SIJIN CHEN

The University of Hong Kong | H: +86 187 1792 9716 | csjch3cook@gmail.com
Homepage | GitHub | Google Scholar

About Me

I am a Ph.D. student at **the University of Hong Kong(HKU-MMLab)** (Sep. 2025 - .), where Prof. Xihui Liu is my advisor. Before that, I spent a wonderful year working on generalist embodied policies as a full-time AI researcher with Dr. Tao Kong at **ByteDance Research (Seed-Robotics)**. I received my Master's degree from **Fudan University** (Sep. 2021 - Jun. 2024), where Prof. Tao Chen is my advisor, and I am also fortunate to work closely with Dr. Hongyuan Zhu from A*STAR, and Dr. Gang Yu, Dr. Xin Chen, and Dr. Chi Zhang from Tencent. Before this, I got my Bachelor's degree also from **Fudan University** (Sep. 2017 - Jun. 2021).

My long-term research goal is to develop robust and generalized multi-modality systems that can **perceive**, **understand**, and **interact** with the physical world. Outside my research, I love sports and music.

Research Interests

Embodied AI, Generative AI, Vision and Language, and Large Language Models.

Employment

ByteDance Research (Seed-Robotics) — Beijing, China
AI Researcher — Jul. 2024 - Aug. 2025
Core member of the GR team led by Dr. Tao Kong, research on generalist embodied policies. Exploring the potential of multi-modal generative pre/co-training for robot manipulation generalization.

Education

The University of Hong Kong (HKU-MMLab) — Hong Kong, China
Ph.D. in Electrical and Electronic Engineering, advised by Prof. Xihui Liu. — Sep. 2025 - .
Fudan University — Shanghai, China
Master in Artificial Intelligence, advised by Prof. Tao Chen. — Sep. 2021 - Jun. 2024
Fudan University — Shanghai, China
Bachelor in Data Science and Big Data Technology. — Sep. 2017 - Jun. 2021

Selected Publications (google scholar)

*** Please see my homepage for demos of the selected publications.**

- GR-3 Technical Report.
  [**Alphabetical**] Chilam Cheang, **Sijin Chen**, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, Yichu Yang.
  [Tech Report 2025 | project | paper]
  [Summary]: The latest VLA model from **ByteDance Seed-Robotics**. GR-3 is able to 1) zero-shot generalize to novel objects and instructions, 2) efficiently adapt to novel settings with few-shot human trajectories, and 3) perform long-horizon and dexterous manipulation.

- OmniSVG: A Unified Scalable Vector Graphics Generation Model.
  Yiying Yang*, Wei Cheng*, **Sijin Chen**, Xianfang Zeng, Jiaxu Zhang, Liao Wang, Gang Yu, Xinjun Ma, Yu-Gang Jiang.
  [ArXiv 2025 | project | paper | github | 2,000+ stars]
  [Summary]: OmniSVG auto-regressively generates SVGs from images and texts.

- MeshXL: Neural Coordinate Field for Generative 3D Foundation Models.
  Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Yijun Fu, Wei Cheng, Fukun Yin, Yanru Wang, Zhibin Wang, Jingyi Yu, Gang Yu, Bin Fu, Tao Chen.
  [NeurIPS 2024 | project | paper | github | 320+ stars]
  [Summary]: Building end-to-end large auto-regressive 3D mesh generation models.

- MeshAnything: Artist-Created Mesh Generation with Autoregressive Transformers.
  Yiwen Chen, Tong He, Di Huang, Weicai Ye, **Sijin Chen**, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, Guosheng Lin, Chi Zhang.
  [ICLR 2025 | project | paper | github | 2,000+ stars]
  [Summary]: MeshAnything mimics human artists in extracting meshes from any 3D representation.
- LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning.
  Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, Tao Chen.
  [CVPR 2024 | project | paper | github | 300+ stars]
  [Summary]: 3D-LLMs respond to visual and text interactions in complex 3D scenes.
- Vote2Cap-DETR++: Decoupling Localization and Describing for End-to-End 3D Dense Captioning.
  Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, Tao Chen.
  [T-PAMI 2024 | paper | github | 90+ stars]
  [Summary]: Decoupled feature extraction for localizing and describing objects in 3D scenes.
- End-to-End 3D Dense Captioning with Vote2Cap-DETR.
  Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, Tao Chen.
  [CVPR 2023 | paper | github | 90+ stars]
  [Summary]: Addressing 3D dense captioning as a set prediction problem with parallel decoding.
- M3DBench: Let's Instruct Large Models with Multi-modal 3D Prompts.
  Mingsheng Li, Xin Chen, Chi Zhang, **Sijin Chen**, Hongyuan Zhu, Fukun Yin, Gang Yu, Tao Chen.
  [ECCV 2024 | project | paper | github | 60+ stars]
  [Summary]: A large scale dataset querying 3D language models with text, 2D, and 3D prompts.
- 3DET-Mamba: Causal Sequence Modelling for End-to-End 3D Object Detection.
  Mingsheng Li, Jiakang Yuan, **Sijin Chen**, Lin Zhang, Anyu Zhu, Xin Chen, Tao Chen.
  [NeurIPS 2024 | paper]
  [Summary]: Exploring state space model's potential as both encoder and decoder for 3D detection.
- WI3D: Weakly Incremental 3D Detection via Visual Prompts.
  Mingsheng Li, **Sijin Chen**, Shengji Tang, Hongyuan Zhu, Xin Chen, Fukun Yin, Tao Chen.
  [T-MM 2024 | paper]
  [Summary]: Introducing new categories to 3D detectors with 2D foundation models.

