Foundations of Multisensory Artificial Intelligence

Paul Pu Liang, Machine Learning Department, School of Computer Science, Carnegie Mellon University

Most of today's AI systems only perceive a narrow slice of the world. For example, AI assistants infer human intents only from the words we say but not our tone of voice and subtle facial expressions. In healthcare, while AI can perceive singular modalities like medical scans, they are unable to holistically understand medical histories, read scans and vitals, and monitor patient behaviors to assist healthcare professionals. My scientific goal is to **understand the machine learning principles of multisensory intelligence in order to design practical AI systems that can integrate, learn from, and interact with a diverse range of real-world sensory modalities**. Multisensory AI can have many practical benefits, including (1) autonomous agents that can perceive language, vision, audio, and touch to communicate with humans and execute complex tasks, (2) technologies that safely improve human physical, emotional, and social wellbeing from verbal and nonverbal communication, smartphones, and wearable devices, and (3) analyzing sensors that monitor the health of computer and information systems.

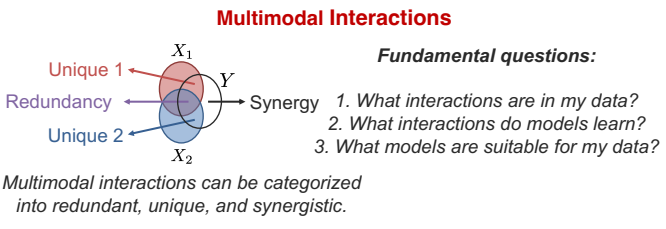
Multisensory AI introduces core challenges in quantifying the useful information in each data source, modeling how modalities interact to create new information when integrated, and handling real-world heterogeneity across diverse modalities. My research has developed the foundations of multisensory AI spanning three pillars:

1. I have developed a new theoretical framework formalizing the **useful information** in each modality and how **modalities interact** with each other to give rise to new information for a task [14], such as sarcasm identified from the incongruity between spoken words and vocal expressions. Quantifying the types of interactions a multimodal task requires enables researchers to decide which modality to collect [14, 16], design suitable approaches to learn these interactions [15, 20], and analyze whether their model has succeeded in learning [12].
2. I have also built methods to learn multimodal interactions across time from sequences such as dialog, speech, videos, medical time-series, and physical sensors. Methods I designed such as **cross-modal attention** [2, 18] and **multimodal transformers** [27] progressively capture interactions across time, such as learning that eye-rolling and sighing long after positive spoken speech can indicate sarcasm. Cross-modal attention is a scalable backbone that underpins today's multimodal foundation models such as ViLBERT, Flamingo, and UniVL.
3. Finally, I have built **multisensory and multitask foundation models** that handle heterogeneity at scale across many real-world modalities. Collaborating with psychiatrists at the Univ. of Pittsburgh, Columbia, and the Univ. of Oregon, my methods to predict mood fluctuations in patients using recorded smartphone data, typed text, and typing patterns now aid **mental health** practitioners in their treatment [6]. Similar technologies I developed are used by doctors at Harvard Medical School for **cancer prognosis** using genomics data and pathology images [11, 22], and by roboticists to **control robot arms** based on cameras and touch sensors [8].

My research has been recognized by the Siebel Scholars Award, Waibel Presidential Fellowship, Facebook PhD Fellowship, Center for ML and Health Fellowship, Rising Stars in Data Science, and 3 best paper/honorable mention awards at ICMI [18] and NeurIPS workshops [3, 21]. A taxonomy and pedagogy for multimodal ML I co-created [9] now serves as a foundational resource, and I have led the instruction of core AI courses at CMU and tutorials at international conferences. The scientific impact of multisensory AI is immense, and I look forward to interdisciplinary collaborations between CS, statistics, robotics, psychology, and healthcare as a professor.

