Xingjian Diao

Email: xingjian.diao.gr@dartmouth.edu | Website: https://xid32.github.io/ | Phone: 1-4123700998

Linkedin | Github | Google Scholar

Research interests

My research interests focus on multimodal learning. I have published papers and developed code on temporal modeling, efficient training, and audio-video-language integration, advancing state-of-the-art multimodal large language models (**MLLMs**) applications such as multimodal question answering and video captioning.
My GitHub repositories on MLLMs have received **700+ stars ★** .

Education

- **Dartmouth College** *Sep 2022 - Apr 2026 (Expected)*
  Ph.D. student in Computer Science Hanover, USA
  - **Advisor**: Prof. Soroush Vosoughi and Prof. Jiang Gui
- **Northwestern University** *Sep 2020 - Dec 2021*
  Master of Science, Computer Science Evanston, IL
- **University of Pittsburgh** *Aug 2016 - Apr 2020*
  Bachelor of Science, Computer Science Pittsburgh, PA

Internship

- **Amazon** *June 2025 - Sept 2025 (Expected)*
  Applied Scientist Intern Santa Cruz, USA

  IMU2Safety for Safety-Related Human Activity Recognition

  Design and develop a **novel video-to-IMU generation pipeline that addresses the scarcity of large-scale annotated IMU datasets**. While existing SOTA methods primarily target routine daily activities, our approach focuses on synthesizing realistic IMU signals for rare and safety-critical human activities recognition (Safety HAR) such as falling, losing balance, fighting, shoving, dragging, jumping from height, and fainting.

  Computer Vision Ubiquitous Computing Action Recognition

Selected Publications

[Preprint 2025] SoundMind: RL-Incentivized Logic Reasoning for Audio-Language Models
Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, Jiang Gui
Pdf | Github Code; ★ **Starred 500+** | Introduce the Audio Logical Reasoning (ALR) dataset, consisting of 6,446 text-audio annotated samples (e.g., CoT in Audio) specifically designed for complex reasoning tasks and proposes SoundMind, a rule-based reinforcement learning (RL) algorithm tailored to endow audio language models (ALMs) with deep bimodal reasoning abilities.
Large Scale Audio Language Modelling RL Audio Reasoning

[NAACL 2025] Temporal Working Memory: Query-Guided Temporal Segment Refinement for Enhanced Multimodal Understanding
Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, Jiang Gui
Pdf | Github Code; ★ **Starred 300+** | Propose and develop a plug-and-play module to enhance multimodal foundation models' (MFMs) temporal modeling capabilities, using query-guided attention to retain task-relevant video-audio segments, achieving significant performance gains in video captioning, question answering, and video-text retrieval across nine state-of-the-art models.
Multimodal Alignment Video Understanding Audio Modelling

[EMNLP 2024] Learning Musical Representations for Music Performance Question Answering
Xingjian Diao, Chunhui Zhang, Tingxuan Wu, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, Jiang Gui
Pdf | Github Code | Propose and develop a specialized framework for audio-visual modeling in music performances, addressing underexplored multimodal interactions, distinctive music characteristics, and temporal alignment. Introduce annotated rhythmic and music source features, achieving state-of-the-art performance on Music AVQA datasets.
Multimodal QA Multimodal Alignment Representation Learning Audio Modelling

[ACL 2025] Learning Sparsity for Effective and Efficient Music Performance Question Answering
Xingjian Diao, Tianzhen Yang, Chunhui Zhang, Weiyi Wu, Ming Cheng, Jiang Gui
Pdf | Design and develop a sparse learning framework for music AVQA, integrating sparsification strategies to optimize efficiency and accuracy, achieving state-of-the-art performance, 28.32% faster training, and introducing a key-subset selection algorithm that reduces computational resources by 75%.
Multimodal QA Sparsity Learning Multimodal Representation Learning

[WACV 2025] FT2TF: First-Person Statement Text-To-Talking Face Generation

Xingjian Diao, Ming Cheng, Wayner Barrios, SouYoung Jin

Pdf | Github Code | Propose and develop a one-stage end-to-end text-to-talking face generation pipeline driven by first-person statement text, requiring only visual and textual inputs during inference. Experiments on LRS2 and LRS3 demonstrate state-of-the-art performance, showing its ability to generate realistic talking faces effectively from text inputs.

🏷 Multimodal Alignment 🏷 AIGC 🏷 Multimodal Representation Learning

[Preprint 2025] On The Design Choices of Next Level LLMs

Pdf | Github | Yijun Tian, **Xingjian Diao**, Ming Cheng, Chunhui Zhang, Jiang Gui, Soroush Vosoughi, Xiangliang Zhang, Nitesh V. Chawla, Shichao Pei

Provides a comprehensive analysis of current LLM design choices across model architecture, attention mechanisms, post-training strategies, optimization techniques, and data selection, identifying key trends and proposing future research directions for next-generation large language models.

🏷 LLMs 🏷 Attention Mechanisms 🏷 Post-Training 🏷 Optimization 🏷 RL 🏷 Data Selection

[Preprint 2025] Enhancing Digital Pathology Visual Understanding With Sparse Pyramid Attention Networks

Weiyi Wu, **Xingjian Diao**, Chongyang Gao, Xinwen Xu, Siting Li, Jiang Gui

Pdf | Proposes Sparse Pyramid Attention Networks (SPAN) for whole slide image analysis in digital pathology, featuring Spatial-Adaptive Feature Condensation and Context-Aware Feature Refinement modules that preserve spatial relationships while efficiently processing gigapixel-scale images through hierarchical multi-scale representations, achieving superior performance in tumor detection, classification, and segmentation tasks across multiple histopathology datasets.