Projects

- **Generalist Embodied Policies.** Jul. 2024 - Aug. 2025
  Presented **GR-3**, a Vision-Language-Action (VLA) model on bi-manual robots for general pick-place tasks that follows language instructions. Put forward a co-training strategy that elicits zero-shot instruction following on 1) unseen objects and 2) unseen spatial and high-level concepts. Proposed an incremental few-shot and cross-embodiment fine-tuning strategy that efficiently adapts GR-3 to novel objects with minimal human hand trajectories ($\leq$10 hand trajectories per unseen object).
- **Generative 3D Foundation Models.** Jan. 2024 - Jun. 2024
  Put forward **MeshXL**, a family of generative pre-trained transformers for the direct generation of 3D object meshes, accepted to NeurIPS 2024. Proposed **MeshAnything** to mimic human artists in extracting meshes from any 3D representation, accepted to ICLR 2025.
- **Language for 3D Scenes.** Aug. 2021 - Mar. 2024
  Proposed **Vote2Cap-DETR**, a set-to-set method for localizing and describing objects in 3D scenes, accepted to CVPR 2023 and won the Scan2Cap challenge at ICCV 2023. Proposed an advanced method, **Vote2Cap-DETR++**, which is accepted to T-PAMI 2024. Presented **LL3DA**, a large language 3D assistant responding to both text and visual interactions with complex 3D scenes, accepted to CVPR 2024. Put forward **M3DBench**, a large-scale multi-modal 3D dataset covering 327k lines of annotations for diverse 3D vision and language tasks, accepted to ECCV 2024.
- **Class-Incremental 3D Detection.** Apr. 2023 - Dec. 2023
  Proposed WI3D, learning to detect new categories from 2D images, accepted to T-MM 2024.
- **Earlier R&D Projects.** Before Sep. 2021
  Self-Supervised Pre-training on 3D Point Clouds. Developed a self-supervised learning algorithm that learns global- and patch-level contrastive representations for 3D point clouds.
  A Smart Advertisement Display System. Developed a human perception system that detects faces, recognizes facial expressions, estimates eye gaze, age, and gender for advertisement recommendation.

Scholarships and Awards

Spot Bonus at ByteDance AI-Lab-Research (Breakthrough in new fields, 2025-Q1). Apr. 2025
Outstanding Graduate Student Award (rank 1/24). Apr. 2024
First place winner of the Scan2Cap Challenge at ICCV 2023. Oct. 2023
National Scholarship (rank 1/46). Sep. 2023
Second prize of the Scholarship for Outstanding Students. Sep. 2022
Award for the Scholarship for Outstanding Students. Sep. 2021
Second prize of the Scholarship for Outstanding Students. Jun. 2021

Research Intern

Tencent. Jan. 2024 - Jun. 2024
Research Intern, advised by Dr. Xin Chen and Dr. Gang Yu, working on generative 3D foundation models. Proposed MeshXL, a family of generative pre-trained transformers for the direct generation of 3D object meshes, accepted to NeurIPS 2024.

Invited Talks

- "**MeshXL**: Neural Coordinate Field for Generative 3D Foundation Models" Jul. 2024
  Under a well-defined ordering strategy, the direct generation of 3D meshes can be modeled as a "next-coordinate generation" paradigm, and can be seeminglessly addressed by modern large language model techniques. One technical report at miHoYo and another at AnySyn3D.
- "**Vote2Cap-DETR**: A Set-to-Set Perspective Towards 3D Dense Captioning" Oct. 2023
  By treating 3D Dense Captioning as a translation task from a set of object queries into a set of "box-caption" pairs, we present a set-to-set perspective towards 3D Dense Captioning. A winner presentation for the Scan2Cap challenge at ICCV 2023.
- "End-to-End 3D Dense Captioning with Vote2Cap-DETR" Jun. 2023
  We present an end-to-end transformer model for localizing and describing objects in parallel within diverse 3D environments. A paper presentation at VALSE 2023, Wuxi, China.

Skills

Languages: Chinese (native), English (proficient), Shanghai dialect
Programming: Python, R, C, Matlab, SQL
Tools: PyTorch, Blender, Visual Studio, Spyder, Jupyter Notebook

Academic Services

I am excited about the rapid development of the AI society. Currently, I serve as a reviewer for **NeurIPS** (2024, 2025), **ICCV** 2025, **ACM MM** 2025, **ICLR** 2025, **AAAI** 2025, **ICML** 2025, and **T-MM**.