1. Foundations of multimodal interactions: redundancy, uniqueness, and synergy

Multimodal interactions can be categorized into **redundancy**, **uniqueness**, and **synergy**: *redundancy* quantifies information shared between modalities, such as smiling while telling an overtly humorous joke; *uniqueness* quantifies the information present in only one, such as each medical sensor designed to provide new information; and *synergy* quantifies the emergence of new information using both, such as conveying sarcasm through disagreeing verbal and nonverbal cues [9]. I have developed new mathematical frameworks quantifying the interactions required for a task and those learned by trained models [8, 10, 13, 14], yielding principled approaches to learn interactions from data [15, 16].



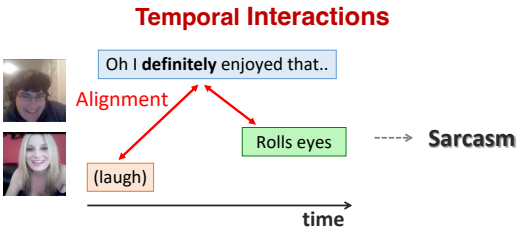
Multimodal interactions can be categorized into redundant, unique, and synergistic.

1.1 Quantifying multimodal interactions: By introducing a new connection between information theory and multimodal interactions [14], I designed *scalable estimators to quantify the interactions in large-scale multimodal datasets and those learned by multimodal models*. These estimators are based on max-entropy convex optimization and a scalable end-to-end estimator suitable for high-dimensional continuous data. We implemented these methods in two real-world case studies: (1) By collaborating with psychiatrists who study their patient's daily mood from smartphone typing patterns and application usage, we found that words like '*love*', '*thank*', and '*haha*' were typed faster during positive moods, while '*don't*' and '*talk*' were typed slower during negative moods [6]. (2) With computational pathologists, we found that cancer of the lower-grade glioma was predictive from genomics readings alone, while pancreatic cancer showed synergy between genomics and pathology images [22]. Using these insights, we designed practical models to learn synergistic interactions that significantly outperformed prior work using individual modalities. Furthermore, domain experts appreciated the transparency that these methods convey as opposed to black-box neural networks, resulting in trust and adoption in real-world practice.

1.2 Implications on performance, modeling, and visualizations: Quantifying interactions enables us to prove *theoretical guarantees of model performance* [16], which helps users decide which modalities to collect to improve performance. It also inspires new multimodal approaches: (1) going beyond shared information to capture task-relevant *unique information* [15, 20], and (2) capturing synergy via the *disagreement* between modality predictors beyond agreement as is typically done [16]. Finally, we designed MULTIVIZ [12], a *tool enabling humans to visualize multimodal interactions in trained models*. Users found MULTIVIZ useful for understanding the features learned by neural networks and performing error analysis on misclassifications. Users even used it to find real bugs in open-source models, and it is currently used for interpretation efforts in affective computing and healthcare.

2. Learning interactions across time in multimodal sequences

It is increasingly common to see temporal modalities such as dialog, speech, videos, medical time-series, and physical sensors. This introduces a critical challenge of **learning interactions across time**. In the example on the right, can AI understand that eye-rolling occurring long after positive spoken speech and laughter can indicate sarcasm? My research has advanced the state-of-the-art in modeling multimodal sequential data via cross-modal attention methods [2, 4, 24] and multimodal transformers [11, 27].

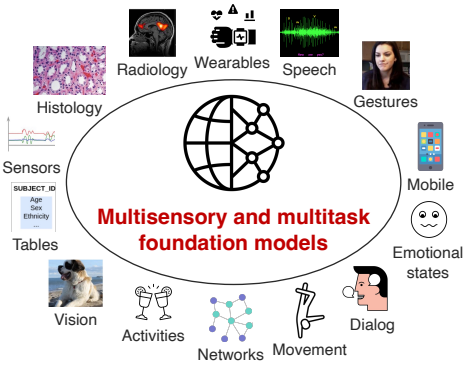


Learning interactions across time from multimodal sequences: dialog, speech, videos, physical sensors, medical data etc.

