Xingjian Diao

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

Linkedin | Github | Google Scholar

Research interests

My research focuses on **multimodal large language models (MLLMs)** for video, audio, and language understanding. I have developed methods for **multimodal reasoning, efficient multimodal learning, and generative multimodal modeling**, aiming to build scalable and generalizable MLLMs that advance multimodal question answering, video understanding, and audio–visual reasoning across complex real-world scenarios and dynamic environments.
My GitHub repositories on MLLMs have received 1.5k+ Stars ★ .

Education

- **Dartmouth College** *Sep 2022 - Jun 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
  - **Advisor**: Prof. Nabil Alshurafa
- **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
  - **Mentor**: Dr. Heba Aly and **Manager**: Dr. Hongda Mao

High-Frequency Video-to-IMU Synthesis via Physics-Guided Simulation and Hybrid U-Net Refinement
Pdf | Proposed PrimeIMU, a physics-guided video-to-IMU generation framework that fuses low-frequency kinematic cues from 3D video poses with physics-inspired simulated inertial initialization through a hybrid U-Net refinement module, effectively bridging the anatomical–inertial gap to produce high-fidelity, sensor-faithful IMU signals that generalize across activities, devices, and datasets, enabling synthetic-only training, cross-domain adaptation, and scalable deployment of wearable sensing models.
Video Understanding Computer Vision Ubiquitous Computing

Selected 1st-Author Publications [Full List]

[EMNLP 2025 (Oral)] 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 1k+** | Dataset | Introduced the Audio Logical Reasoning (ALR) task containing 6,446 text–audio CoT-annotated samples to enable complex reasoning over spoken content, and proposed SoundMind, a rule-based reinforcement learning algorithm that enhances deep cross-modal reasoning in audio-language models.
Large Audio-Language Model RL Reward Modelling Audio Reasoning

[NAACL 2025] Temporal Working Memory: Query-Guided Temporal Segment Refinement for Enhanced Multimodal Understanding
Guarini Graduate Student Travel Award, Dartmouth College
Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, Jiang Gui
Guarini School of Graduate and Advanced Studies Travel Award, Dartmouth College
Pdf | Github Code; ★ **Starred 300+** | Proposed Temporal Working Memory, a plug-and-play query-guided segment refinement module that maintains dynamic temporal memory to effectively preserve task-relevant video–audio segment and enhance long-range temporal reasoning, achieving consistent performance gains when integrated into nine recent state-of-the-art multimodal large language models (MLLMs) across AVQA, video captioning, and retrieval tasks.
MLLM Video Understanding Audio Modelling

[EMNLP 2024] Learning Musical Representations for Music Performance Question Answering
BMDS Travel Award, Dartmouth College
Xingjian Diao, Chunhui Zhang, Tingxuan Wu, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, Jiang Gui
Pdf | Github Code | Proposed a specialized framework for audio–visual modeling in music understanding, addressing underexplored multimodal interactions, distinctive musical characteristics, and temporal alignment, and introduced annotated rhythmic and source features, with the framework achieving state-of-the-art on Music-AVQA 1.0 and 2.0.
Multimodal QA Representation Learning Multimodal Alignment Audio Modelling

Selected Core-Contribution Publications

[EMNLP 2025 (Oral)] ProtoVQA: An 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 | Proposed ProtoVQA, an adaptable prototypical VQA framework that learns question-aware prototypes and uses spatially-constrained greedy matching to ground answers in semantically coherent image regions, unifying answering and grounding via a shared backbone and introducing the VLAS metric to quantify visual–linguistic alignment.
🏷 Multimodal QA 🏷 Interpretability

[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 | Proposed Sparsify, a sparse learning framework for Music Audio-Visual Question Answering that integrates Sparse Masking, Adaptive Sparse Merging, and Sparse Subset Selection to reduce multimodal redundancy, highlight task-critical tokens, and accelerate training convergence, achieving efficiency and accuracy gains across Music-AVQA benchmarks.
🏷 Multimodal QA 🏷 Sparsity Learning

[WACV 2025] FT2TF: First-Person Statement Text-To-Talking Face Generation
Xingjian Diao, Ming Cheng, Wayner Barrios, SouYoung Jin
Pdf | 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.
🏷 AIGC 🏷 Multimodal Alignment

[Preprint 2025] On The Design Choices of Next Level LLMs
Pdf | 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 🏷 Post-Training 🏷 Optimization 🏷 RL

[Preprint 2025] SPAN: Unlocking Pyramid Representations for Gigapixel Histopathological Images
Weiyi Wu, **Xingjian Diao**, Chongyang Gao, Xinwen Xu, Siting Li, Jiang Gui
Pdf | Introduced 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 | Presents the first comprehensive survey of Music Audio-Visual Question Answering, 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 QA 🏷 Video Understanding 🏷 Music Modelling

Selected Machine Learning Engineering 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:** ICLR, NeurIPS, CVPR, ICCV, ACL, ACL Industry Track, EMNLP, ACMMM, ACMMM Datasets Track, WACV, ICASSP, AISTATS, EACL, IUI, 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*