LI KAI

Age: 27 · Homepage: https://likaiucas.github.io/
WeChat: lkbb1213 · Email: kevinlikai@foxmail.com
Phone: (+86) 132-5820-0037

HIGHLIGHTS

Research focus: **large multimodal models** (MLLM / VLM) · **video world models** · reinforcement learning · 3D spatial understanding. Two research internships at Tencent; currently with the **Tencent Omni Research Team** working on **memory mechanisms for video world models**, alongside reinforcement-learning-driven sparse-view 3D indoor scene reconstruction.
Publications: 2 first-author papers in IEEE Transactions journals (CAS Tier-1 top, CCF-B) and 4 first-author papers under review; 19 papers in total, 152 citations, h-index 7 (Google Scholar).
Competitions: **1st place** in the WWW 2025 (CCF-A) Multimodal Dialogue Intent Recognition Challenge; 2nd place in the Tianzhi Cup 2023 Change Detection Track (RMB 300k prize); 3rd prize in the PIESAT Cup 2022.
Honours: nominee for the IEEE GRSS Chapter Excellence in Technical Communication Award (10 nominees worldwide, the only Chinese nominee); National Second Prize, China Undergraduate Mathematical Contest in Modeling; National Scholarship; full scholarship at City University of Hong Kong.
Deployment: the proposed algorithms have been deployed at surveying and mapping bureaus in several regions of China.

EXPERIENCE

Tencent – Omni Research Team – Research Intern (Multimodal LLM / Video World Models) Oct. 2025 – Present

Topics: memory in video world models; reinforcement learning and multimodal LLMs for sparse-view 3D scene generation and reconstruction

- **Memory mechanisms for video world models (current focus, ongoing)**: video world models predict how a scene evolves but largely lack a persistent representation of what they have already seen. I study how to give a generative video model an explicit **memory** of scene state, so that long-horizon rollouts stay self-consistent, revisited viewpoints agree with what was generated before, and the cost of remembering does not grow unchecked with sequence length.
- **Reinforcement learning for cross-view consistency (ongoing)**: reconstructions from sparse, uncalibrated views are often geometrically and semantically inconsistent across viewpoints. I treat the execution result of an *executable scene program* as a verifiable reward signal and replace purely supervised training with a **reinforcement learning** objective, optimising cross-view consistency directly instead of fitting each view in isolation.
- **Scenix (first author, under review)**: proposed the **Executable Scene Program** paradigm, removing the reliance on camera parameters and dense views. A VLM jointly predicts room layout, object attributes and semantic descriptions from only a few RGB observations, and the prediction is instantiated as an **editable** 3D scene.
- **Data and training**: built a 3D layout dataset with 6-DoF pose annotations, trained LayoutVLM and combined it with two agentic workflows to achieve a fully decoupled 3D indoor reconstruction pipeline. Unlike conventional 2D-detection-plus-depth-estimation pipelines, it reasons about spatial structure explicitly.

Tencent – Computer Vision Algorithm Intern May 2021 – Sep. 2021

Internship between finishing my bachelor's degree and starting my PhD. Surveyed **3D object detection** and **3D reconstruction** algorithms, which laid the groundwork for my later research on 3D spatial understanding.

Aerospace Information Research Institute, Chinese Academy of Sciences – PhD Research Sep. 2021 – Present

Independently established the Offset Token research line (see Publications), completing a full cycle from proposing the concept, to vector-level extraction, to reformulating the task and finally to efficient computation. The resulting algorithms have been deployed at surveying and mapping bureaus in several regions of China.

EDUCATION

City University of Hong Kong – PhD in Data Science – Advisor: Prof. Xiangyu Zhao Sep. 2024 – Jun. 2027

University of Chinese Academy of Sciences – PhD in Signal and Information Processing – Advisor: Prof. Zhongming Zhao (Party Secretary and Deputy Director, Institute of Remote Sensing, CAS) Sep. 2021 – Jun. 2027

The University of Texas at Austin, USA – Summer School in Artificial Intelligence Jul. 2019 – Aug. 2019

Sophia University, Japan – Short-term Visiting Programme Jan. 2018 – Feb. 2018

University of Electronic Science and Technology of China (Project 985) – B.Eng. in Space Information and Digital Technology Sep. 2017 – Jun. 2021

GPA 3.92/4, ranked 2/53 in the major; recommended for direct PhD admission without entrance examination.

PUBLICATIONS

Two first-author papers published in leading IEEE Transactions journals (CCF-B), with four more first-author papers under review. My doctoral research addresses projection displacement correction for elevated objects (buildings) in oblique, off-nadir imagery, organised around the concept I proposed, the **Offset Token**. The work spans proposing the concept, vector-level extraction, reformulating the problem and efficient computation, and the algorithms have been deployed at surveying and mapping bureaus across several regions of China.

Published:

[1] Li, Kai, et al. "Prompt-Driven Building Footprint Extraction in Aerial Images with Offset-Building Model." IEEE Transactions on Geoscience and Remote Sensing (2024).