2.1 Temporal multimodal fusion: While prior work summarized temporal modalities into a single static feature before fusion, I developed a new method for *temporal fusion at the fine-grained level* of individual words, gestures, and vocal expressions [2, 18]. At each time step, temporal fusion first highlights a subset of signals via learned attention weights before fusing with previous signals, which captures progressively more complex multimodal interactions. For example, the first stage learns happiness from laughter, the second stage fuses this with the word 'definitely enjoyed' which creates strong happiness, and the third stage fuses this with the eye-rolling expression that together reflect sarcasm. Attention weights are learned recursively based on the current multimodal input and previous signals, and we call this module **cross-modal attention**. Using these methods, we created systems that can perceive human emotions, navigate social interactions, and even identify humor and sarcasm. My line of research in fine-grained temporal fusion [2, 4, 18, 29] has been cited more than 1000 times and has been applied broadly for temporal fusion of video, multiomics, medical sensors, and physical sensors.

2.2 Multimodal transformers: While I initially built temporal fusion upon recurrent models, in subsequent work I co-created the first **multimodal transformer** to enable *multimodal temporal learning using parallel transformers* [27]. The multimodal transformer learns a cross-modal attention matrix to highlight related signals across time (e.g., rolling eyes and sighing). This matrix is used to learn a new representation for each modality fused with other modalities in parallel over the entire sequence, which provides huge efficiency gains when trained on modern GPUs. Cross-modal attention is a powerful backbone that has been applied to a variety of applications, including today's multimodal foundation models such as ViLBERT, Flamingo, and UniVL.

3. Towards multisensory foundation models



Learning from extreme heterogeneity across diverse modalities.

Finally, I have led open-source efforts to build the next generation of **foundation models that can process the multisensory world**. By extending the impressive dialog, reasoning, and generalization abilities of large language models, real-world sensory processing can enable general-purpose digital literacy, physical dexterity, and social intelligence [11, 17]. At the same time, I research new methods to improve the safety and fairness of these large foundation models so that they can be deployed reliably in the real world.

3.1 Multisensory foundation models: I created HighMMT [11], a model with an *unprecedented range of modalities and skills* including answering queries about visual images, comprehending human communication, classifying diseases from various medical sensors, understanding multimodal user interfaces in HCI, and predicting robot movement based on visual and tactile sensors. Due to the extreme heterogeneity across diverse modalities, a critical component of HighMMT is **heterogeneity quantification**: measuring the similarity and differences between modalities. For example, should I model human speech and gestures the same way as medical data? We found that sharing interactions for similar modalities yields benefits of scale, whereas different modalities should be learned separately [11]. Through this new approach, HighMMT performance consistently improves with more modalities. It even transfers to entirely new modalities and tasks, which is especially useful in the healthcare and affective computing domains where data is limited, so directly learning on them is difficult.

3.2 Safety of language models: I am currently building new **multisensory language models** that can answer open-ended questions about the digital, physical, and social worlds beyond fixed-label classification [17]. However, their capacity for free-form text generation can be a double-edged sword since they can be easily prompted to generate *unsafe stereotypes, derogatory language, and biased content*. I was one of the early researchers to investigate how large language models pretrained on vast amounts of data may reflect the racial and gender biases contained in those datasets, creating one of the first benchmarks to **measure social biases** and methods to **mitigate biases** in language models [5, 7]. Our work is now part of standard LLM evaluation suites like BIG-Bench and has been used to benchmark the safety of ChatGPT, GPT-4, Lamda, Flan, Palm, and more.

3.3 Multisensory resources: Finally, I have created the MultiBench benchmark [8] and model zoo [10] enabling large-scale training of multisensory foundation models for affective computing, HCI, robotics, healthcare, multimedia, and education [23]. Our resources and models have become the community standard at ML, language, and vision conferences, including starting a new research field in human multimodal language [1, 19, 30], and inspiring many theoretical studies and modeling paradigms in multimodal affect, robotics, HCI, and IoT.

