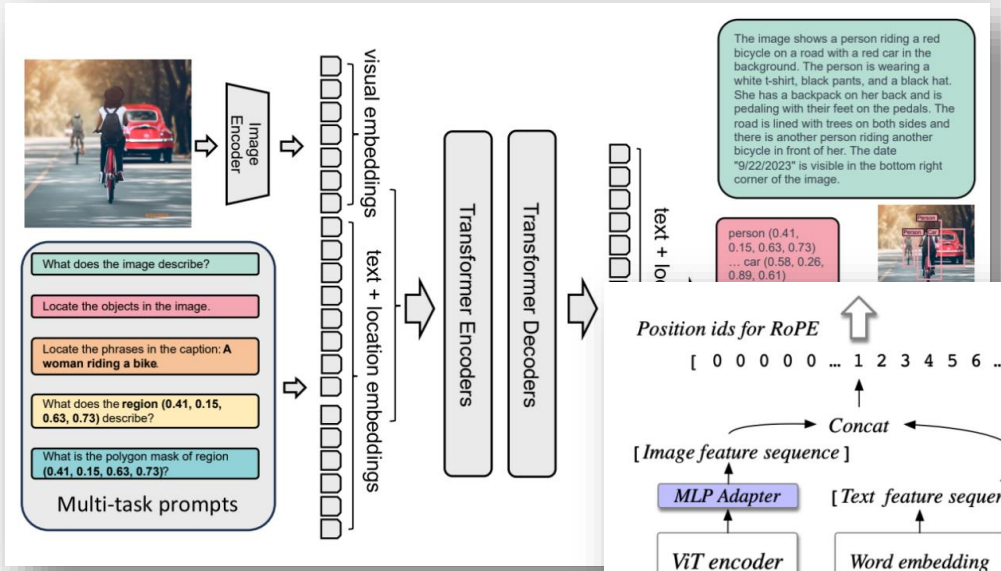


Remote Sensing Foundational Models

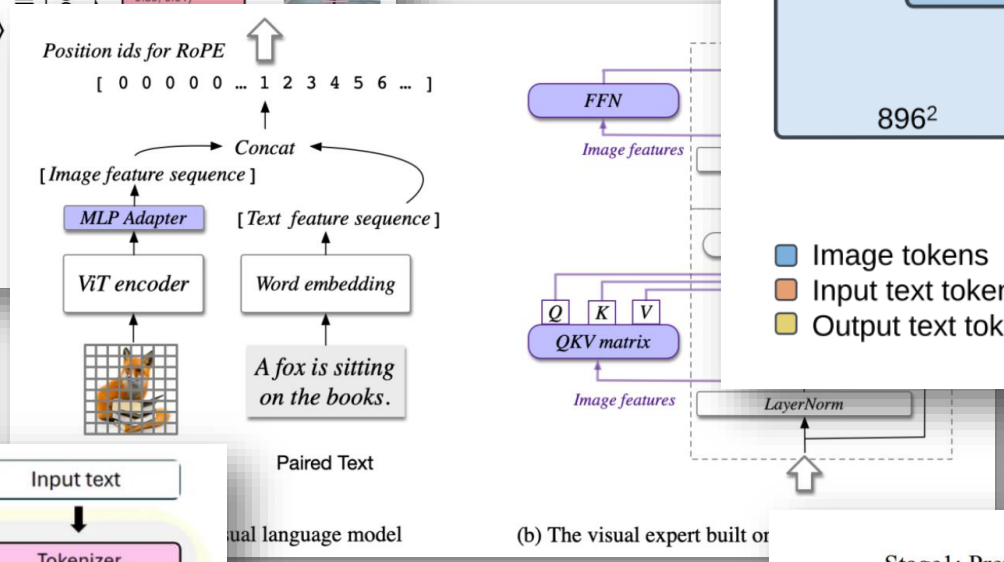
Satellite imagery

Foundation models (Multi-modality)

