HAIYU WU

haiyupersonal@gmail.com | Google Scholar | (574)-413-7096

SUMMARY

The long-term goal of my research is to improve the transfer learning, zero-shot, and few-shot capabilities of foundation models, and to understand how models can effectively learn useful knowledge from noisy data. With this motivation, I led the team at Altos Labs to build a world model for cells that can capture essence of biology from low-quality datasets. Meanwhile, I have published papers at top-tier conferences which covers VLM, PETF, image generation, etc. Applying my knowledge and kills to the most impactful project is always my passion.

Highlighted Research Experience and Publications

World Model, **Self-supervised Learning**, **VLM**, **Diffusion model**

WORK EXPERIENCE

Full time - Multi-modal Research Scientist **July 2025 - Present**
Manager: Dr. Morgan Levine *Altos Labs*

- *Goal*: Build a world model that capture essence of biology from low-quality omics datasets.
- *Achievement 1*: Led a team to build a self-supervised multi-modal foundation model (non-autoregressive) for noisy biological data. Resolved extreme sequence length disparities (60K RNA to 3B ATAC) and achieved SOTA zero-shot performance. Pioneered robust age prediction at the single-cell resolution.
- *Achievement 2*: Invented a novel self-supervised learning method, HiJEPA, that does not rely on training heuristics. It outperforms the SOTA method, DINO, on **out-of-domain** classification, **transfer learning**, **segmentation**, and **generation** aspects. Moreover, it is efficient to scale up and easy to tune the hyper-parameters.
- *Achievement 3*: Contributed to building a Llava-like LLM for age, disease, and cell type prediction. We trained Qwen2.5-7B and used a finite-state machine to produce formatted JSON outputs. Our team won the **first prize** in an internal competition, outperforming the other teams by over **15%**.

Research Intern - Generic object generation **April 2024 - January 2025**
Advisor: Dr. Liang Zheng *The Australian National University*

- *Goal*: Generate synthetic data that is even more effective than real data.
- *Achievement*: Proposed a generative algorithm that efficiently and effectively generates large-scale face recognition datasets with **synthetic identities**. The generated dataset achieves a higher accuracy than datasets with images from **real identities**.

EDUCATION

Ph.D degree in Computer Engineering **May 2025**
University of Notre Dame
Advisor: Dr. Kevin W. Bowyer

FEATURED RESEARCH

[1] HiJEPA: Hardened improved JEPA via embedding space regularization decoupling (**Coming soon**)

- Keywords: Self-supervised learning, World model, OOD generalizability, transfer learning.
- Achievement: Invented a novel self-supervised learning method that needs **ZERO** training heuristics for training stability, being easy to scale up, and having a better performance on OOD datasets, transfer learning, segmentation, and generation than DINOv1. With the same backbone and pretraining dataset, it achieves **1.5%** higher average accuracy than any existing methods. This enables a whole new research direction in effectively building world models in various areas.

[2] Vec2Face+ for Face Dataset Generation. (**Under review T-PAMI**)

- Keywords: Conditional generation, LoRA, identity preservation and generation.

- Achievement: Developed a novel generative model that produced the first dataset with pure synthetic identities to surpass the real-world CASIA-WebFace dataset in average accuracy across **9** benchmarks. This resolves the GDPR policy violation of the existing FR training sets.

[3] Prompt-OT: An Optimal Transport Regularization Paradigm for Knowledge Preservation in Vision-Language Model Adaptation. (**WACV**)

- Keywords: VLM, catastrophic forgetting mitigation, PEFT.
- Achievement: Proposed a structural regularization method using Optimal Transport to align joint image-text distributions in CLIP, theoretically proving larger feasible parameter spaces and outperforming existing point-wise alignment techniques.

[4] Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors. (**ICLR**)

- Keywords: Large-scale synthetic dataset generation, identity preservation and generation, privacy.
- Achievement: Generated over 300K synthetic identities; improved model accuracy by **1.79**% across 5 test sets with a **5x** smaller and **311x** faster model than the second-best model.

[5] Logical consistency and greater descriptive power for facial hair attribute learning. (**CVPR**)

- Keywords: correlation alignment, adversarial learning, fine-grained level dataset design.
- Achievement: Improved accuracy by **47.03**% on logical relationship alignment.

AWARDS

Best paper: Consistency and Accuracy of CelebA Attribute Values — VDU @ CVPR2023

SKILLS

Models	JEPA models, Flow models, LLMs
Knowledge	Self-supervised Learning, Reinforcement learning, World Model
Distributed Training	Accelerate, PyTorch DDP, Arrow, Webdataset, HF dataset

SERVE AS

Industrial Chair: IEEE AVVS 2027
Workshop Organizer: DataCV@ICCV25
Reviewer: TPAMI, CVPR, ECCV, ICCV, BMVC, AAAI, PRLETTERS, WACV, F&G, IJCB, ICPR