YUNXIANG PENG

+1(917) 558-5904 ◇ Newark, DE

yxpengcs@udel.edu ◇ https://www.linkedin.com/in/yunxiang-peng/ ◇ https://jefferyy-peng.github.io/

EDUCATION

University of Delaware, Ph.D. in Computer Science — Jan 2024 - Present

Columbia University, M.S. in Electrical Engineering — Sep 2022 - Dec 2023

Xidian University, B.S in Electrical Engineering — Sep 2018 - Jul 2022

SKILLS

Programming Languages	Python, C++, JavaScript
ML Tools	PyTorch, TensorFlow, Hugging Face Transformers, DeepSpeed, PEFT, TRL
Multimodal Modeling	CLIP, BLIP/BLIP-2, MAE, Flamingo, LLaVA, Stable Diffusion, Flux
LLM/Agent Tools	LangChain, LlamaIndex, vLLM

PUBLICATIONS

- **Yunxiang Peng**, Mengmeng Ma, Ziyu Yao, Xi Peng. *Inside-Out: Measuring Generalization in Vision Transformers Through Inner Workings.* **Under Review at CVPR 2026**.
- Mengmeng Ma, **Yunxiang Peng**, Tang Li, Xi Peng. *Seeing Is Not Believing: Detect and Interpret Cancer Segmentation Failures.* **Under Review at CVPR 2026**.
- **Yunxiang Peng**, et al. *Learning-based Synthetic MRI Post-Processing Framework for Automated Contrast Optimization and Brain Segmentation.* **Under Review at ISMRM 2026**.
- Mengmeng Ma, Tang Li, **Yunxiang Peng**, Lu Lin, Volkan Beylergil, Binsheng Zhao, Oguz Akin, Xi Peng. *"Why Is There a Tumor?": Tell Me the Reason, Show Me the Evidence.* **ICML 2025**.
- Ziwen Xie, **Yunxiang Peng**, et al. *Physiolabxr: A python platform for real-time, multi-modal, brain–computer interfaces and extended reality experiments.* **JOSS 2024**.

RESEARCH EXPERIENCE

Research Assistant — *University of Delaware*, 2024 – Present

Research Topcis: Multi-modal Learning, Mechanistic Interpretability, OOD Generalization

- Existing medical segmentation models either provide mask predictions without interpretability or textual justifications without visual evidence, limiting clinical trust. To bridge this gap, we constructed a **dataset** of 180k image–mask–justification triplets, which we used to build a tumor segmentation model. Our model achieved a **5% improvement in segmentation performance** and **enabled textual justifications with visual evidence**. [ICML'25]
- Proposed to **evaluate model generalization** on unlabeled test data via its **inner workings** (circuits). The resulting circuit-based metrics **improved performance prediction accuracy by up to 60%** over existing baselines across diverse OOD benchmarks. [Under review at CVPR 2026]
- Proposed to **use interpretable features from Sparse Autoencoders (SAEs) to detect and correct failure cases** in medical segmentation models. We observed that the SAE features capture rich signals indicative of model failures. Building on this, we developed a pipeline leveraging these features, achieving a 21% improvement in failure detection on medical benchmarks. [Under review at CVPR 2026]

Research Assistant — *Columbia University*, 2023 – 2024

Research Topcis: Multi-modal Learning, Brain-Computer Interface

- **Learning generalizable EEG representations and align it to Large Vision-Language Models**. Using contrastive pretraining to train the EEG encoder, and integrated it and projection layers into Google's Paligemma model. Performed supervised fine-tuning for token alignment and optimized semantic alignment using GRPO (BERTScore reward).

INDUSTRY EXPERIENCE

CHDI Foundation — May 2024 – Sep 2024

AI Research Intern

- Proposed a **data-driven postprocessing framework** for Synthetic MRI (SyMRI), implementing a deep learning–based contrast parameter adjustment module that jointly optimizes contrast generation and downstream segmentation tasks, achieving a 6% improvement in segmentation accuracy comparing to empirically selected contrast parameters.
- Collected and preprocessed SyMRI data to ensure compatibility with deep learning pipelines.
- Collaborated with radiologists to evaluate the learned MRI contrasts and downstream segmentation performance, observing improved visual contrast around key anatomical structures.