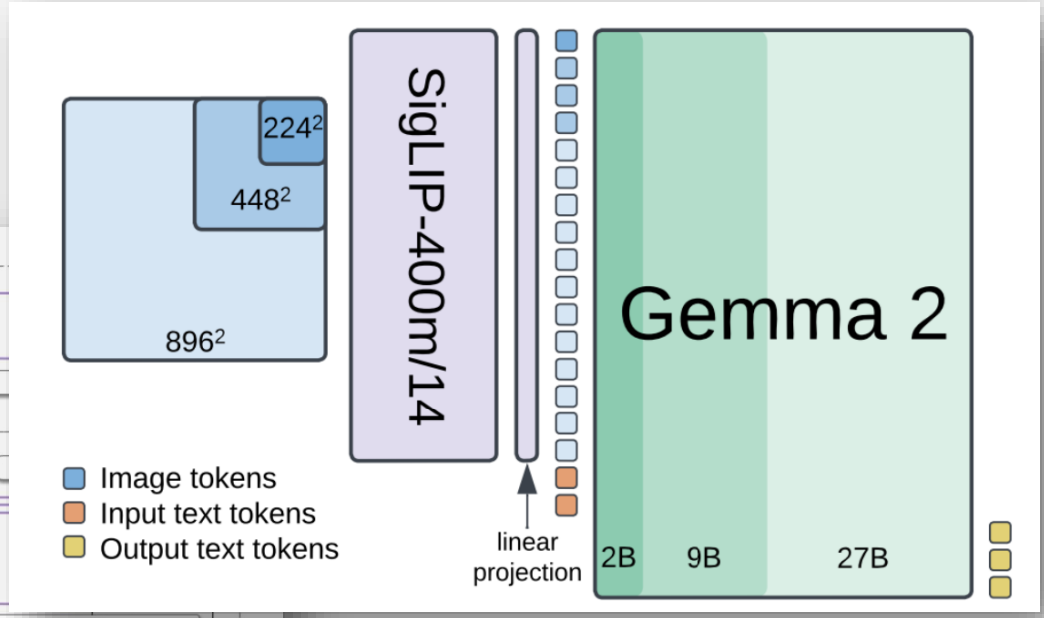




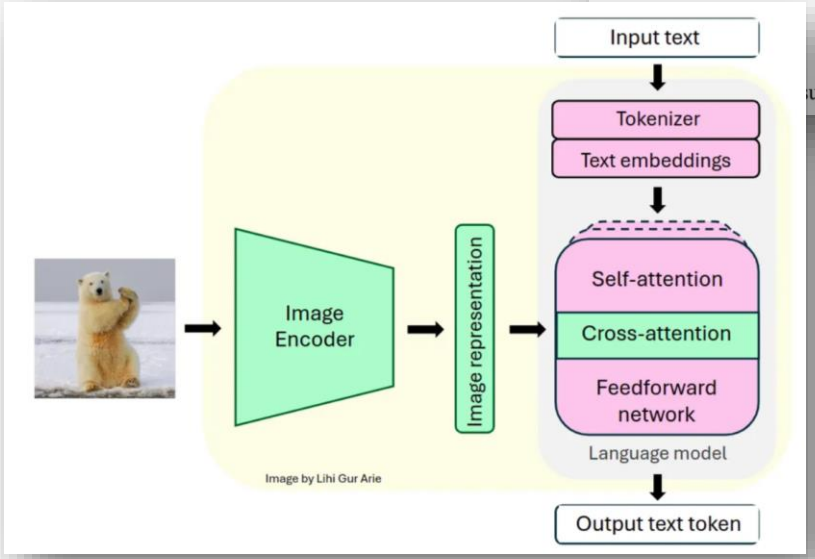
Florence-2



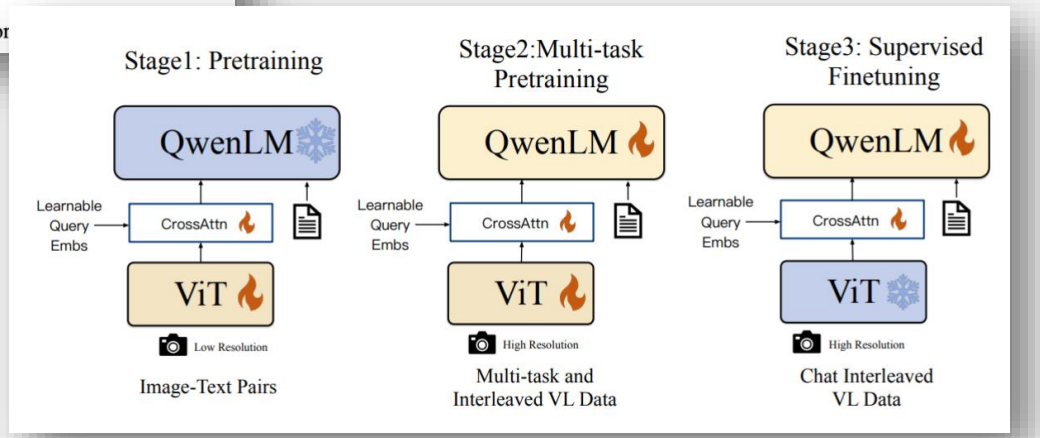
CogVLM



PaliGemma-2



Llama3.2-Vision



Qwen-VL

Meet Aquila

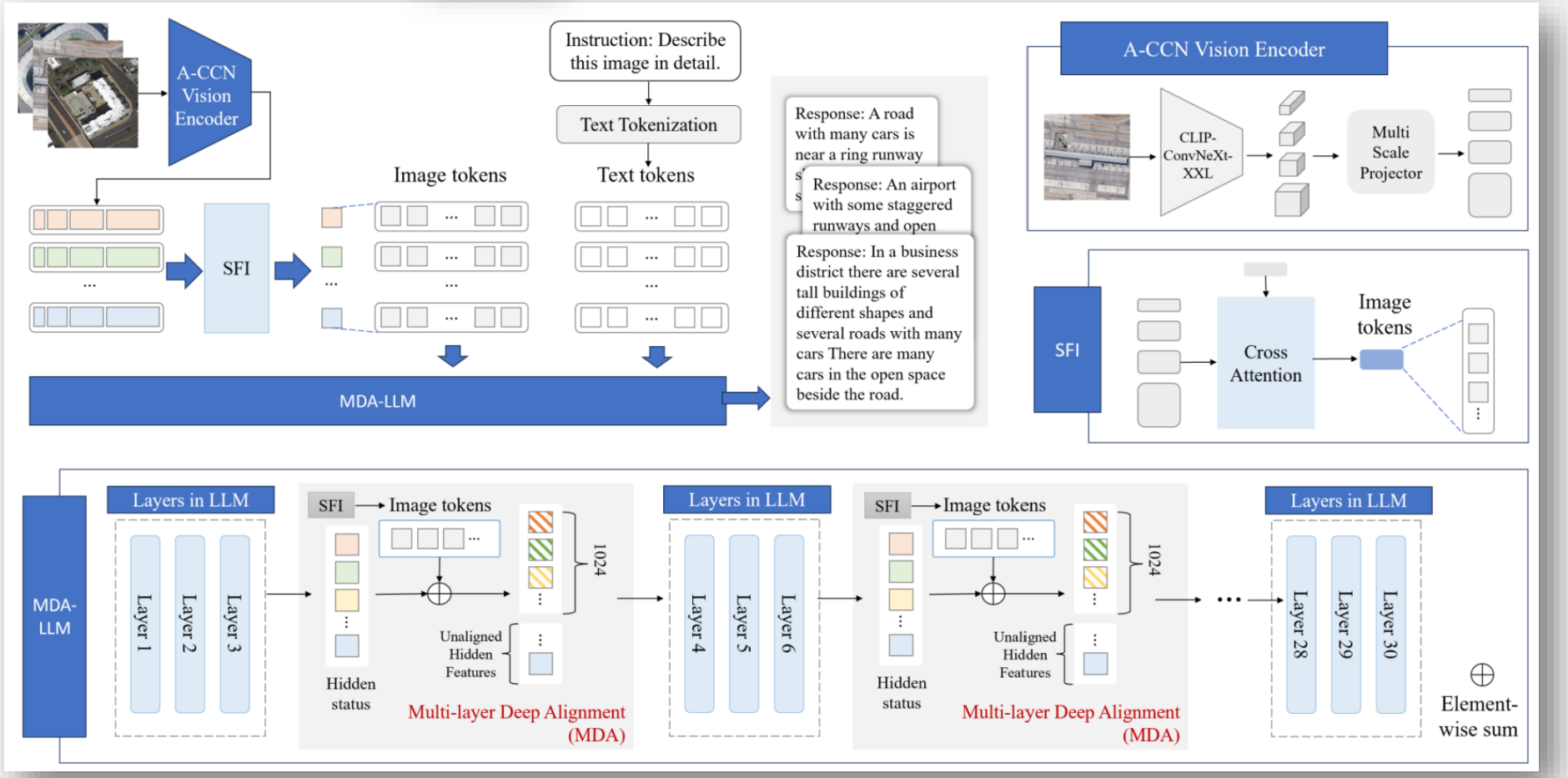
A Hierarchically Aligned Visual-Language Model for Enhanced Remote Sensing Image Comprehension