The future: Dynamic, interactive, and affective multisensory intelligence

My long-term goal is to create systems that can **dynamically interact with and assist humans in the real world** in order to improve their physical, emotional, and social wellbeing. This includes communicating with humans through verbal and nonverbal channels to understand our intents, assisting us with complex tasks when necessary, suggesting physical and social activities to improve our overall wellbeing, and engaging those who need social and emotional support through personalized interactions. These goals require new fundamental advances in dynamic, interactive, and affective multisensory intelligence, while addressing safety, robustness, personalization, and privacy concerns. I am excited to significantly extend my research in the following directions:

Foundations of dynamic multimodal learning: A major challenge in real-world settings is to choose what modalities to collect for the task: doctors sequentially decide what medical tests to administer based on prior readings, and autonomous agents are carefully designed to balance the utility and cost of their onboard sensors. This new paradigm of **dynamic modalities** will require rethinking the current way benchmarks and methods are designed in multimodal research. Learning over dynamic modalities will require quantifying information gain

and performance, while balancing the drawbacks of increased complexity, increased heterogeneity, and decreased robustness due to more modalities. These insights can also be applied to other CS subfields with heterogeneous and interconnected data, including distributed systems, federated learning, and active learning.

Interactive multisensory agents: I aim to build multisensory agents that can safely interact with the world, such as navigating multimedia content on the web to assist humans in online shopping, travel bookings, content management, and other web tasks. While there is great interest in agents that can code and solve language-based computer tasks, unlocking their full capabilities with rich image, video, and audio processing is still a grand challenge. These applications require learning representations across sequentially changing multimodal inputs during decision-making. By building on my current work in multimodal foundations and future work in dynamic learning, I aim to develop **new foundations of interactive multimodal systems** with theoretical guarantees to ensure real-world safety [25]. Finally, in collaboration with roboticists and engineers, I aim to transfer insights from the virtual world to real-world physical agents that can interact based on sight, sound, dialog, force, and other wireless Internet-of-things sensors including GPS, WiFi, depth, thermal, and capacitance.

Multisensory AI for physical, emotional, and social wellbeing can have an enormous impact in helping us track indicators of wellbeing [6], and using this information to suggest physical and social activities, or even engage users (especially the youth, elderly, and those who need social and emotional support) through personalized interactions [28] to improve their wellbeing. I plan to continue my existing collaborations with medical schools [6, 14, 22] and build new collaborations to study how non-invasive, regularly collected sensors (e.g., verbal and nonverbal communication, smartphones, wearable devices, daily mood, physical activities, and social interaction) can complement doctor appointments, lab tests, medical imaging, and therapy sessions. Key challenges include **social-emotional AI** that can understand behavioral cues, engage in social interaction, and respect social norms and commonsense, as well as **personalization** while maintaining the **privacy** of personal information [3, 26].

I look forward to leading a group to push forward this research agenda, which will be enriched by existing and new collaborations with experts in various CS subfields, statistics, robotics, medicine, and psychology. I hope to make significant contributions to the scientific understanding of multisensory intelligence in order to create the next generation of AI systems that can assist and augment human capabilities.

References

[1] **Paul Pu Liang**, Ruslan Salakhutdinov, and Louis-Philippe Morency. Computational Modeling of Human Multimodal Language: The MOSEI Dataset and Interpretable Dynamic Fusion. *Carnegie Mellon University*, 2018.

[2] **Paul Pu Liang**, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal Language Analysis with Recurrent Multistage Fusion. In *EMNLP*, 2018.

[3] **Paul Pu Liang**, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think Locally, Act Globally: Federated Learning with Local and Global Representations. *NeurIPS Workshop on Federated Learning (distinguished student paper)*, 2019.

[4] **Paul Pu Liang**, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization. In *ACL*, 2019.

