Tzu-Heng (Brian) Huang

thuang273@wisc.edu zihengh1 zihengh1.github.io zihengh1

Education

2021 – 2026 (Expected) **Ph.D. in Computer Science. University of Wisconsin-Madison.**
Advised by Prof. Frederic Sala.

2016 – 2020 **B.S. in Computer Science. National Chengchi University.**
Advised by Prof. Man-Kwan Shan and Dr. Ling-Jyh Chen. Major GPA: *3.96*

Research Interests

My research focuses on ***data-centric AI for multimodal models***, enabling models to learn more with less supervision. Several works I have worked on, including (i) *online data mixing* for multimodal models, (ii) *fine-grained data selection* for efficient pretraining (e.g., training CLIP on 100+ million samples), (iii) *data curation* using ensemble and objective detection ***(1st place on the DataComp'23 leaderboard)***, (iv) a 500x cheaper *auto-labeling* system over LLM annotators ***(NeurIPS'24 Spotlight)***, and (v) a new accelerated training framework via *model merging*.

Research Experience

May. 2025 – Sep. 2025 **Research Intern. Meta GenAI Llama Team.**
advised by Dr. Ernie Chang.

May. 2024 – Dec. 2024 **AIML Research Intern. Apple Inc.**
advised by Dr. Javier Movellan and Manjot Bilkhu.
— *Automated Model-aware Data Selection for Efficient Pretraining.*
— *Optimizing Domain Mixtures for MLLM Pretraining.*

Aug. 2021 – Present **Graduate Research Student. UW-Madison.**
advised by Prof. Frederic Sala.
— *Data-centric AI for Foundation Models: Auto-labeling and Data Curation.*
— *Parameter Marketplace: Through Model Merging and Auction Agents.*

May. 2023 – May. 2024 **Co-Founder. Awan.AI LLC.**
— *LLM for Traditional Chinese Medicine and Tongue Syndrome Diagnosis.*
— *Automating TCM Diagnosis: RAG-Enhanced LLMs with Reasoning Agents.*

Jun. 2019 – Sep. 2019 **Research Intern. Argonne National Laboratory.**
advised by Dr. Charlie Catlett and Dr. Rajesh Sankaran.
— *Ensemble-based Time Series Calibration for Low-cost Sensors.*

Sep. 2018 – Aug. 2021 **Research Assistant. National Chengchi University.**
advised by Prof. Man-Kwan Shan.
— *Spatio-temporal Modeling in Large-scale Sensor Networks.*

Feb. 2018 – Jul. 2020 **Research Intern. Academia Sinica.**
advised by Dr. Ling-Jyh Chen.
— *Large-scale Air Quality Sensor Network Development.*

Research Publications

1. A. Ge, **T.-H. Huang**, J. Cooper, *et al.*, "R&B: Domain Regrouping and Data Mixture Balancing for Efficient Foundation Model Training," in *submission*, 2025.
2. **T.-H. Huang**, M. Bilkhu, F. Sala, and J. Movellan, "Evaluating Sample Utility for Data Selection by Mimicking Model Weights," in *submission*, 2025. URL: https://arxiv.org/abs/2501.06708.

3. **T.-H. Huang**, C. Cao, V. Bhargava, and F. Sala, "The ALCHEmist: Automated Labeling 500x CHEaper than LLM Data Annotators," in *Neural Information Processing Systems (NeurIPS)* ***[Spotlight (Top 3%)]***, 2024. URL: https://arxiv.org/abs/2407.11004.

4. W. Tan, N. Roberts, **T.-H. Huang**, *et al.*, "MoRe Fine-Tuning with 10x Fewer Parameters," in *ICML Workshop: Efficient Systems for Foundation Models (ES-FoMo) and ICML Workshop: Foundation Models in the Wild.*, 2024. URL: https://arxiv.org/abs/2408.17383.

5. N. Roberts, X. Li, D. Adila, *et al.*, "Geometry-Aware Adaptation for Pretrained Models," in *Neural Information Processing Systems (NeurIPS)*, 2023. URL: https://arxiv.org/abs/2307.12226.

6. **T.-H. Huang**, C. Cao, S. Schoenberg, H. Vishwakarma, N. Roberts, and F. Sala, "ScriptoriumWS: A Code Generation Assistant for Weak Supervision," in *ICLR Workshop: Deep Learning For Code (DL4C)*, 2023. URL: https://arxiv.org/abs/2502.12366.

7. **T.-H. Huang**, C. Shin, S. J. Tay, D. Adila, and F. Sala, "Multimodal Data Curation via Object Detection and Filter Ensembles," in *ICCV Workshop: Towards the Next Generation of Computer Vision Datasets (TNGCV)* ***[1st place on the Datacomp leaderboard (small-scale filtering track)]***, 2023. URL: https://arxiv.org/abs/2401.12225.

8. **T.-H. Huang**, H. Vishwakarma, and F. Sala, "Train 'n Trade: Foundations of Parameter Markets," in *Neural Information Processing Systems (NeurIPS)*, 2023. URL: https://arxiv.org/abs/2312.04740.

9. N. Roberts, X. Li, **T.-H. Huang**, *et al.*, "AutoWS-Bench-101: Benchmarking Automated Weak Supervision with 100 Labels," in *Neural Information Processing Systems (NeurIPS)*, 2022. URL: https://arxiv.org/abs/2208.14362.

10. **T.-H. Huang**, C.-H. Tsai, and M.-K. Shan, "Key Sensor Discovery for Quality Audit of Air Sensor Networks," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2020. URL: https://dl.acm.org/doi/abs/10.1145/3386901.3396606.

Miscellaneous

Awards

- 2024 — **Spotlight Paper (Top 3%): The Alchemist**, selected by NeurIPS'24.
- 2023 — **ICCV Datacomp Competition**, won the first place in the small-scale filtering track.
- **Scholar Award**, granted by NeurIPS'23.
- 2021 — **First-year Departmental Scholarship**, granted by UW-Madison.
- 2020 — **Research Intern Scholarship**, granted by National Chengchi University.
- **Undergrad Research Scholarship**, granted by Ministry of Science and Technology.

Invited Talks

- Dec. 2019 — **IoT Project Development**, invited by Nangang High School (Taiwan).
- Sep. 2019 — **Internship Research Talk**, invited by National Chengchi Univerisity.
- Jul. 2019 — **LASS Conference: International Session**, invited by Academia Sinica.
- Mar. 2019 — **Techbang Magazine: PiM25 Project**, invited by Techbang Magazine.
- **Raspberry Pi Jam: PiM25 Project**, invited by Raspberry Pi Foundation (Taiwan).
- Jan. 2019 — **Raspberry Pi Meetup: PiM25 Project**, invited by Raspberry Pi Foundation (Taiwan).

Academic Services

- 2021 – Present — **Paper Reviewer**, NeurIPS, ICLR, ICML, CVPR, CoLLAs, and DMLR.
- 2023 — **Co-organizer**, AutoML Cup in AutoML Conference.
- 2022 – 2023 — **President of Student Association of Taiwan**, UW-Madison.

Miscellaneous (continued)

2021 – 2022 **Vice President of Student Association of Taiwan**, UW-Madison.

Skills

Programming Languages Python, R, C++/C, SQL, LaTeX, and Shell Programming.

Technologies (Distributed) PyTorch, Tensorflow, Keras, ShinyApp, PostgreSQL, and Vim.