MINGFEI CHEN

+1(206) 945-4355 ◇ lasiafly@uw.edu ◇ Seattle, WA
◇ Personal Website ◇ Google Scholar ◇ LinkedIn

EDUCATION

University of Washington Sep 2023 -
Ph.D. in Electrical and Computer Engineering, GPA: 3.98/4.00 *Seattle, WA*
Google PhD Fellowship 2025.

University of Washington Sep 2021 - Jun 2023
M.S. in Electrical and Computer Engineering, GPA: 3.98/4.00 *Seattle, WA*

Huazhong University of Science and Technology Sep 2016 - Jun 2020
B.S. in Computer Science and Technology, GPA: 3.96/4.00 *Wuhan, China*
Experimental Program for Exemplary Engineers.
Undergraduate excellent student and outstanding undergraduate graduation thesis awards.

PROJECTS

Research Interest: Multi-modal research for spatial reasoning and interaction in 3D scenes, as well as the related applications on multi-modal LLMs, XR devices and robotics.

Research Skills: Multimodal Learning, Audio-Visual, 3D Vision, Computer Vision, Large Language Models, Robotic Systems, Deep Learning, Python, PyTorch, Distributed Training, Linux.

NVIDIA Research (AMRI and Seattle Robotics Lab)
Aug 2026 – Present
Mentors: Koki Nagano, Amrita Mazumdar, Shalini De Mello *Seattle, WA*
Project: Multimodal Foundation Model for Human–Robot Interaction in Dynamic 4D Audiovisual World

- Developing an HRI foundation model that fuses spatial vision, audio, speech, and motion into persistent 4D representations of multiple people, objects, and their directed interactions.
- Designing a joint world–action model that forecasts future entity states and robot actions, grounding dialogue, navigation, and manipulation within an evolving audiovisual environment.
- Building a multi-rate front-end–back-end architecture that combines real-time on-device perception and control with iterative planning and replanning, evaluated through physical and social interaction scenarios.

Meta Reality Labs Research
Jun 2025 – Jan 2026
Mentors: Yue Liu, Homanga Bharadhwaj, Yifan Wang, Zhengqin Li *Redmond, WA*
Project: Flowing from Reasoning to Motion: Learning 6DoF Hand Trajectory Prediction from Egocentric Human Interaction Videos

- Built EgoMAN, the first large-scale egocentric benchmark for *stage-aware 3D hand trajectory prediction* with structured visual–linguistic–motion reasoning supervision.
- Proposed a reasoning-to-motion architecture that connects VLM intent reasoning with 6DoF trajectory generation via structured token interfaces and progressive training.
- Achieved state-of-the-art 6DoF motion forecasting, with improved interpretability and cross-task generalization on diverse, unseen egocentric manipulation scenarios.

NeuroAI Lab, University of Washington, Seattle
Dec 2021 - Present
Advisor: Eli Shlizerman *Seattle, WA*
Project: Omni-Modal Foundation Models for Realistic 4D Dynamic Scenes (Ongoing)

- Developing omni-modal foundation models that integrate vision, binaural audio, language, motion, and 3D geometry for semantic–spatial understanding, retrieval, grounding, and reasoning in realistic dynamic scenes.

- SceneBind: Proposed a scene representation combining global embeddings with object-centric slots that encode semantics, azimuth, elevation, distance, and uncertainty; designed semantic–spatial matching for cross-modal retrieval, object grounding, and zero-shot audio-visual localization.
- SAVVY (NeurIPS 2025 Oral): Introduced SAVVY-Bench, the first benchmark for 3D spatial reasoning in dynamic audio-visual scenes, and a training-free AV-LLM pipeline that integrates egocentric spatial tracks with dynamic global maps to improve spatial awareness.

Project: 3D Audio-Visual Represnetation and Reconstruction from Sparse Audio-Visual Samples

- Explored audio-visual scene representation for 3D world to support the cross-modal tasks such as spatial audio scene reconstruction.
- AV-Cloud: Proposed the Audio-Visual Cloud, a set of sparse AV anchor points derived from camera calibration, to learn the audio-visual scene representation and generate spatial audio for arbitrary listener locations. Developed a novel approach for rendering high-quality spatial audio in 3D scenes that is synchronized with the visual stream but does not rely on explicit visual cues. (NeurIPS24)
- BEE: Developed a real-time end-to-end integrated rendering pipeline, for dynamic audio reconstruction from sparse audio-visual samples. (ICCV23)
- INRAS: Proposed a method for high-fidelity impulse responses modeling at any emitter-listener positions. (NeurIPS22)

Meta Codec Avatars Lab Jun 2024 - Dec 2024
Mentors: Israel D. Gebru, Alexander Richard *Pittsburgh, PA*
Project: Visually-guided Ambient Sound Field Modeling for Navigable Codec Spaces (CVPR25 Highlight)

- Developed a novel approach, SoundVista, to reconstruct target binaural sound using reference ambisonic microphones at arbitrary locations, enhancing ambient sound modeling in diverse environments.
- Created benchmarks in both simulated Soundspace environment and real-world captured room, conducting research from synthetic agents to real.
- Pioneered the integration of 3D visual capture data with acoustic parameters to create a pre-trained visual representation model, facilitating improved acoustic scene analysis. Utilized the model to sample optimal reference microphone positions, which can significantly boost performance within the same microphone budget.
- Implemented adaptive reweighting of input audio channels using the pretrained visual model and room geometry, optimizing reconstruction across varying room sizes and layouts without constraints on microphone quantity.