[5] **Paul Pu Liang**, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards Debiasing Sentence Representations. In *ACL*, 2020.

[6] **Paul Pu Liang**, Terrance Liu, Anna Cai, Michal Muszynski, Ryo Ishii, Nick Allen, Randy Auerbach, David Brent, et al. Learning Language and Multimodal Privacy Preserving Markers of Mood from Mobile Data. In *ACL*, 2021.

[7] **Paul Pu Liang**, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards Understanding and Mitigating Social Biases in Language Models. In *ICML*, 2021.

[8] **Paul Pu Liang**, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[9] **Paul Pu Liang**, Amir Zadeh, and Louis-Philippe Morency. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv preprint arXiv:2209.03430*, 2022.

[10] **Paul Pu Liang**, Yiwei Lyu, Xiang Fan, Arav Agarwal, Yun Cheng, Louis-Philippe Morency, and Ruslan Salakhutdinov. MultiZoo & MultiBench: A Standardized Toolkit for Multimodal Deep Learning. In *JMLR*, 2022.

[11] **Paul Pu Liang**, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, et al. High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning. In *TMLR*, 2022.

[12] **Paul Pu Liang**, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. MultiViz: Towards Visualizing and Understanding Multimodal Models. In *ICLR*, 2023.

[13] **Paul Pu Liang**, Yun Cheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal Fusion Interactions: A Study of Human and Automatic Quantification. In *ICMI*, 2023.

[14] **Paul Pu Liang**, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, et al. Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. In *NeurIPS*, 2023.

[15] **Paul Pu Liang**, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *NeurIPS*, 2023.

[16] **Paul Pu Liang**, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, et al. Multimodal Learning Without Labeled Multimodal Data: Guarantees and Applications. *arXiv preprint arXiv:2306.04539*, 2023.

[17] **Paul Pu Liang**, Haofei Yu, Shentong Mo, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multisense: A multisensory foundation model. *in progress*, 2023.

[18] Minghai Chen*, Sen Wang*, **Paul Pu Liang***, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning. In *ICMI (best paper honorable mention)*, 2017.

[19] Hai Pham*, **Paul Pu Liang***, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities. In *AAAI*, 2019.

[20] Yao-Hung Hubert Tsai*, **Paul Pu Liang***, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning Factorized Multimodal Representations. In *ICLR*, 2019.

[21] Xiang Fan, Yiwei Lyu, **Paul Pu Liang**, Ruslan Salakhutdinov, and Louis-Philippe Morency. Nano: Nested Human-in-the-Loop Reward Learning for Few-shot Language Model Control. *NeurIPS Workshop on Human in the Loop Learning (best paper nomination)*, 2022.

[22] Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, **Paul Pu Liang**, and Faisal Mahmood. Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction. *arXiv preprint arXiv:2304.06819*, 2023.

[23] Dong Won Lee, Chaitanya Ahuja, **Paul Pu Liang**, Sanika Natu, and Louis-Philippe Morency. Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos. In *ICCV*, 2023.

[24] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, **Paul Pu Liang**, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. In *ACL*, 2018.

[25] Ziyin Liu, Zhikang Wang, **Paul Pu Liang**, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep Gamblers: Learning to Abstain with Portfolio Theory. In *NeurIPS*, 2019.

[26] Liangqiong Qu, Yuyin Zhou, **Paul Pu Liang**, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning. In *CVPR*, 2022.

[27] Yao-Hung Hubert Tsai, Shaojie Bai, **Paul Pu Liang**, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*, 2019.

[28] Amir Zadeh, Michael Chan, **Paul Pu Liang**, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR*, 2019.

[29] Amir Zadeh, **Paul Pu Liang**, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention Recurrent Network for Human Communication Comprehension. In *AAAI*, 2018.

[30] AmirAli Bagher Zadeh, **Paul Pu Liang**, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACL*, 2018.