- **Aquila-CCN Vision Encoder (ConvNeXt-CLIP backbone)**
  - Most RSVLMs rely on ViT-based CLIP encoders that struggle with very large inputs. Aquila-CCN keeps convolutional inductive biases and is pre-trained for high-resolution imagery, so it natively handles multi-kilometre scenes without tiling or heavy down-sampling.

- **Hierarchical Spatial Feature Integration (SFI)**
  - Traditional models flatten visual tokens and bolt them onto an LLM through a single linear or Q-Former layer (“shallow alignment”). SFI repeatedly aggregates multi-scale feature maps with learnable queries, preserving spatial structure and fusing details from object to landscape scale before any language interaction.

- **Multi-layer Deep Alignment (MDA) inside the LLM**
  - Instead of passing visual features to the LLM once, Aquila injects SFI outputs at several transformer layers, allowing iterative cross-modal attention and reasoning. This deep fusion yields stronger grounding of spatial relationships and markedly better performance on remote-sensing captioning and VQA tasks.

Aquila architecture



- SFI = Spatial Feature Integration
- MDA = Multi-layer Deep Alignment

Table 2. Hyperparameters for alignment pretraining and instruction finetuning.

Hyperparameters	Alignment pretraining	Instruction finetuning
Visual encoder	CLIP-ConvNeXt-XXL	CLIP-ConvNeXt-XXL
Textual encoder	Llama-3-8b	Llama-3-8b
Datasets	1M	1.8M
Epoch	1	1
GPUs	4 NVIDIA A800	4 NVIDIA A800
Training time	70h	120h
Input resolution	1024^2	1024^2
Image token length	1024	1024
Global batch size	32	32
Learning rate	1e-3	4e-5
Learning rate scheduler	Cosine	Cosine
Optimizer	AdamW	AdamW
Adam β	(0.9, 0.95)	(0.9, 0.95)
Weight decay	0.05	0.1
Warmup ratio	0.06	0.03
Lora rank	--	64
Lora alpha	--	16
Lora drop	--	0.05

Table 3. Performance of VLMs on remote sensing caption datasets

Model	RSICD	Sydney	UCM	FIT_RSFG-Captions
LLaVA1.5-7B	55.26	52.21	57.58	15.38
GeoChat	45.86	39.79	44.09	8.79
RSGPT	70.32	82.26	86.12	--
SkySenseGPT	--	--	--	27.31
Aquila	**74.6**	**83.42**	**88.25**	**35.08**

Table 4. Performance of VLMs on remote sensing visual question answer datasets

Model	RSVQA-LR	RSVQA-HR	FIT_RSFG-VQA
LLaVA1.5-7B	65.86	53.99	58.59
GeoChat	91.81	70.82	53.47
RSGPT	92.29	92.15	--
LHRS-Bot	89.19	92.55	--
H^2RSVLM	89.12	74.35	--
SkyEyeGPT	88.23	86.87	--
SkySenseGPT	92.69	82.56	79.76
Aquila	**92.72**	**92.64**	**83.87**

Visual input examples		
User 4	How many planes do this image contain?	Where is the power tower located in the image? A fully visible large lattice tower is on the bottom right part of the image.
User	Describe the scene. The scene depicts an airport, with a grassy area and a runway in the background. Multiple airplanes can be seen in the image. One airplane is located in the lower right corner, partially visible, with another airplane parked beside it. Facing to the right, a third airplane is parked next to the second airplane. Additionally, a fourth airplane is situated in the upper right corner of the image. These airplanes are neatly lined up in a row, each parked alongside another airplane. The airplanes are arranged neatly, creating a sense of order in the scene.	Describe the scene. In the forest scene, there is a lattice tower standing tall among the trees. The lattice tower is located towards the right side of the image, slightly closer to the foreground. It is positioned near the edge of the forest, overlooking the trees and the power line that runs through the area. The lattice tower is fully visible in the image, with all its intricate details clearly visible. The power line extends from the lattice tower, disappearing into the distance. The tower stands out against the backdrop of the dense forest, adding a man-made element to the natural landscape.
User The airplanes are white in color.	What colors are on this airplane?	What is the direction of the power lines in the image? The power lines appear to extend from the bottom left to the top right.

