XIN WEN

The University of Hong Kong | (+852) 6211-6716 | wenxin@eee.hku.hk | Google Scholar | Github | wen-xin.info

EDUCATION

The University of Hong Kong — **Hong Kong SAR**
Ph.D. in Electrical and Electronic Engineering — Sept. 2021 - Mar. 2026 (expected)
- Advisor: Prof. Xiaojuan Qi

Tongji University — **Shanghai, China**
B.Eng. in Computer Science (Minor: Applied Mathematics) — Sept. 2017 - July 2021
- Advisor: Prof. Yin Wang

EXPERIENCE

Meta FAIR — Perception Team — **New York, USA**
Research Scientist Intern — Aug. 2025 - Jan. 2026 (expected)
- Will work with Daniel Bolya and Christoph Feichtenhofer on large-scale multi-modal representation learning.

Huawei Noah's Ark Lab & Imperial College London — **London, UK**
Research Intern — Oct. 2024 - May 2025
- Advised by Ismail Elezi and Jiankang Deng, I developed a visual tokenizer that decomposes images to their *semantic principal components*, enabling a PCA-like visual representation similar to human language [2].

Shanghai AI Laboratory — Embodied AI Center — **Shanghai, China**
Research Intern — Aug. 2023 - Oct. 2024
- Advised by Yilun Chen and Jiangmiao Pang, I systematically studied the effect of *long-tailed/non-object-centric/uncurated pre-training data* for VLMs [4] and robotics [3] and provided corresponding solutions.

MEGVII Research — Foundation Model Group — **Beijing, China**
Research Intern — Apr. 2022 - June 2023
- Advised by Xiangyu Zhang, I designed a self-supervised method that learns *object discovery* and visual representations together from scratch [7]. I also proposed a simple yet strong baseline for generalized category discovery [16].

ByteDance AI Lab — Visual Computing Group — **Shanghai, China**
Research Intern — Jan. 2020 - June 2021
- Advised by Jie Shao, I worked on *content-based video retrieval* [8] and *large-scale video-language pre-training*.
- Training these models required processing million-scale video data. Contributed to TikTok's video de-duplication service.

RESEARCH OUTLINE

I build **core representations** that enable machines to *perceive, model, and act*, forging structure from the chaos of real-world data.

- **Pre-Training on Unstructured Data:** Developing methods to learn from non-object-centric [3, 7], long-tailed [4], and heterogeneous data [5] across diverse modalities including image [3, 7, 9], video [8], vision-language [4], and point cloud [5, 6].
- **Generation and Action:** Building novel visual tokenizers for autoregressive generation from scratch [2] or from pre-trained representations [1], and deploying pre-trained representations for robot interaction with the physical world [3].
- **Open-World Generalization:** Demonstrating robust performance in challenging scenarios including OOD [10, 12], long-tail [11, 14], open-world understanding [15, 16, 17], and counterfactual reasoning [13].

PUBLICATIONS (*equal contribution, †project lead)

Visual Representation Learning and Pre-Training

[1] Vision Foundation Models as Effective Visual Tokenizers for Autoregressive Image Generation — Preprint 2025 [PDF]
Anlin Zheng, **Xin Wen**, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, Xiaojuan Qi

[2] "Principal Components" Enable A New Language of Images — **ICCV 2025** [PDF] [Code]
Xin Wen*, Bingchen Zhao*, Ismail Elezi, Jiankang Deng, Xiaojuan Qi

[3] A Data-Centric Revisit of Pre-Trained Vision Models for Robot Learning — **CVPR 2025** [PDF] [Code]
Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, Xiaojuan Qi

[4] What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights — **NeurIPS 2024** [PDF] [Code]
Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, Xiaojuan Qi

[5] Towards Larger-scale 3D Representation Learning with Multi-dataset Point Prompt Training — **CVPR 2024** [PDF] [Code]
Xiaoyang Wu, Zhuotao Tian, **Xin Wen**, Bohao Peng, Xihui Liu, Kaicheng Yu, Hengshuang Zhao

[6] Masked Scene Contrast: A Scalable Framework for Unsupervised 3D Representation Learning — **CVPR 2023** [PDF] [Code]
Xiaoyang Wu, **Xin Wen**, Xihui Liu, Hengshuang Zhao

[7] Self-Supervised Visual Representation Learning with Semantic Grouping — **NeurIPS 2022** [PDF] [Code]
Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, Xiaojuan Qi

[8] Temporal Context Aggregation for Video Retrieval with Contrastive Learning — **WACV 2021** [PDF] [Code]
Jie Shao*, **Xin Wen***, Bingchen Zhao, Xiangyang Xue

[9] Distilling Visual Priors from Self-Supervised Learning — **ECCV 2020 VIPriors Workshop** [PDF] [Code]
Bingchen Zhao, **Xin Wen**

Generalization and Robustness (open-world, OOD, long-tail, counterfactual, etc.)