National University of Singapore & Sea AI Lab Jun 2021 - Nov 2021
Advisor: Shuicheng Yan, Jiashi Feng *Singapore*
Project: Controllable High-Fidelity 3D Human Rendering Under Sparse Views (ECCV22)

- Developed a geometry-guided progressive NeRF for efficient human body rendering from sparse views (3 cameras, 120° apart), with a multi-view feature integration approach for enhanced occluded information accuracy using geometry guidance.
- Curated a large-scale dataset for pretraining and validated the method's real-world robustness.
- Achieved 70% rendering time reduction, taking just 175ms on RTX 3090, outperforming industry standards.

Sensetime & University of Washington, Seattle Nov 2020 - Jun 2021
Advisor: Jenq-Neng Hwang *Beijing, China*
Project: Online Multi-Object Tracking (MOT)

- Proposed an innovative online MOT framework for tailored feature aggregation in detection and association.
- Crafted a track association module for combined location and appearance matching using motion and appearance embeddings.

- Achieved state-of-the-art online MOT tracking on MOT17, enhancing association effectiveness and maintaining top detection accuracy.

Sensetime Research Jul 2020 - Mar 2021
Advisor: Si Liu *Beijing, China*
Project: Human-Object Interaction (HOI) Detection (CVPR21)

- Formulated HOI detection as a set prediction problem, addressing instance-centric and location limitations.
- Developed a one-stage HOI framework with a transformer for optimal feature aggregation and introduced an instance-aware attention module.
- Achieved 31% accuracy improvement over leading one-stage methods on the HICO-DET dataset without additional features.

PUBLICATIONS

[1] **Mingfei Chen**, Zijun Cui, Ruoke Zhang, Hyeonggon Ryu, Eli Shlizerman. **"SceneBind: Binding What and Where Across Vision, Audio and Language."** Under Review.
[2] **Mingfei Chen**, Yifan Wang, Zhengqin Li, Homanga Bharadhwaj, Yujin Chen, Chuan Qin, Yuan Tian, Ziyi Kou, Eric Whitmire, Raj Sodhi, Hrvoje Benko, Eli Shlizerman, Yue Liu.**"EgoMAN: Interaction-Structured Reasoning for Egocentric 3D Hand Trajectory Prediction"** European Conference on Computer Vision (ECCV), 2026.
[3] **Mingfei Chen***, Zijun Cui*, Xiulong Liu*, Jinlin Xiang, Caleb Zheng, Jingyuan Li, Eli Shlizerman. **"SAVVY: Spatial Awareness via Audio-Visual LLMs through Seeing and Hearing."** Neural Information Processing Systems (NeurIPS), 2025 **(Oral - Acceptance Rate $< 0.4\%$)**.
[4] **Mingfei Chen**, Israel D. Gebru, Ishwarya Ananthabhotla, Christian Richardt, Dejan Markovic, Steven Krenn, Todd Keebler, Jacob Sandakly, Eli Shlizerman, Alexander Richard. **"SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding."** Computer Vision and Pattern Recognition (CVPR), 2025 **(Highlight - Acceptance Rate $< 2.9\%$)**.
[5] **Mingfei Chen**, Eli Shlizerman. **"AV-Cloud: Spatial Audio Rendering Through Audio-Visual Cloud Splatting."** Neural Information Processing Systems (NeurIPS), 2024.
[6] **Mingfei Chen**, Kun Su, Eli Shlizerman. **"Be Everywhere - Hear Everything (BEE): Audio Scene Reconstruction by Sparse Audio-Visual Samples."** International Conference on Computer Vision (ICCV), 2023.
[7] Kun Su*, **Mingfei Chen***, Eli Shlizerman. **"INRAS: Implicit Neural Representation for Audio Scenes."** Neural Information Processing Systems (NeurIPS), 2022.
[8] **Mingfei Chen**, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, Shuicheng Yan. **"Geometry-Guided Progressive NeRF for Generalizable and Efficient Neural Human Rendering."** European Conference on Computer Vision (ECCV), 2022.
[9] **Mingfei Chen***, Yue Liao*, Si Liu, Zhiyuan Chen, Fei Wang, Chen Qian. **"Reformulating HOI Detection as Adaptive Set Prediction."** Computer Vision and Pattern Recognition (CVPR), 2021.
[10] Jingyuan Li, Trung Le, Chaofei Fan, **Mingfei Chen**, Eli Shlizerman. **"Brain-to-Text Decoding with Context-Aware Neural Representations and Large Language Models."** *Journal of Neural Engineering*, Oct. 2025.
(* equal contribution.)

AWARDS & HONORS

[2025] Google PhD Fellowship 2025 in Machine Perception (North America).
[2023] International Conference on Computer Vision Diversity, Equity & Inclusion (DEI) Award.
[2020] Huazhong University of Science and Technology Outstanding Undergraduate Graduation Thesis.
[2018, 2019] Huazhong University of Science and Technology Merit Student Scholarship.
[2018] Meritorious Winner of Mathematical Contest In Modeling (MCM/ICM).
[2017] Huazhong University of Science and Technology Undergraduate Excellent Student (Top 1% in 35000).