Core contribution: the first proposal of the Offset Token concept.

(1) We propose and validate the Offset Token to describe how buildings deform under oblique viewing angles, carrying the field from small CNN models to foundation-model research. (2) We identify a general pattern in projection displacement correction: for low-rise buildings the displacement *length* is easier to predict while its *direction* is hard; for high-rise buildings the direction is clearly easier but the length is hard to estimate. Building on this, and inspired by NMS, we propose the D-NMS post-processing algorithm that uses the directions of high-rise buildings to correct those of low-rise ones.

[2] Li, Kai, et al. "PolyFootNet: Extracting Polygonal Building Footprints in Off-Nadir Remote Sensing Images." IEEE Transactions on Geoscience and Remote Sensing (2025).

Core contribution: starting from the Offset Token, the first vector-level extraction of building structure under off-nadir photography.

(1) We achieve, for the first time, automated vector-level building footprint delineation in imagery far from nadir. (2) Following the general correction pattern above, we combine Nadaraya–Watson regression with a **self-offset attention (SOFA)** mechanism to improve displacement prediction. (3) We analyse the common failure cases of footprint prediction in generalisation experiments and exploit the inherent multi-solution nature of buildings to resolve semantic ambiguity in footprint delineation.

Under review:

[3] Li, Kai, et al. "Scenix: Sparse-View 3D Scene Reconstruction via Executable Scene Programs."

Core contribution: work done at Tencent. A new paradigm in which a VLM learns directly from the 3D layout of a room, enabling fully automatic reconstruction from a handful of photographs to a complete 3D scene. Key ingredients: training LayoutVLM plus two agentic workflows.

Conventional 3D reconstruction built on 2D image analysis (detection, depth estimation) cannot genuinely reason about space. We instead construct a 3D layout dataset with 6-DoF annotations and achieve a fully decoupled 3D indoor scene reconstruction pipeline.

[4] Li, Kai, et al. "DragOSM: Extract Building Roofs and Footprints from Aerial Images by Aligning Historical Labels."

Core contribution: starting from the Offset Token, reformulating the mismatch between historical labels and updated imagery as a denoising problem.

We rethink building-related tasks from an OpenStreetMap perspective: we may not need to predict building structure from scratch every time. Open-source maps already cover most regions of the world, and although their label quality is uneven (positional bias is widespread), they provide clean, human-drawn outlines. By correcting the positional bias we obtain building footprint labels. (1) Inspired by denoising models, we interpret the complex systematic positional error as a Gaussian process over a time series. (2) We propose the **Alignment Token** to guide and correct this positional error, together with **DragOSM**, a label-level autoregressive model. (3) Following current thinking on test-time scaling laws for autoregressive LLMs, we design an analogous TTA strategy that strengthens DragOSM's predictions at inference time.

[5] Li, Kai, et al. "ObliCity: A Benchmark and Baseline for Roof-to-Ground Projection Displacement Correction."
Core contribution: combining the Offset Token with mathematical reasoning, advocating offset learning as a task in its own right.
Following the general pattern found in [1], I analysed the length-error component experimentally: for high-rise buildings the predicted displacement is in most cases *shorter* than ground truth, whereas for low-rise buildings it is generally *longer*. Inspired by flow matching in modern generative models, we cast projection displacement correction as a system of ODEs and train the model with conditional normalizing flows to emulate the entire manual annotation process, yielding markedly more accurate length correction.

[6] Li, Kai, et al. "LoDEOT: Low-Dimensional and Efficient Offset Tokens for Building Footprint Extraction from Off-Nadir Imagery."
Core contribution: previous Offset Token designs followed common CV conventions. Starting instead from the mathematics and geometric intuition of 3D off-nadir imaging of buildings, this work redefines a highly efficient Offset Token and naturally incorporates a denoising formulation into training.

Co-authored publications:

- Zhou, Z., **Li, K.**, & Deng, Y. (2026). Remote-Sensing City Layout Extraction with MLLM. arXiv:2608.16484. [first-authored by the master's student I mentor]
- Meng, Y., Deng, L., Xi, Z., Chen, J., Chen, J., Yue, A., Liu, D., **Li, K.,** ... & Sun, X. (2025). IRSAMap: Towards Large-Scale, High-Resolution Land Cover Map Vectorization. IEEE Transactions on Geoscience and Remote Sensing. [CAS Tier-1 top, CCF-B, IF 8.6]
- Wang, C., Xi, Z., Liu, D., Feng, Y., Deng, Y., **Li, K.**, Chen, J., Chen, J., & Meng, Y. (2025). PCP: A Prompt-Based Cartographic-Level Polygonal Vector Extraction Framework for Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing. [CAS Tier-1 top, CCF-B, IF 8.6]
- Li, Z., Wu, B., Zhang, Y., Li, X., **Li, K.**, & Chen, W. (2025, May). CuSMer: Multimodal Intent Recognition in Customer Service via Data Augment and LLM Merge. In Companion Proceedings of the ACM Web Conference 2025 (pp. 3058–3062). [CCF-A; 1st place in the challenge final]
- Wang, C., Chen, J., Meng, Y., Deng, Y., **Li, K.**, & Kong, Y. (2024). SAMPolyBuild: Adapting the Segment Anything Model for polygonal building extraction. ISPRS Journal of Photogrammetry and Remote Sensing, 218, 707–720. [CAS Tier-1 top, IF 12.2]
- Chen, W., Deng, Y., Jin, W., Chen, J., Chen, J., Feng, Y., Xi, Z., Liu, D., **Li, K.**, & Meng, Y. (2025). DGTRSD and DGTRSCLIP: A Dual-Granularity Remote Sensing Image-Text Dataset and Vision-Language Foundation Model for Alignment. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
- Wang, H., Li, Z., Zhang, Y., He, Z., Jiang, L., **Li, K.**, et al. (2026). ConsistNav: Closing the Action Consistency Gap in Zero-Shot Object Navigation with Semantic Executive Control. arXiv:2605.09869.
- Ma, J., Wang, S., Yang, J., Hu, J., Liang, J., Lin, G., **Li, K.**, & Meng, Y. (2025). SayAnything: Audio-Driven Lip Synchronization with Conditional Video Diffusion. arXiv:2502.11515.
- Chen, W., Chen, J., Deng, Y., Chen, J., Feng, Y., Xi, Z., Liu, D., **Li, K.**, & Meng, Y. (2025). LRSCLIP: A Vision-Language Foundation Model for Aligning Remote Sensing Image with Longer Text. arXiv:2503.19311.
- Three further IGARSS 2019 conference papers as an undergraduate (Land Price Assessment Based on Deep Neural Network; Urban Functional Regions Discovering Based on Deep Learning; A WordNet-Based Geospatial Web Services Search Method Supporting Quality of Service Constraints).

Competitions and Honours

During PhD studies

Academic competitions:
[1] WWW 2025 (CCF-A), Multimodal Dialogue Intent Recognition Challenge — 1st place in the final. (1) I found that, just as for conventional computer vision models, multi-scale and rotation augmentation during training markedly improves the predictive power of multimodal LLMs. (2) Combined with my teammates' model-merging strategy, our model took first place in the final multimodal e-commerce task.
[2] Tianzhi Cup 2023, Change Detection Track (RMB 300k prize) — 2nd place in the final. A heavily engineering-oriented contest; we used standard computer vision augmentation together with a cosine annealing schedule. The main difficulty was the on-site 48-hour limit: we first had to configure a workstation offline, from operating system to runtime environment, before tackling the algorithmic task.
[3] PIESAT Cup 2022, Road Intersection and Road Extraction Track (RMB 30k prize) — 3rd prize in the final. We observed that road annotation is inherently subjective — annotators define roads of different widths differently, so some narrow roads are labelled while others are not. We proposed an image compression strategy to blur the distinction between road widths and identified the optimal downsampling scheme experimentally. Thanks to the compression we ranked 1st in the performance track and 4th overall in the final.
[4] IGARSS 2025 Three Minute Thesis (3MT) — Top 10 in the final (with OBM), and recipient of an IGARSS 2025 Travel Grant.

Honours and scholarships: nominee for the IEEE GRSS Chapter Excellence in Technical Communication Award (10 nominees, the only Chinese nominee); "Three-Good" Outstanding Student, 2025; corporate-sponsored scholarship (2 recipients); full scholarship at City University of Hong Kong.

During undergraduate studies

Academic record: GPA 3.92/4, ranked 2/53 in the major; named a "UESTC Elite for Overseas Study" (10 students campus-wide per year), with funding from the university and New Oriental.
Academic competitions: National Second Prize, China Undergraduate Mathematical Contest in Modeling; 3rd prize or above in 12 campus academic competitions.
Scholarships: National Scholarship ($\times$1); Outstanding Student Scholarship ($\times$3).

Mentorship and Teaching

Mentorship (co-mentored with Prof. Xiangyu Zhao, City University of Hong Kong)

- Jiajun Zhang, M.Sc. at City University of Hong Kong; next position: PhD student at Peking University.
- Zigan Zhou, M.Sc. at City University of Hong Kong; interned at the Chinese Academy of Sciences. Mentored his first-author paper *Remote-Sensing City Layout Extraction with MLLM* (arXiv:2608.16484).

Mentorship (co-mentored with Prof. Jingbo Chen, University of Chinese Academy of Sciences)

- Weizhi Chen, University of Chinese Academy of Sciences (AIRCAS); research focus on CLIP-based remote sensing image-text alignment, with 2 papers published in **IEEE JSTARS**; now an LLM Algorithm Researcher at Zhonggong Education.

Teaching assistantships

Optimization Methods, University of Chinese Academy of Sciences (Prof. Zhenjiang Zhao); Academic English (Prof. James).

Additional Information

English: IELTS 6.5 (Speaking 6.5); CET-6 547.