🏷 Digital Pathology 🏷 Whole Slide Image Analysis

[Preprint 2025] Music Performance Audio-Visual Question Answering Requires Specialized Multimodal Designs

Wenhao You, **Xingjian Diao**, Chunhui Zhang, Keyi Kong, Weiyi Wu, Zhongyu Ouyang, Chiyu Ma, Tingxuan Wu, Noah Wei, Zong Ke, Ming Cheng, Soroush Vosoughi, Jiang Gui

Pdf | Github | Presents the first comprehensive survey of Music Performance Audio-Visual Question Answering (Music AVQA), systematically analyzing how specialized multimodal designs with explicit spatial-temporal modeling are essential for reasoning over continuous, densely layered musical performances, through comparative analysis of existing datasets and methods, while proposing music-specific architectural enhancements to advance multimodal understanding in this unique domain.

🏷 Multimodal Modelling 🏷 Video Understanding 🏷 Music Modelling

[Preprint 2025] An End-to-End Adaptable Prototypical Framework for Explainable Fine-Grained Visual Question Answering

Xingjian Diao, Weiyi Wu, Peijun Qing, Keyi Kong, Ming Cheng, Soroush Vosoughi, Jiang Gui

Pdf | Github Code | Propose ProtoVQA, introducing adaptable prototypes for cross-modal tasks, spatially-constrained matching for geometric variations, and systematic evaluation of visual-linguistic alignment.

🏷 Multimodal QA 🏷 Interpretability 🏷 Multimodal Reasoning

Selected research projects

- **Health Aware Bits (HABits) Lab, Northwestern University** *Sep - Dec 2021*
  Research Assistant Evanston, IL
  Intake Detection Tool with Multiple Classifiers | Github Code
  - Helped develop an approach to detect feeding gestures from the wrist-worn sensor with low inference time and less power consumpting.
  - Wrote and deployed applications in Android Studio to real devices for evaluation, including training, loading, and testing predefined machine learning algorithms.
  - Implemented the DTW (Dynamic time warping), CNN-LSTM, random forest, SVM, KNN, and Naïve-bayes algorithms using Android Studio.
  - Improved the inference time from existing model architectures by implementing alternative algorithms for our models and comparing their inference time and accuracy.
- **Health Aware Bits (HABits) Lab, Northwestern University** *Mar - Aug 2021*
  Research Assistant Evanston, IL
  Interactive Active Learning Annotation Tool | Github Code
  - Developed and designed an interactive annotation software that uses human-in-the-loop machine learning concepts known as “Active Learning” to label time sequences, with the goal of reducing labeling expenses and addressing challenges faced during the annotation process (used PyQt5, cv2, sklearn, xgboost, numpy, pandas, and pyqtgraph).
  - Added functionalities such as time synchronization, plotting, autoloading of raw data and videos, rewinding video frame by frame, automatically locating queried time sequences, time-sequence-labeling, and labeling results export.
  - Applied the clustered entropy active learning method to query the maximally informative samples.
  - Modified the code of QtChart package to perform better colored plot segmentation.
- **University of Texas Southwestern Medical Center** *May - Aug 2020*
  Software Development Engineer Intern Dallas, TX
  iPADshiny (integrated Protein Array Data management,analysis and visualization tools) | Github Code

- ◦ Developed framework in R shiny for a desktop application that enables biologists to conduct protein-array profiling analysis.
- ◦ Added functionalities for each step of auto-antibody profiling analysis including data import, quality control, normalization procedure, batch correction, and result visualizations with multiple options for every step.
- ◦ Implemented Alone, ANOVA, ComBat, and PCA algorithms for Batch Correction, as well as Scaling, RLM, Quantile, loaess, and VSN algorithms for Normalization, in order to conduct data preprocessing.

Services and Skills

- **Skills: Programming Languages:** Python (PyTorch, Scikit-learn, NumPy, Pandas), Java, Android Studio, SQL (PostgreSQL, MySQL), R (shiny), Shell script, **Development Tools:** Git, LaTeX, Markdown, TMUX, Bash
- **Reviewrs:** ACMMM, NEURIPS, EMNLP, ACL Industry Track, ACL, ICCV, CVPR, ICLR, WACV, ICASSP, MINT 2024, ISBI, IJCNN, ICME, AVSS.
- **Teachings:** Graduate TA for Video Understanding (CS89/189, Spring 2024), Machine Learning (CS74/274, Winter 2024), Database Systems (COSC61, Summer 2023), Object-Oriented Programming (COSC10, Spring 2023 & Fall 2022), and Applied Cryptography (COSC62/162, Winter 2023)

Awards

- **Dartmouth Fellowship,** Dartmouth College. *2022-2025*
- **Guarini School of Graduate and Advanced Studies Travel Award,** Dartmouth College. *2025*
- **Biomedical Data Science Travel Award,** Dartmouth College. *2025*
- **IEEE EMBC NextGen Scholar Award,** IEEE EMBC. *2024*