KAI CHEN

HKUST, Clear Water Bay, New Territories, Hong Kong SAR
Email: kai.chen@connect.ust.hk ◇ Homepage: www.cse.ust.hk/kchenbf

RESEARCH OVERVIEW

My research aims at constructing reliable **Multi-modal** AI systems from a **data-centric** perspective. Recent deep learning has witnessed the success of the **"pre-training fine-tuning"** pipeline, enabled by training on massive amounts of datasets. Although remarkable, the intrinsic nature of supervised learning still poses the AI systems with severe risks, especially when encountering unseen **"corner cases"** during inference time. Thus, an effective and continual **"corner case collection and generation"** process is essential to obtain ultimate reliability of AI systems. This methodology has been explored in unimodal domains (*e.g.*, LLM only), which instead remains an open challenge for multi-modal models. Now, I'm trying to answer the following questions,

- *How to construct end-to-end Multi-modal LLMs with frontier visual, textual, and speech capabilities?*
- *How to construct 3D visual world models in a controllable and scalable manner?*
- *How to enhance Multi-modal LLMs via training with synthetic data and world models?*

Research Areas: Omni-modal LLMs, Visual World Modeling, Mixture-of-Experts (MoE)

EDUCATION

Hong Kong University of Science and Technology, Hong Kong SAR *Sep 2020 - June 2026 (Expect)*
Ph.D. in **Computer Science and Engineering**
GPA: 4.10/4.0
Advisor: Prof. Dit-Yan Yeung

Fudan University(FDU), Shanghai, China *Sep 2016 - June 2020*
B.S. in **Computer Science**, Minor in **Economics** (Outstanding Graduates of Shanghai)
Overall GPA: 3.70/4.0, Major GPA: 3.90/4.0, Ranking: 3/32
Advisor: Prof. Yanwei Fu

University of Manchester, Manchester, UK *Sep 2018 - Jan 2019*
Exchange student in the **Department of Computer Science**
Advisor: Dr. Tingting Mu

EXPERIENCE

Mobile Intelligence Group (MIG), SenseTime Oct 2019 - April 2020
Research Intern *Advisor:Dr. Wenxiu Sun, Sensetime*

- Research on real-time (portrait) instance segmentation deployable on mobile devices.

Computer Vision Lab, Indiana University Bloomington (IUB) June 2019 - Sep 2019
Global Talent Attraction Program (GTAP) Visiting Scholar *Advisor:Prof. David Crandall, IUB*

- Research on semi-supervised semantic segmentation and indoor 3D reconstruction.

SELECTED HONORS

CVPR 2025 Travel Awards	May 2025
HKUST Research Travel Grant	2023-2025
HKUST Postgraduate Scholarship	Sep 2020
Outstanding Graduates of Shanghai [Wechat Post] (5%, by Shanghai Government)	April 2020
Scholarship for Outstanding Graduates (5%, by Fudan University)	April 2020
Oversea Visiting Student Stipend of (15,000 CNY, Fudan University)	Dec 2019

Joel & Ruth Spira Scholarship (1%, by Lutron Electronics) Mar 2019
National Scholarship (1%, by Ministry of Education of P.R.China) Sep 2018
Scholarship for Outstanding Undergraduate Students (5%, by Fudan University) Oct 2017

PUBLICATIONS

Full publication list on my Google Scholar. (* denotes equal contribution)

I. Multi-modal Foundation Models
RQ: How to construct multi-modal LLMs with visual, textual, and speech reasoning abilities simultaneously?

[C23] **EMOVA: Empowering Language Models to See, Hear and Speak with Vivid Emotions** **CVPR 2025**
Kai Chen*, Yunhao Gou*, Runhui Huang*, Zhili Liu*, Daxin Tan*, *and other 26 authors* [link]

[C22] **Perceptual Decoupling for Scalable Multi-modal Reasoning via Reward-Optimized Captioning** **Arxiv 2025**
Yunhao Gou*, **Kai Chen***, Zhili Liu*, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, Yu Zhang. [link]

II. Multi-modal Foundation Models - Mixture of Cluster-conditional Experts (MoCE)
RQ: Does more data always result in better performance during model pre-training and fine-tuning?

