XINYU LI

7324066689 ◇ lixinyu.arthur@outlook.com ◇ https://www.arthurlxy.com

EXPERIENCE

AGI, Amazon 2022 – Present
Principal Applied Scientist *Bellevue, WA*

- Designed and implemented Amazon's in-house M-LLM training and ablation pipeline in native PyTorch, enabling end-to-end encoder-LLM joint optimization with full 3D parallelism. The framework was adopted across AGI teams for encoder pre-training, ablations, and research exploration.
- Led the design and pre-training of multi-scale visual encoders powering the Nova model family and AWS RAG systems serving core services such as S3 and Amazon photos. Achieved state-of-the-art performance with 40% token reduction compared to the previous generation.
- Directed video and multimodal understanding initiatives, designing and implementing the multimodal sequence construction, dataloader, and modeling pipelines spanning data, pre-training, and post-training workflows. Nova models with state-of-the-art video understanding performance are now deployed to customers in media, entertainment, and surveillance.
- Spearheaded next-generation Nova research on a joint understanding–and-generation framework and long video reasoning, with outcomes published at CVPR, ICCV, ECCV, NeurIPS, and ICLR.

AML, Tiktok/ByteDance 2021 - 2022
Staff Research Scientist *Bellevue, WA*

- Led development of ByteDance's large multimodal model integrating vision, audio, and language understanding, powering TikTok's compliance, search, and recommendation systems.
- Developed ModelZoo, a unified framework for implementing and benchmarking state-of-the-art vision and multimodal architectures, accelerating research and developing cycles.
- Conducted research in video and multimodal understanding, resulting in publications at top-tier venues such as CVPR, ICASSP, and ICLR.

AWS AI, Amazon 2018 - 2021
Senior Applied Scientist *Seattle, WA*

- Led video understanding research at AWS AI, delivering production models for compliance, tracking, and ad insertion within Rekognition Video and related AWS services, used by Prime Video and Amazon Stores.
- Published in CVPR, ICCV, ECCV, and NeurIPS; contributed to open-source frameworks including GluonCV-Torch, GluonMM, and Decord.

EDUCATION

Ph.D. in Computer Engineering, Rutgers University *Sep 2013 - Feb 2018*
B.S. in CE, Uni. of Electronic Science and Technology of China *Sep 2009 - June 2013*

SELECTED PUBLICATION

Selected publications in past 5 years, full list of publication can be find at Google Scholar
* denotes equally contributed.

1. Wenjin Zhang, Xinyu Li, Chenyang Gao, Ivan Marsic. " SemiVisBooster: Boosting Semi-Supervised Learning for Fine-Grained Classification through Pseudo-Label Semantic Guidance" ICCV 2025.
2. Core Contributor. "The Amazon Nova family of models: Technical report and model card."

3. Core Contributor. “Amazon Nova Premier: Technical report and model card.”
4. Xianhang Li, Peng Wang, Xinyu Li, Heng Wang, Hongru Zhu, and Cihang Xie. “Efficient VideoMAE via Temporal Progressive Training.” CVPR 2025 workshop.
5. Yicheng Wang, Zhikang Zhang, Jue Wang, David Fan, Zhenlin Xu, Linda Liu, Xiang Hao, Vimal Bhat, Xinyu Li. “GEXIA: Granularity Expansion and Iterative Approximation for Scalable Multi-grained Video-language Learning.” WACV 2025.
6. Seon-Ho Lee, Jue Wang, David Fan, Zhikang Zhang, Linda Liu, Xiang Hao, Vimal Bhat, Xinyu Li. “Now You See Me: Context-Aware Automatic Audio Description.” WACV 2025.
7. Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, Xinyu Li. “Video Token Merging for Long Video Understanding.” NeurIPS 2024.
8. David Fan, Jue Wang, Shuai Liao, Zhikang Zhang, Vimal Bhat, Xinyu Li. “Text-Guided Video Masked Autoencoder.” ECCV 2024.
9. David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, Xinyu Li. “Motion-Guided Masking for Spatiotemporal Representation Learning.” ICCV 2023.
10. Najmeh Sadoughi, Xinyu Li, Avijit Vajpayee, David Fan, Bing Shuai, Hector Santos-Villalobos, Vimal Bhat, Rohith MV. “MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation.” ICCV 2023.
11. Chen, Yuxiao, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N. Metaxas, and Hongxia Yang. “Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens.” CVPR 2023.
12. Xiaoyu Liu, Hanlin Lu, Jianbo Yuan, and Xinyu Li. “CAT: Causal Audio Transformer for Audio Classification.” ICASSP 2023.
13. Jiaojiao Zhao*, Yanyi Zhang*, Xinyu Li*, Hao Chen, Shuai Bing, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G.M. Snoek, Joseph Tighe. “TubeR: Tubelet Transformer for Video Action Detection.” CVPR 2022 Oral.
14. A S M Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, Davide Modolo. “What to look at and where: Semantic and Spatial Refined Transformer for detecting human-object interactions.” CVPR 2022 Oral.
15. Feng Cheng, Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Li, Wei Xia. “Stochastic Backpropagation: A Memory Efficient Strategy for Training Video Models.” CVPR 2022 Oral.
16. Bing Shuai, Xinyu Li, Kaustav Kundu, Joseph Tighe. “Id-Free Person Similarity Learning.” CVPR 2022.
17. Xinyu Li*, Yanyi Zhang*, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. ”VidTr: Video Transformer Without Convolutions.” ICCV 2021.
18. Xinyu Li*, Chunhui Liu*, Hao Chen, and Joseph Tighe “Selective Feature Compression for Efficient Activity Recognition Inference.” ICCV 2021.
19. Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu and Stefano Soatto. ”Long Short-Term Transformer for Online Action Detection.” NeurIPS 2021 (Spotlight).
20. Zhang, Yanyi, Xinyu Li, and Ivan Marsic. “Multi-Label Activity Recognition using Activity-specific Features.” CVPR 2021.
21. Shuai, Bing, Andrew G. Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. “Multi-object tracking with Siamese track-RCNN.” CVPR 2021.

22. Zhu, Yi, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. “A Comprehensive Study of Deep Video Action Recognition.” arXiv preprint arXiv:2012.06567 (2020).

23. Li, Xinyu, Bing Shuai, and Joseph Tighe. “Directional temporal modeling for action recognition.” In European Conference on Computer Vision, pp. 275-291. Springer, Cham, ECCV 2020.

SELECTED OPEN-SOURCE PACKAGES

GluonCV-Torch Project Link

- GluonCV, with 5.9K stars, provides implementations of state-of-the-art (SOTA) deep learning algorithms in computer vision. Available in mxnet and pytorch.

GluonMM Project Link

- GluonCV-Transformer is an open-sourced library based on PyTorch, providing a list of SOTA transformer-based research implementations on various image tasks (image classification, object detection, semantic segmentation), video tasks (video classification, spatio-temporal action detection, long-video reasoning), and multimedia tasks (sound classification, video to text, retrieval, etc.).

TubeR: Tubelet Transformer Project Link

- This repo contains the supported code to reproduce spatio-temporal action detection results of TubeR: Tubelet Transformer for Video Action Detection.

PROFESSIONAL SERVICES

Reviewer: CVPR, ICCV, ECCV, NeurIPS, ICLR, WACV, ICASSP, Interspeech
Journal Reviewer: Pattern Recognition Letters, IMWUT, IEEE TMC