VQA demonstrations of the proposal Aquila

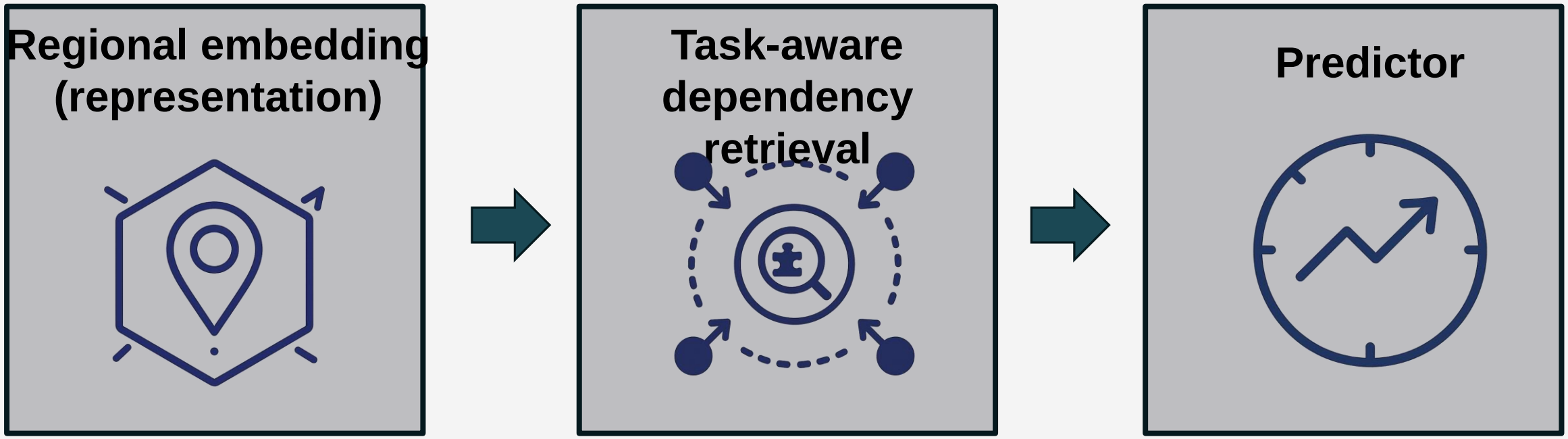
Questions 1

How can multimodal foundation models be adapted to improve predictive analytics in complex urban environments?

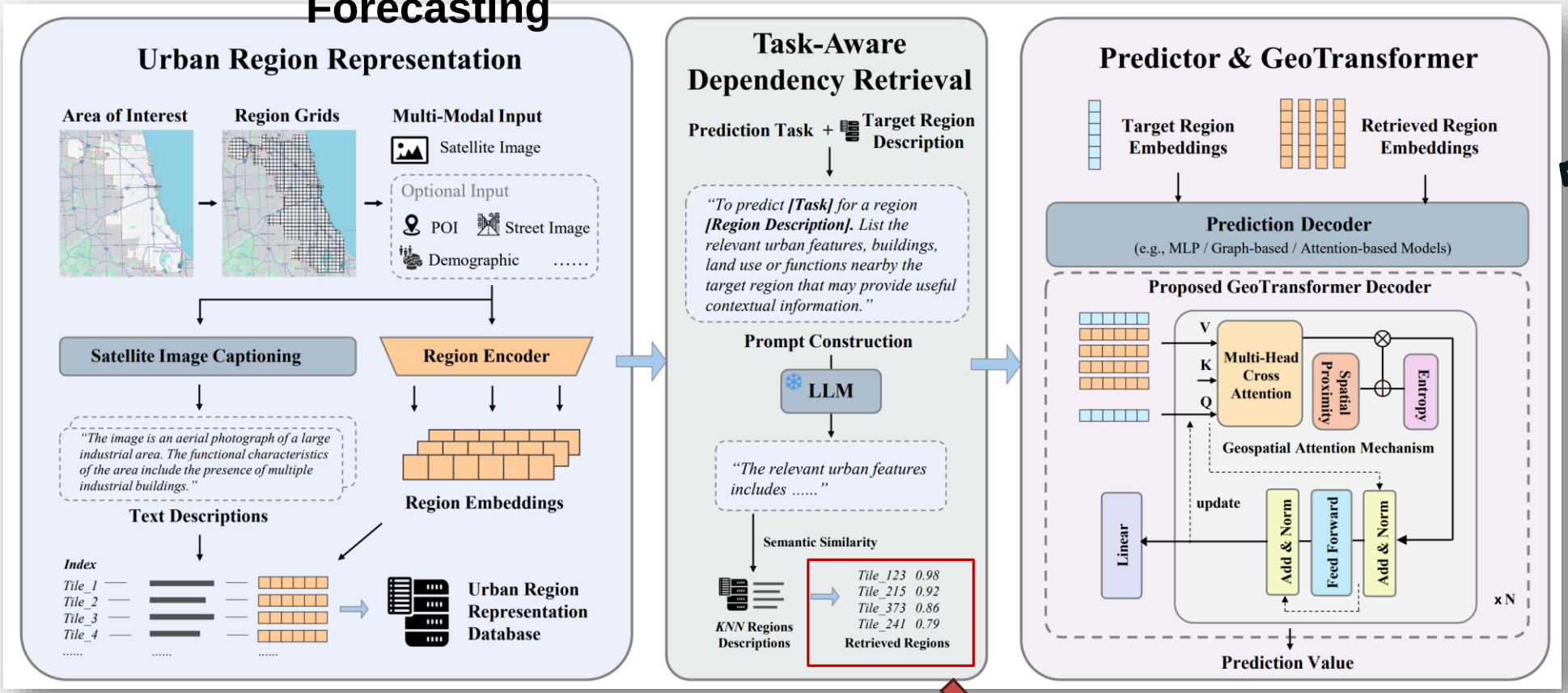
Questions 2

Can we develop a unified forecasting framework that fuses very-high-resolution remote-sensing imagery with plug-and-play auxiliary datasets for predicting urban dynamics?