[C21] **Mixture of Cluster-conditional LoRA Experts for Vision-language Instruction Tuning** **Arxiv 2023**
Yunhao Gou*, Zhili Liu*, **Kai Chen***, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James Kwok, Yu Zhang. [link]

[C20] **Task-customized Masked Autoencoder via Mixture of Cluster-conditional Experts** **ICLR 2023 Spotlight**
Zhili Liu*, **Kai Chen***, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, James Kwok. [link]

[C19] **Task-Customized Self-Supervised Pre-training with Scalable Dynamic Routing** **AAAI 2022**
Zhili Liu, Jianhua Han, **Kai Chen**, Lanqing Hong, Hang Xu, Chunjing Xu, Zhenguo Li. [link]

III. Multi-modal Foundation Models - (M)LLM Reliability via Self-alignment
RQ: How reliable are frontier (M)LLMs? Can we enhance (M)LLM reliability with its intrinsic capabilities?

[C18] **Corrupted but Not Broken: Rethinking the Impact of Corrupted Data in Visual Instruction Tuning** **Arxiv 2025**
Yunhao Gou, Hansi Yang, Zhili Liu, **Kai Chen**, Yihan Zeng, Lanqing Hong, Zhenguo Li, Qun Liu, James T Kwok, Yu Zhang. [link]

[J2] **Unified Triplet-Level Hallucination Evaluation for Large Vision-Language Models** **TMLR 2025**
Junjie Wu*, Tsz Ting Chung*, **Kai Chen***, Dit-Yan Yeung. [link]

[C17] **Mixture of insighTful Experts (MoTE): The Synergy of Thought Chains and Expert Mixtures in Self-Alignment** **ACL 2025**
Zhili Liu*, Yunhao Gou*, **Kai Chen***, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, James T. Kwok. [link]

[C16] **Eyes Closed, Safety On: Protecting Multimodal LLMs via Image-to-Text Transformation** **ECCV 2024**
Yunhao Gou*, **Kai Chen***, Zhili Liu*, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James Kwok, Yu Zhang. [link]

[C15] **Gaining Wisdom from Setbacks: Aligning Large Language Models via Mistake Analysis** **ICLR 2024**
Kai Chen*, Chunwei Wang*, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, Dit-Yan Yeung, Lifeng Shang, Xin Jiang, Qun Liu. [link]

IV. Visual World Models - Corner Cases for Autonomous Driving (CODA)
RQ: How to enhance the robustness of self-driving agents towards road corner cases?
A: 1) multi-modal reasoning, 2) corner case generation, and 3) corner case collection

[C14] **ECCV 2024 W-CODA: 1st Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving** — **ECCV 2024**
Kai Chen*, Ruiyuan Gao*, Lanqing Hong*, Hang Xu, Xu Jia, Holger Caesar, Dengxin Dai, Bingbing Liu, Dzmitry Tsishkou, Songcen Xu, Chunjing Xu, Qiang Xu, Huchuan Lu, Dit-Yan Yeung. [link]

[C13] **Automated Evaluation of Large Vision-Language Models on Self-driving Corner Cases** — **WACV 2025**
Kai Chen*, Yanze Li*, Wenhua Zhang*, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, Dit-Yan Yeung, Huchuan Lu, Xu Jia. [link]

[C12] **CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving** — **ECCV 2022**
Kaican Li*, **Kai Chen***, Haoyu Wang*, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, Hang Xu. [link]

V. Visual World Models - Geometric-controllable Visual Generation (GeoDiffusion)
RQ: How to generate the 3D visual world in a controllable and scalable manner?

[C11] **MagicDrive3D: Controllable 3D Generation for Any-View Rendering in Street Scenes** — **Arxiv 2024**
Ruiyuan Gao, **Kai Chen**, Zhihao Li, Lanqing Hong, Zhenguo Li, Qiang Xu. [link]

[C10] **MagicDrive-V2: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control** — **ICCV 2025**
Ruiyuan Gao, **Kai Chen**, Bo Xiao, Lanqing Hong, Zhenguo Li, Qiang Xu. [link]

[C9] **Implicit Concept Removal of Diffusion Models** — **ECCV 2024**
Zhili Liu*, **Kai Chen***, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James Kwok. [link]

[C8] **DetDiffusion: Synergizing Generative and Perceptive Models for Enhanced Data Generation and Perception** — **CVPR 2024**
Yibo Wang*, Ruiyuan Gao*, **Kai Chen***, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, Kai Zhang. [link]

[C7] **MagicDrive: Street View Generation with Diverse 3D Geometry Control** — **ICLR 2024**
Ruiyuan Gao*, **Kai Chen***, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, Qiang Xu. [link]

[C6] **TrackDiffusion: Tracklet-Conditioned Video Generation via Diffusion Models** — **WACV 2025**
Pengxiang Li*, **Kai Chen***, Zhili Liu*, Ruiyuan Gao, Lanqing Hong, Dit-Yan Yeung, Huchuan Lu, Xu Jia. [link]

[C5] **GeoDiffusion: Text-Prompted Geometric Control for Object Detection Data Generation** — **ICLR 2024**
Kai Chen*, Enze Xie*, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung. [link]

VI. Representation Learning - Object-level Self-supervised Learning (SSL)
RQ: How to perform object-level SSL for better transferability on downstream dense perception tasks?

