# Xingjian Diao

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

Linkedin | Github | Google Scholar

## Research Interests

My research focuses on **multimodal learning** 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 multimodal models that advance multimodal question answering, video understanding, and audio–visual reasoning across complex real-world scenarios and dynamic environments.
My GitHub repositories on multimodal large language models (MLLMs) have received 1.5k+ Stars ⭐ .

## Education

- **Dartmouth College** *Sep 2022 - Dec 2026 (Expected)*
  *Ph.D. candidate 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*
  *Applied Scientist Intern* Santa Cruz, USA

  **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.
  Computer Vision Ubiquitous Computing

  **IMU2Reason: A Multimodal LLM for Safety-Aware Activity Understanding**
  Pdf | Proposed MHIA-LM, a multimodal LLM trained on fused video and IMU representations, where spatiotemporal video embeddings from InternVideo2 and IMU motion features encoded by a PatchTST-based encoder are integrated through a multi-layer Gated Hierarchical Interaction module and mapped into the LLM token embedding space for instruction tuning, enabling fine-grained activity recognition and safety reasoning.
  MLLM Multimodal Reasoning

## Selected 1ST-Author Publications [Full List]

**[Preprint 2026] Addressing Overthinking in Large Vision-Language Models via Gated Perception-Reasoning Optimization**
**Xingjian Diao**, Zheyuan Liu, Chunhui Zhang, Weiyi Wu, Keyi Kong, Lin Shi, Kaize Ding, Soroush Vosoughi, Jiang Gui
Pdf | Introduced Gated Perception–Reasoning Optimization (GPRO), a token-level adaptive computation framework that leverages large-scale perception–reasoning failure attribution (≈790K samples) to train a meta-reasoning controller via multi-objective reinforcement learning, dynamically routing between fast execution, visual re-perception, and slow reasoning to mitigate overthinking while improving both accuracy and efficiency in vision–language models.
Large Vision-Language Model Overthinking Mitigation Reinforcement Learning

**[EMNLP 2025] SoundMind: RL-Incentivized Logic Reasoning for Audio-Language Models**
**Oral Presentation Award (top 4.35%)**, 30th EMNLP
**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 Logical Reasoning Reinforcement Learning

**[EMNLP 2025] ProtoVQA: An Adaptable Prototypical Framework for Explainable Fine-Grained Visual Question Answering**
**Oral Presentation Award (top 4.35%)**, 30th EMNLP
**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

**[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
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

**[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

**[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 | Proposed and developed 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

## Selected Collaborative Publications

**[EACL 2026] Tailoring Memory Granularity for Multi-Hop Reasoning over Long Contexts**
Peijun Qing, **Xingjian Diao**, Chiyu Ma, Saeed Hassanpour, Soroush Vosoughi
Pdf | Proposed Tailoring Memory Granularity, a reward-guided framework that adaptively composes hybrid memory across multiple granularities, including chunks, triples, atomic facts, and summaries, to enhance large language models in long-context multi-hop reasoning through dynamic query-specific memory selection.
LLM Hybrid Memory Multi-hop Reasoning

**[AACL 2025] Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge**
**Oral Presentation Award (top 3.30%)**, 14th IJCNLP-AACL
Lin Shi, Chiyu Ma, Wenhua Liang, **Xingjian Diao**, Weicheng Ma, Soroush Vosoughi
Pdf | Proposed a systematic position-bias evaluation framework for LLM-as-a-Judge that integrates Repetition Stability, Position Consistency, and Preference Fairness to measure bias robustness, expose primacy–recency behaviors under prompt permutations, and extend analysis from pairwise to list-wise comparisons on MTBench and DevBench.
LLM-as-a-Judge

**[EMNLP 2025] Knowing More, Acting Better: Hierarchical Representation for Embodied Decision-Making**
Chunhui Zhang, Zhongyu Ouyang, **Xingjian Diao**, Zheyuan Liu, Soroush Vosoughi
Pdf | Proposed a hierarchical action probing framework that aggregates layer-wise MLLM representations to enhance spatial grounding, contextual integration, and abstract reasoning for embodied decision-making, achieving substantial gains across language-guided rearrangement tasks.
MLLM Vision–Language–Action

**[EMNLP 2025] Assessing and Mitigating Medical Knowledge Drift and Conflicts in Large Language Models**
Weiyi Wu, Xinwen Xu, Chongyang Gao, **Xingjian Diao**, Siting Li, Lucas A Salas, Jiang Gui
Pdf | Proposed ConflictMedQA, a guideline-grounded benchmark pairing up-to-date and outdated clinical recommendations, along with ECDA/IKCR metrics to evaluate medical knowledge drift and assess RAG, DPO, and RoD

for resolving temporal conflicts in LLMs.
Medical LLM Knowledge Drift RAG

**[EMNLP 2024] AlphaLoRA: Assigning LoRA Experts Based on Layer Training Quality**
Peijun Qing, Chongyang Gao, Yefan Zhou, **Xingjian Diao**, Yaoqing Yang, Soroush Vosoughi
Pdf | Proposed AlphaLoRA, a training-free layer-wise expert allocation strategy for LoRA-MoE that leverages Heavy-Tailed Self-Regularization to quantify layer training quality and assign experts accordingly, reducing redundancy while improving performance across NLP and reasoning benchmarks.
LLM Mixture-of-Experts