A Unified Framework for Next-Gen Urban Forecasting



A Unified Framework for Next-Gen Urban Forecasting



Task-aware representation

To predict crimes for a given target region described as follows: *[Region Description]*. List the relevant urban features, buildings, land use, or functions near the target region that may provide useful contextual information.

Variable	Primary data set & provider	Native spatial / temporal resolution	How it was prepared for the model	Notes
(Input) **Satellite imagery**	National Agriculture Imagery Program (NAIP) four-band (RGB + NIR) mosaic for the Greater Chicago Area, downloaded through Google Earth Engine	0.6 m pixel size; acquisition: September 2019	Tiled into 512 × 512-pixel chips (≈ 307 m × 307 m on the ground) before being fed to the representation encoder	Provides the visual backbone from which every region-level embedding and caption is derived
(Prediction tasks) **GDP (grid-level economic output)**	"Global 1 km × 1 km gridded revised real GDP" data set (1992–2019) constructed from DMSP/OLS & VIIRS night-time lights, Chen et al. 2022	1 km grid; annual snapshots; 2019 layer selected	Values interpolated from the 1 km raster onto the 500 m × 500 m analysis grid	Captures broad economic intensity independent of local survey data
Housing price	Official "House Prices" tables on the City of Chicago Data Portal (2019 transactions)	Point records of individual sales; timestamps to the day	Median sale price calculated for each 500 m grid cell via spatial join	Reflects real-estate market value at neighbourhood scale
Ride-share demand	Transportation Network Provider (TNP) & taxi trip logs published by the City of Chicago (2019)	Point pick-up / drop-off events with exact times	Trip counts aggregated to the 500 m grid (separately for pick-ups and drop-offs)	Serves as a proxy for short-distance human mobility
Traffic crashes	Chicago "Traffic Crashes – Crashes" record set (2019 subset)	Point incidents with precise lat/long and timestamp	Total crash count per 500 m cell	Captures safety-related road conditions
Crimes	Chicago "Crimes – 2001-Present" database (records for 2019)	Point incidents; offence codes & times	Incident count per 500 m grid cell	Indicator of social disorder & policing demand
Municipal service demand	Chicago 311 Service Request collections (2019)	Point requests with service category	Request count per 500 m grid cell	Represents resident-reported infrastructure & maintenance needs

- **A city-wide 500 m × 500 m fishnet was generated; all non-imagery variables were spatially joined to this grid and, where necessary, interpolated so that every raster/tile and every structured attribute**

Best regional encoding method among the three

Model	R^2					
	GDP	Housing Price	Ride-share	Traffic Crashes	Crimes	Services
Panel 1 Region-based Encoding Methods						
1 Tile2Vec	0.484/0.320	0.504/0.341	0.551/0.498	0.427/0.318	0.427/0.284	0.675/0.645
2 SatMAE^{++}	0.616/0.403	0.757/0.558	0.719/0.550	0.689/0.425	0.608/0.473	0.836/0.769
3 DHM	0.721/0.493	0.923/0.326	0.668/0.571	0.813/0.281	0.712/0.121	0.856/0.212
Panel 2 Traditional Dependency Modeling + Graph-based Methods						
4 SatMAE^{++}+ GAT (grid)	0.700/0.435	0.882/0.570	0.817/0.501	0.539/0.465	0.594/0.521	0.787/0.745
5 SatMAE^{++}+ GAT (sparse)	0.754/0.561	0.815/0.431	0.781/0.694	0.672/0.325	0.419/0.211	0.891/0.726
Panel 3 Our Framework						
6 SatMAE^{++}+ GAT	0.801/0.612	0.787/0.641	0.825/0.771	0.773/0.453	0.619/0.520	0.832/0.797
7 SatMAE^{++}+ GeoTransformer	**0.811/0.783**	**0.923/0.912**	**0.920/0.901**	**0.716/0.638**	**0.669/0.597**	**0.891/0.824**

Table 2: Predictive performance of baselines and our framework. Each entry is represented as training/testing performance.

- GAT = Graph Attention Network
- DHM = Deep Hybrid Model

Model Variant	R^2		
	GDP	Ride-share	Crimes
Encoder Compatibility			
Tile2Vec + GeoTransformer	0.65/0.55	0.72/0.69	0.60/0.53
SatMAE^{++}+ GeoTransformer	0.81/0.78	0.92/0.90	0.67/0.59
DHM + GeoTransformer	0.87/0.83	0.91/0.89	0.85/0.34
Decoder Compatibility			
SatMAE^{++}+ GAT	0.80/0.61	0.83/0.77	0.62/0.52
SatMAE^{++}+ GeoTransformer	0.81/0.78	0.92/0.90	0.67/0.60

Table 3: Framework compatibility experiments on three representative tasks. Each entry shows training/testing performance.

- GeoTransformer's architecture is genuinely ***plug-and-play***—you can swap in any tested encoder or decoder module and still obtain strong, comparable accuracy across tasks.
- Using an **LLM-guided, task-aware neighbour-retrieval strategy** consistently delivers the highest prediction accuracy, decisively outperforming purely data-driven or random retrieval baselines.
- The model's top performance depends on applying **both spatial-distance and information-entropy weights** in the attention layer; omitting either weight causes a clear drop in predictive power.