[C4] **Mixed Autoencoder for Self-supervised Visual Representation Learning** — **CVPR 2023**
Kai Chen*, Zhili Liu*, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung. [link]

[C3] **MultiSiam: Self-supervised Multi-instance Siamese Representation Learning for Autonomous Driving** — **ICCV 2021**
Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung. [link]

[C2] **SODA10M: A Large-Scale 2D Self/Semi-Supervised Object Detection Dataset for Autonomous Driving** **NeurIPS 2021**
Jianhua Han, Xiwen Liang, Hang Xu, **Kai Chen**, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, Chunjing Xu. [link]

Early Works

[J1] **Automatic Dense Annotation for Monocular 3D Scene Understanding** **IEEE Access 2020**
Md. Alimoor Reza, **Kai Chen**, Akshay Naik, David Crandall, Soon-Heung Jung. [link]

[C1] **Automatic Annotation for Semantic Segmentation in Indoor Scenes** **IROS 2019**
Md Alimoor Reza, Akshay Naik, **Kai Chen**, David Crandall. [link]

ACADEMIC SERVICES

Program Committee / Organizer

- The 1st W-CODA Workshop at ECCV 2024 on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving. 2024
- The 2nd SSLAD workshop at ECCV 2022. 2022
- The 1st SSLAD workshop at ICCV 2021 on Self-supervised Learning for Next-generation Industry-level Autonomous Driving. 2021

Area Chair

- International Joint Conferences on Artificial Intelligence (IJCAI) 2025

Conference Reviewer

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2022-2025
- IEEE International Conference on Computer Vision (ICCV) 2023-2025
- European Conference on Computer Vision (ECCV) 2022-2024
- International Conference on Learning Representations (ICLR) 2023-2025
- International Conference on Machine Learning (ICML) 2025
- Neural Information Processing Systems (NeurIPS) 2021-2025
- International Joint Conferences on Artificial Intelligence (IJCAI) 2023-2025
- AAAI Conference on Artificial Intelligence (AAAI) 2022
- International Conference on Robotics and Automation (ICRA) 2022
- ACM International Conference on Multimedia (ACM MM) 2025
- IEEE Winter Conference on Applications of Computer Vision (WACV) 2026
- Asian Conference on Computer Vision (ACCV) 2024

Journal Reviewer

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)
- IEEE Transactions on Image Processing (TIP)
- IEEE Access

PATENTS

- [CN116665219A] **GeoDiffusion: Text-Prompted Geometric Control for Object Detection Data Generation**. Enze Xie, **Kai Chen**, Lanqing Hong, Zhenguo Li. *Published in May 26th, 2023.*
- [CN115731530A] **MultiSiam: Self-supervised Multi-instance Siamese Representation Learning for Autonomous Driving**. **Kai Chen**, Lanqing Hong, Hang Xu, Zhenguo Li. *Published in Aug. 24th, 2021.*

TEACHING

- **HKUST** COMP 2012 - Object-Oriented Programming and Data Structures, Teaching Assistant, Fall 2024.
- **HKUST** COMP 2012 - Object-Oriented Programming and Data Structures, Teaching Assistant, Fall 2021.
- **HKUST** COMP 2012 - Object-Oriented Programming and Data Structures, Teaching Assistant, Spring 2021.

INVITED TALKS

- [AI TIME Online] EMOVA: Empowering Language Models to See, Hear and Speak with Vivid Emotions. [Recording]
- [VALSE Webinar] Geometric-controllable Visual Generation: A Systemetic Solution.[Recording]
- [AIDriver Online] Controllable Corner Case Generation for Autonomous Driving.[Recording]
- [AI TIME Online] Gaining Wisdom from Setbacks: Aligning Large Language Models via Mistake Analysis. [Recording]
- [TechBeat Online] Gaining Wisdom from Setbacks: Aligning Large Language Models via Mistake Analysis. [Recording]
- [VALSE 2023@Wuxi] Mixed Autoencoder for Self-supervised Visual Representation Learning. [Recording]
- [VALSE 2023@Wuxi] CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving. [Recording]

TECHNICAL SKILLS

Program Languages	Python, Matlab, C/C++/C#, SQL, LaTeX
Framework	Pytorch, Tensorflow
Language	Native in Mandarin, Fluent in English and Japanese CET-4(649), CET-6(619), TOEFL-iBT(101)