**[EMBC 2024] GluMarker: A Novel Predictive Modeling of Glycemic Control Through Digital Biomarkers**
**EMBC NextGen Scholar Award**, IEEE 46$^{th}$ EMBC
Ziyi Zhou, Ming Cheng, **Xingjian Diao**, Yanjun Cui, Xiangling Li
Pdf | Proposed GluMarker, a digital biomarker framework that integrates broader daily factors (meals, insulin doses, and CGM-derived metrics) to predict next-day glycemic control and identify key digital biomarkers that reveal how everyday behaviors shape diabetes management.
Digital Biomarkers

**[Preprint 2025] On The Design Choices of Next Level LLMs**
Yijun Tian, **Xingjian Diao**, Ming Cheng, Chunhui Zhang, Jiang Gui, Soroush Vosoughi, Xiangliang Zhang, Nitesh V. Chawla, Shichao Pei
Pdf | Provided 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.
LLM Post-Training Reinforcement Learning

**[Preprint 2025] What Makes a Good Curriculum? Disentangling the Effects of Data Ordering on LLM Mathematical Reasoning**
Yaning Jia, Chunhui Zhang, **Xingjian Diao**, Xiangchi Yuan, Zhongyu Ouyang, Chiyu Ma, Soroush Vosoughi
Pdf | Proposed a unified offline curriculum learning framework that systematically disentangles five curriculum dimensions and isolates the causal impact of data ordering to provide principled guidance on curriculum design.
LLM Curriculum Learning

**[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 gigapixel whole-slide pathology, combining spatial-adaptive condensation with context-aware refinement to preserve spatial structure and enable efficient multi-scale tumor detection, classification, and segmentation.
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 | Presented the first survey of Music Audio-Visual Question Answering, providing a systematic analysis of how specialized multimodal architectures with spatio-temporal modeling enable reliable reasoning over musical performances.
Multimodal QA Video Understanding

## Teachings

**Graduate Teaching Assistant**

- COSC89/189 (Video Understanding), Dartmouth College, Spring 2024
- COSC74/274 (Machine Learning), Dartmouth College, Winter 2024
- COSC61 (Database Systems), Dartmouth College, Summer 2023
- COSC10 (Object Oriented Programming), Dartmouth College, Spring 2023
- COSC62/162 (Applied Cryptography), Dartmouth College, Winter 2023
- COSC10 (Object Oriented Programming), Dartmouth College, Fall 2022

## Services

**Reviewer / Program Committee Member**

- ICML (International Conference on Machine Learning) 2026
- NeurIPS (Annual Conference on Neural Information Processing Systems) 2025
- ICLR (International Conference on Learning Representations) 2025, 2026
- CVPR (Conference on Computer Vision and Pattern Recognition) 2025, 2026
- ICCV (International Conference on Computer Vision) 2025
- ACL (Annual Meeting of the Association for Computational Linguistics) 2025
- ACL (ACL Industry Track) 2025
- EMNLP (Empirical Methods in Natural Language Processing) 2025
- ACMMM (ACM International Conference on Multimedia) 2025
- ACMMM (Datasets Track) 2025

- TMLR (Transactions on Machine Learning Research) 2025
- IMWUT (Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies) 2026
- WACV (IEEE/CVF Winter Conference on Applications of Computer Vision) 2025, 2026
- AISTATS (International Conference on Artificial Intelligence and Statistics) 2026
- EACL (European Chapter of the Association for Computational Linguistics) 2026
- ICASSP (International Conference on Acoustics, Speech & Signal Processing) 2025, 2026
- IUI (ACM International Conference on Intelligent User Interfaces) 2026
- ISBI (IEEE International Symposium on Biomedical Imaging) 2025, 2026
- IJCNN (International Joint Conference on Neural Networks) 2024, 2025
- ICME (IEEE International Conference on Multimedia & Expo) 2024, 2025, 2026

## AWARDS

- **Dartmouth Fellowship,** Dartmouth College. *2022-Present*
- **EMNLP Oral Presentation Award (top 4.35%),** $30^{\text{th}}$ EMNLP. *2025*
- **IJCNLP-AACL Oral Presentation Award (top 3.30%),** $14^{\text{th}}$ IJCNLP-AACL. *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*