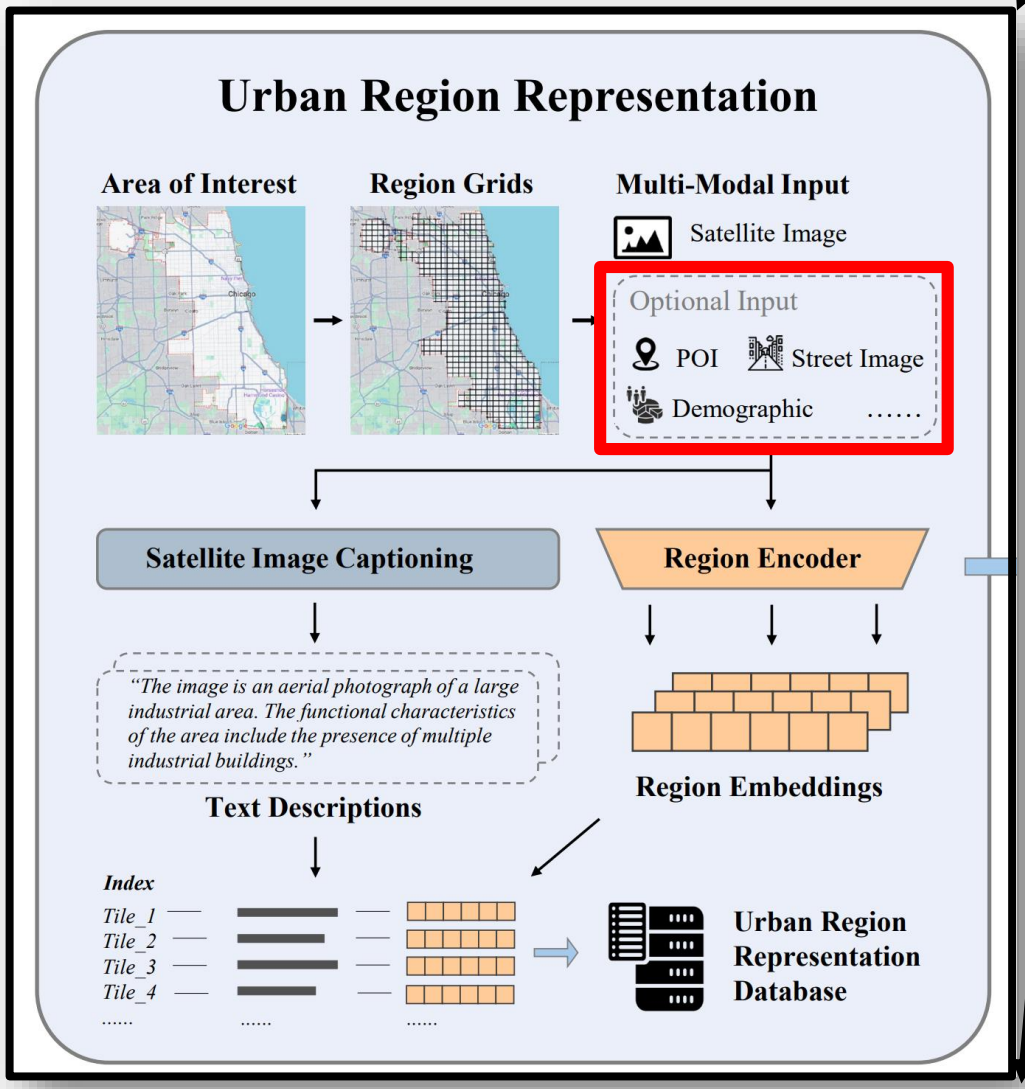
Table 4: Ablation results of retrieval mechanisms (R^2). All models use DHM encoder and GeoTransformer decoder.

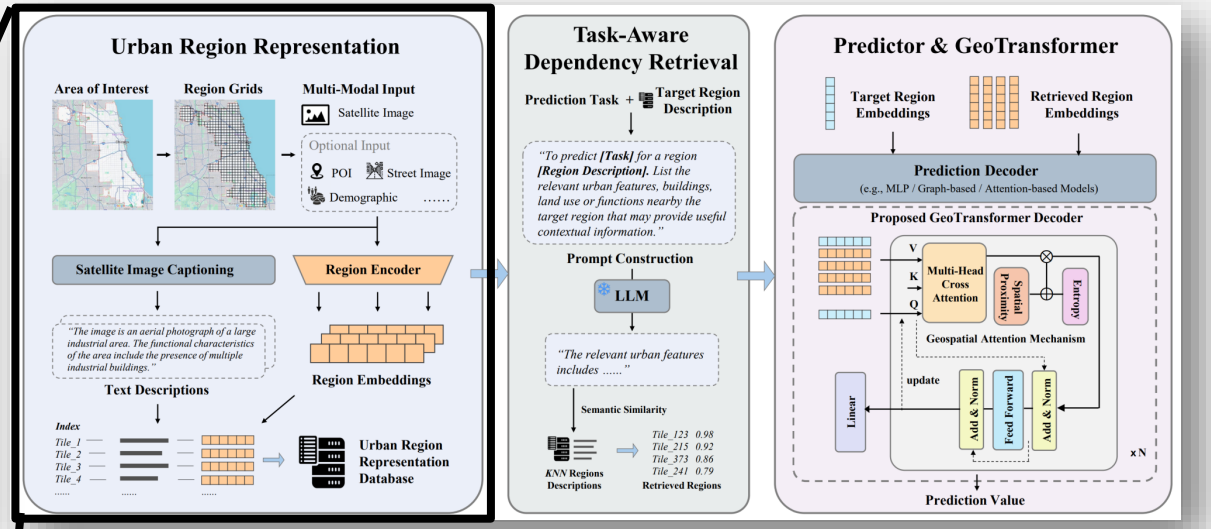
Retrieval Method	GDP	Ride-share	Crimes
Random Retrieval	0.66/0.63	0.72/0.70	0.60/0.45
Similarity-based Retrieval	0.75/0.69	0.88/0.79	0.63/0.51
Sparse Retrieval	0.79/0.72	0.84/0.81	0.70/0.46
Task-aware Retrieval (Ours)	0.81/0.78	0.92/0.90	0.67/0.56

Table 5: Ablation results of geospatial attention weighting (R^2). All models use DHM encoder and task-aware retrieval.