[10] Rethinking Out-of-Distribution Detection in Vision Foundation Models — **ICCV 2025**
Shizhen Zhao, Jiahui Liu, **Xin Wen**, Xiaojuan Qi

[11] Learning from Neighbors: Category Extrapolation for Long-Tail Learning — **CVPR 2025**
Shizhen Zhao, **Xin Wen**, Jiahui Liu, Chuofan Ma, Chunfeng Yuan, Xiaojuan Qi [PDF]

[12] Can OOD Object Detectors Learn from Foundation Models? — **ECCV 2024**
Jiahui Liu, **Xin Wen**, Shizhen Zhao, Yingxian Chen, Xiaojuan Qi [PDF] [Code]

[13] What If the TV Was Off? Examining Counterfactual Reasoning Abilities of Multi-modal Language Models — **CVPR 2024**
Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, **Xin Wen**†, Bingchen Zhao† [PDF] [Code]

[14] Classes Are Not Equal: An Empirical Study on Image Recognition Fairness — **CVPR 2024**
Jiequan Cui, Beier Zhu, **Xin Wen**, Xiaojuan Qi, Bei Yu, Hanwang Zhang [PDF]

[15] CoDet: Co-Occurrence Guided Region-Word Alignment for Open-Vocabulary Object Detection — **NeurIPS 2023**
Chuofan Ma, Yi Jiang*, **Xin Wen***, Zehuan Yuan, Xiaojuan Qi [PDF] [Code]

[16] Parametric Classification for Generalized Category Discovery: A Baseline Study — **ICCV 2023**
Xin Wen*, Bingchen Zhao*, Xiaojuan Qi [PDF] [Code]

[17] Learning Semi-supervised Gaussian Mixture Models for Generalized Category Discovery — **ICCV 2023**
Bingchen Zhao, **Xin Wen**, Kai Han [PDF] [Code]

TALKS

"Principal Components" Enable A New Language of Images
- Huawei Noah's Ark Lab, London — May 2025

Generalization, Composition, Objectness, and Shortcuts: A Revisit of CLIP and DINO
- The University of Hong Kong, Hong Kong — Aug. 2024
- Shanghai AI Laboratory, Shanghai — Dec. 2023

Examining Counterfactual Reasoning Abilities of Multi-modal Language Models
- ICCV 2023 VLAR Workshop, Paris — Oct. 2023

Self-Supervised Visual Representation Learning with Semantic Grouping
- CCVL Lab, Johns Hopkins University, Baltimore — Nov. 2022

AWARDS AND HONORS

- ICLR 2025 Notable Reviewer — May 2025
- NeurIPS 2022 Scholar Award — Oct. 2022
- SmartMore PhD Fellowship — 2021 – 2025
- Outstanding Graduates of Shanghai — June 2021
- **2nd** place, ECCV 2020 VIPriors Image Classification Challenge — July 2020
- Qidi Scholarship of Tongji University (**Top 1%**) — June 2020
- **Regional Champion (China)**, Covestro International Data Science Hackathon — Nov. 2019
- **Silver Medal**, 43rd ACM International Collegiate Programming Contest (ICPC) Asia-East Continent Final — Dec. 2018

ACADEMIC SERVICES

Journal Reviewer
- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) — 2023 – Present
- International Journal of Computer Vision (IJCV) — 2024 – Present

Conference Reviewer
- Neural Information Processing Systems (NeurIPS) — 2023, 2024, 2025
- International Conference on Learning Representations (ICLR) — 2024, 2025
- International Conference on Machine Learning (ICML) — 2024, 2025
- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) — 2022, 2023, 2024, 2025
- IEEE/CVF International Conference on Computer Vision (ICCV) — 2023, 2025
- European Conference on Computer Vision (ECCV) — 2022, 2024
- IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) — 2022, 2023, 2024

Workshop Reviewer
- Out-of-Distribution Generalization in Computer Vision (OOD-CV) — 2023, 2024
- AI for Visual Arts (AI4VA) — 2024
- Computer Vision in the Wild (CVinW) — 2023

SKILLS

Languages: C/C++, Python, Chinese, English **Platform and Toolkits:** Bash, Git, Vim, Slurm, PyTorch, TensorFlow, OpenCV