Huong Ngo

Email: zoengo2002@gmail.com Phone: (508) 488-7263 GitHub: github.com/huongngo-8 Website: huongngo-8.github.io

Education

University of Washington, Seattle **Seattle, WA**
Applied Computational Mathematical Sciences: Data Science & Statistics B.Sc. 3.71 GPA Sep 2020–Aug 2024

Publications

OLMoASR: Open Models and Data for Training Robust Speech Recognition Models
Huong Ngo, Matt Deitke, Martijn Bartelds, Sarah Pratt, Josh Gardner, Matt Jordan, Ludwig Schmidt

Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models **CVPR 2025**
Matt Deitke, Christopher Clark, Sangho Lee, ..., Huong Ngo, et al. (90 citations)

Objaverse-XL: A Universe of 10M+ 3D Objects **NeurIPS 2023**
Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, et al. (685 citations)

Work Experiences

Vercept (acquired by Anthropic) **Seattle, WA**
Member of Technical Staff Jun 2025–Mar 2026

- Led data curation research to improve computer-use agents (planner + grounding VLM); iterated on data mixtures and training recipes, built pipeline processing 414K+ images and 2.28M labels to address distribution shift; improved GUI grounding by up to 17% and agent task success by up to 18%
- Improved grounding VLM on external benchmarks through large-scale label enrichment and test-time compute scaling; achieved 20+ point gains on ScreenSpot-Pro (52%→72%) and 2+ points on ScreenSpotV2 (92%→94%)
- Designed planner synthetic data collection system to collect 5K+ trajectories (1.2K+ annotation hours) and automated agent task construction (2.1K tasks) for planner model fine-tuning; built annotation infrastructure with AWS SQS and deduplication
- Architected agent evaluation framework (14 categories) with automatic environment generation and programmatic verification; built benchmarks to analyze failure modes across planner and grounding models to prioritize data collection efforts

Allen Institute for AI **Seattle, WA**
Pre-Doctoral Researcher Aug 2024–Jun 2025

- Led OLMoASR, the first fully open reproduction of OpenAI's Whisper; designed and implemented web-scale data curation pipeline to filter 3M hours of web-scraped audio down to 1M-hour high-quality training set
- Developed data quality filters (text heuristics, language alignment, fuzzy deduplication); iterated on 50+ data and training recipes by ablating controlled experiments on downstream tasks; isolated filter contributions (up to 16.5% improvement) to guide data curation decisions
- Built multi-node distributed training infrastructure scaling from 39M to 1.5B parameters using PyTorch DDP/FSDP; debugged distributed training stability issues across multi-node GPU clusters with PyTorch Profiler
- Pretrained models competitive with Whisper across all scales on 21 benchmarks; released models on HuggingFace with 470+ GitHub stars

Allen Institute for AI **Seattle, WA**
Research Intern, PRIOR Team Oct 2023–Dec 2023

- Developed a scalable video filtering pipeline by training a classifier on CLIP visual embeddings with human-annotated storyboard quality rankings as labels, replacing costly manual annotation with model-based filtering
- Built end-to-end data collection pipeline on AWS MTurk for large-scale video quality annotation (5K+ annotations) to guide data curation decisions for pretraining
- Designed evaluation benchmark to assess video generation model output quality

University of Washington, Allen Center for Computer Science and Engineering **Seattle, WA**
Deep Learning Research Assistant, *Supervisor: Matt Deitke* Mar 2023–Aug 2024

- Conducted data-centric machine learning and multimodal research in the Reasoning, AI, Vision Lab (RAIVN)
- Worked on Objaverse-XL by constructing web-scale data processing pipelines using CLIP to annotate 120M 3D objects

- Initiated OLMoASR; built data preprocessing pipeline to segment audio-transcript pairs into WebDataset format; ran SLURM job arrays to process 70K+ shards across compute cluster
- Developed distributed training infrastructure with PyTorch DDP; debugged multi-GPU synchronization issues (hanging, deadlocks) and optimized multi-GPU training throughput with PyTorch Profiler

Teaching Experiences

Machine Learning and Database Teaching Assistant — **Seattle, WA**
Paul G. Allen Center for Computer Science and Engineering — Sep 2022–June 2023

Statistics Tutor — **Seattle, WA**
University of Washington, Department of Statistics — Sept 2021–June 2022

Reviewing

CVPR 2026 – Conference on Computer Vision and Pattern Recognition
ICML 2026 – International Conference on Machine Learning

Skills

Languages: Python, SQL, R, Java
Technologies: PyTorch, PyTorch FSDP/DDP, NumPy, pandas, matplotlib, PyTorch Lightning, Weights and Biases, AWS SQS, AWS MTurk, webdataset, HuggingFace, ffmpeg, vllm, Ray, SparkSQL, BeautifulSoup, scikit-learn, PySpark, Docker
Developer Tools: Jupyter, GitHub/Git, Slurm, CUDA

Relevant Coursework

University of Washington, Seattle — **Seattle, WA**
Data-Centric Machine Learning, Machine Learning Systems, Machine Learning for Big Data, Machine Learning, Artificial Intelligence, Databases, Data Structures & Algorithms, Linear Algebra, Statistics & Probability

References

Ludwig Schmidt
Assistant Professor of Computer Science at Stanford University
Member of Technical Staff at Anthropic
Best paper awards at ICML and NeurIPS; led LAION-5B, OpenCLIP, and DataComp

Matt Deitke
AI Researcher at Meta Superintelligence Lab (MSL)
Co-Founder of Vercept
Outstanding Paper at NeurIPS 2022 (ProcTHOR); Best Paper Honorable Mention at CVPR 2025 (Molmo)

Kiana Ehsani
CEO and Co-Founder at Vercept
Previously led Embodied AI team at Allen Institute for AI
Core contributor to AI2-THOR; 31 publications in embodied AI and visual reasoning

Ross Girshick
Co-Founder at Vercept
Previously Research Scientist at Meta AI (FAIR)
Creator of SAM, R-CNN, Faster R-CNN, Mask R-CNN; 18th most cited researcher in science (660K+ citations)