Weighting Variant	GDP	Ride-share	Crimes
No Spatial Weight (W_S off)	0.77/0.75	0.90/0.84	0.60/0.59
No Entropy Weight (W_E off)	0.80/0.75	0.88/0.87	0.63/0.56
No Weighting	0.70/0.66	0.82/0.73	0.59/0.52
Full GeoTransformer (Ours)	0.81/0.78	0.92/0.90	0.67/0.60

Unified Framework for Next-Gen Urban Forecasting





- Satellite imagery alone already provides a surprisingly strong signal, yielding high-quality predictions across a wide range of urban-scale tasks.
- Enriching the model with complementary context, e.g., street-view scenes, POI inventories, and demographic profiles, is expected to boost accuracy and robustness.
- The auto-generated text captions offer human-readable explanations of each region, helping stakeholders interpret and validate model outputs.

A Unified Framework for Next-Gen Urban Forecasting via LLM-driven Dependency Retrieval and GeoTransformer

Yuhao Jia
yuhao.jia@emory.edu
Emory University, University of Pennsylvania
USA

Zile Wu
wuzile@alumni.upenn.edu
University of Pennsylvania
Philadelphia, USA

Shengao Yi
shengao@upenn.edu
University of Pennsylvania
Philadelphia, USA

Yifei Sun
sophiasun@alumni.upenn.edu
University of Pennsylvania
Philadelphia, USA

Xiao Huang
xiao.huang2@emory.edu
Emory University
Atlanta, USA

arXiv:2408.08852v4 [cs.AI] 17 Jun 2025

Abstract

Urban forecasting has increasingly benefited from high-dimensional spatial data through two primary approaches: graph-based methods which rely on predefined spatial structures, and region-based methods that focus on learning expressive urban representations. Although these methods have laid a strong foundation, they either rely heavily on structured spatial data, struggle to adapt to task-specific dependencies, or fail to integrate holistic urban context. Moreover, no existing framework systematically integrates these two paradigms and overcome their respective limitations. To address this gap, we propose a novel, unified framework for high-dimensional urban forecasting, composed of three key components: (1) the Urban Region Representation Module that organizes latent embeddings and semantic descriptions for each region, (2) the Task-aware Dependency Retrieval module that selects relevant context regions based on natural language prompts, and (3) the Prediction Module, exemplified by our proposed GeoTransformer architecture, which adopts a novel geospatial attention mechanism to incorporate spatial proximity and information entropy as priors. Our framework is modular and supports diverse representation methods and forecasting models, and can operate even with minimal input. Quantitative experiments and qualitative analysis across six urban forecasting tasks demonstrate strong task generalization and validate the framework's effectiveness.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Information systems applications**; **Information retrieval**.

Keywords

urban representation, transformer, dependency retrieval, geospatial attention

1 Introduction

In urban forecasting tasks, classical methods usually rely statistical and machine learning methods that operate on low-dimensional, hand-engineered features [6, 13, 17, 19, 20, 24]. While effective in constrained settings, these approaches struggle to model the complexity of urban systems.

Recent advances in spatial representation learning, remote sensing, and deep neural architectures have introduced a new paradigm in urban modeling: transforming urban regions into high-dimensional latent representations to better capture complex urban dynamics. Such representations are commonly derived from text embedding [4, 12], spatial representation learning [14, 16, 26] or by encoding satellite imagery data [10, 23].

High-dimensional urban forecasting applications can be broadly categorized into two directions. The first utilizes graph-based modeling with spatial feature embeddings, then using Graph Neural Networks (GNNs) or Graph Attention Networks (GATs) for predictions [4, 7, 12, 14, 30]. While effective, these methods depend heavily on predefined spatial structures and high-quality spatial data, which limits their flexibility in data-sparse or dynamically changing environments. The second direction focuses on region-based methods, which derive high-dimensional representations directly from satellite imagery or other high-resolution spatial data [10, 18, 23]. These methods produce compact representations that preserve built environment features and support downstream tasks. However, these approaches only utilize local information within each patch for prediction and lack the capability to incorporate global urban context [23], which is crucial for tasks requiring holistic understanding.

The limitations and incompatibility of the two paradigms ultimately reflect a structural divergence rooted in whether spatial dependency is available—either built into the input or entirely absent. Several studies have explored automated mechanisms for capturing spatial dependencies for high-dimensional representations, including spatial autocorrelation, proximity, or sparse regression [8, 11, 15]. However, these approaches remain task-agnostic. To date, no unified framework exists that systematically integrates the two modeling paradigms through task-aware dependency modeling to address their respective limitations.

To address these gaps, we propose a novel, unified and modular framework for high-dimensional urban forecasting. It consists of three functional modules: (1) the Urban Region Representation Module encodes each region into high-dimensional embeddings and semantic descriptions; (2) the Task-Aware Dependency Retrieval Module identifies spatial dependencies among regions by matching task-specific prompts with semantic descriptions; and (3)

References

Huang, X., Lu, K., Wang, S., Lu, J., Li, X., & Zhang, R. (2024). Understanding remote sensing imagery like reading a text document: What can remote sensing image captioning offer?. International Journal of Applied Earth Observation and Geoinformation, 131, 103939.

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., ... & Yuan, L. (2024). Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4818-4829).

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., ... & Tang, J. (2024). Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems, 37, 121475-121499.

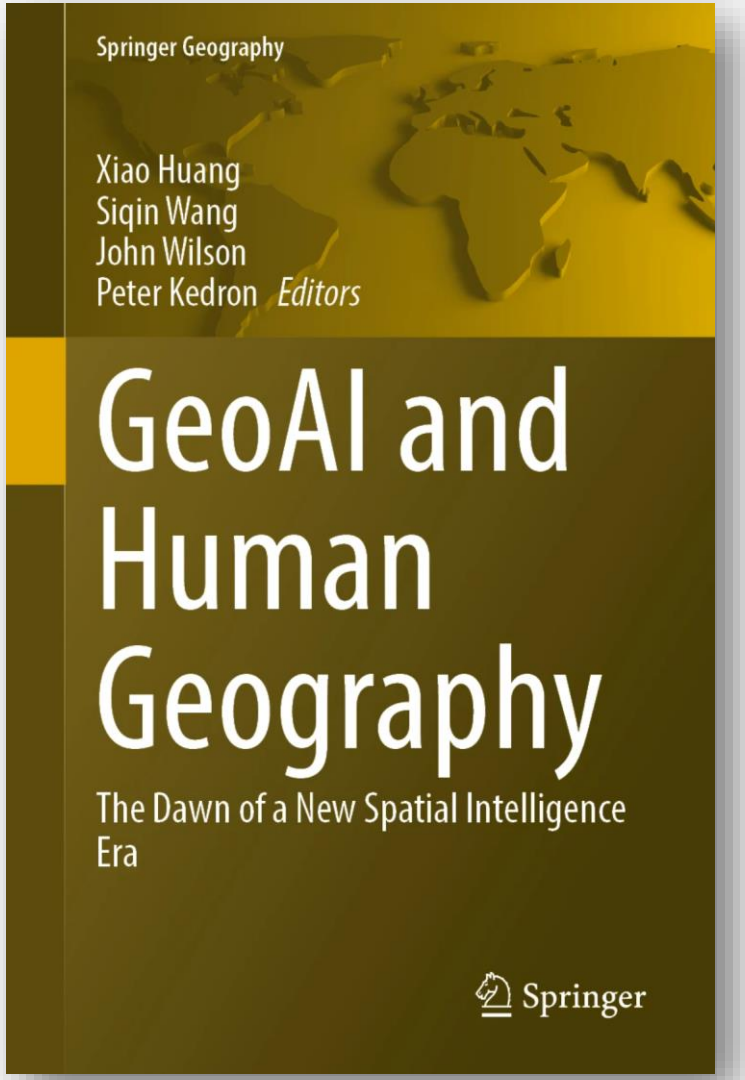
Steiner, A., Pinto, A. S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., ... & Zhai, X. (2024). Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... & Lin, J. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

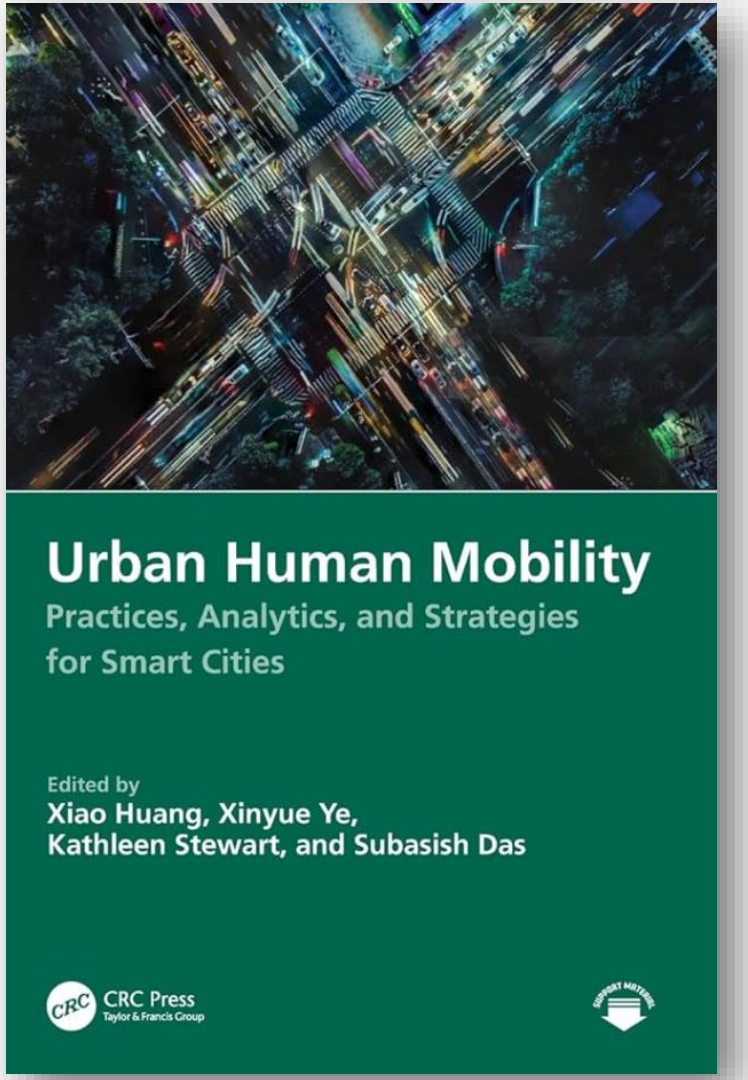
Zhang, Z., Shu, D., Liao, C., Liu, C., Zhao, Y., Wang, R., ... & Gong, J. (2025). FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification using high-resolution multimodal remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing.

Zhang, Z., Shu, D., Liao, C., Liu, C., Zhao, Y., Wang, R., ... & Gong, J. (2025). FlexiSAM: A flexible SAM-based semantic segmentation model for land cover classification using high-resolution multimodal remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing.

Zhang, Z., Zhang, M., Gong, J., Hu, X., Xiong, H., Zhou, H., & Cao, Z. (2023). LuoJiaAI: A cloud-based artificial intelligence platform for remote sensing image interpretation. Geo-Spatial Information Science, 26(2), 218-241.



On sale, Springer Nature



On sale, CRC Press

Data-Driven Earth Observation for Disaster Management: From Theory to Practical Applications

ELSEVIER

April 2026

In production, ELSEVIER

EMORY UNIVERSITY

THANK YOU / QUESTIONS ?

Xiao Huang (xiao.huang2@emory.edu)
Department of Environmental Sciences
Emory University
Personal website: www.xiaohuang116.com

ISDE Seminar Series
July 10, 2025