Zhaoyang Xia

jeffery9707.github.io zx149@rutgers.edu 908-217-3973

Intern Availability: Start: May 5th, End: September 9th

RESEARCH INTERESTS

- **Diffusion Models & Video Editing** (e.g., Diffusion models for video transformation, motion-guided image reenactment, T2I diffusion model optimization)
- **Multi-modal Large Language Models** (e.g., MLLMs applications)
- **Human Action Modeling** (e.g., Human action detection & recognition)

EDUCATION

Rutgers University Sept. 2021 – Expected May. 2026

- Ph.D. in Computer Science, GPA: 4.0/4.0
- Advisor: Dimitris Metaxas

Rutgers University Sept. 2019 – May. 2021

- M.S. in Computer Science (Data Science), GPA: 3.91/4.0

Fudan University Sept. 2015 – Jun. 2019

- B.S. in Information and Computing Science (Data Science & Technology)

SELECTED PUBLICATIONS

Multi-modal Large Language Models

Xia, Zhaoyang. VISIAR: Empower MLLM for Visual Story Ideation. 2024 [Submitted to ACL'25]

Diffusion Models & Video Editing

Xia, Zhaoyang, Yang Zhou, Ligong Han, Carol Neidle, and Dimitris N Metaxas. Diffusion models for sign language video anonymization. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources (**LREC**)*, pages 395–407, 2024 [PDF] [Demo]

Xia, Zhaoyang, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitri Metaxas. Sign Language Video Anonymization. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources (**LREC**)*, pages 202–211, 2022 [PDF] [Demo]

Sooyeon Lee, Abraham Glasser, Becca Dingman, **Xia, Zhaoyang**, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (**ASSET**)*, pages 1–13, 2021 [PDF] [Demo]

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, ... **Xia, Zhaoyang**, Akash Srivastava, and Dimitris N Metaxas. Improving Negative-Prompt Inversion via Proximal Guidance. In *IEEE Winter Conference on Applications of Computer Vision (**WACV**)*, 2024 [PDF]

Human Action Modeling

Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, **Xia, Zhaoyang**, Shijie Geng, Ligong Han, and Dimitris N Metaxas. Hierarchically Self-supervised Transformer for Human Skeleton Representation Learning. In *European Conference on Computer Vision (**ECCV**)*, pages 185–202. Springer, 2022 [PDF]

Yang Zhou, **Xia, Zhaoyang**, Yuxiao Chen, Carol Neidle, and Dimitris Metaxas. A multimodal spatio-temporal gcn model with enhancements for isolated sign recognition. In *Proceedings of the {**LREC**-COLING} 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*. ELRA Language Resources Association (ELRA) and the International Committee ..., 2024 [PDF]

Medical Imaging

Qilong Zhangli, Jingru Yi, Di Liu, Xiaoxiao He, **Xia, Zhaoyang**, Qi Chang, Ligong Han, Yunhe Gao, Song Wen, Haiming Tang, et al. Region proposal rectification towards robust instance segmentation of biological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*, pages 129–139. Springer, 2022 [PDF]

Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, **Xia, Zhaoyang**, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*, pages 485–495. Springer, 2022 [PDF]

SELECTED RESEARCH EXPERIENCE

VISIAR: Empower MLLM for Visual Story Ideation [In Submission]
Research Scientist/Engineer Intern, Adobe Inc.

- Proposed a novel task: Visual Story Ideation, which aims to select and rearrange videos for potential storyline generation from a collection of assets.
- Proposed VISIAR, a framework leveraging ***Multi-modal Large Language Models*** enhanced by Graph Clustering method through novel ideation ***Graph Construction***.
- Collected new dataset and built benchmark. Surpass GPT4o by 25 % in user study.

Diffusion models & Video Editing
PhD Student, Rutgers University

- ***Diffusion models for Sign Language video anonymization*** [PDF] [Demo]
  - Proposed zero-shot text-guided sign language anonymization, which alters the signer's identity through text-guided video editing. Designed methods for accurate gestures and facial expression transferring for ASL videos with ***Stable Diffusion*** and the ***Image Animation*** module.
  - Applied cross-frame attention mechanism and optical flow guided latent fusion method with ControlNet for consistent video editing.
- ***Sign Language Video Anonymization*** [PDF] [Demo]
  - Developed a motion-based **Image Animation** model for sign language video anonymization, generating high-resolution videos with altered signer identities while preserving essential motions and facial expressions.
  - Designed an asymmetric encoder–decoder image generator for high-resolution outputs. Designed loss to improve hands and face generation through bounding boxes.
- ***American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard-of-Hearing Users*** [PDF] [Demo]
  - Applied First Order Motion Model for ***Face Swap*** to automatically disguise the face in sign language videos while preserving essential facial expressions and natural human appearance.
  - Incorporated segmentation model and color-based method for ***Skin Segmentation*** to enhance anonymization.

Diffusion Models & Image Generation
Research Scientist/Engineer Intern, Adobe Inc. & PhD Student, Rutgers University

- ***Improving Diffusion Models with Human Preference***

  - Designed algorithm for improving diffusion model with ***Human Preference*** data.
  - Enhanced image generation quality by utilizing representations from the UNet for guidance during inference stage, focusing on aligning the output with human aesthetic preferences.

- ***Improving Tuning-free Real Image Editing with Proximal Guidance*** [PDF]
  - Improved the DDIM inversion ability for diffusion models.
  - Developed a regularization term introduced by the proximal function to reduce artifacts. Proposed inversion guidance using one-step gradient descent to enhance editing quality.

Human Action Understanding
PhD Student, Rutgers University

- ***Hierarchically Self-supervised Human Skeleton Representation Learning*** [PDF]
  - Develop Hierarchically Self-supervised Transformers for Human Skeleton Representation Learning using various tasks on frame, clip, and video levels.
  - Improved the downstream tasks such as human action recognition and detection.
- ***Multi-modal Spatio-temporal Sign Recognition*** [PDF]
  - Proposed a multi-modal network using skeletons and handshapes as input to recognize individual signs and detect their boundaries in American Sign Language (ASL) videos

Medical Imaging
PhD Student, Rutgers University

- ***Instance Segmentation*** [PDF]
  - Designed models to detect grammar markers in videos containing continuing sentences of ASL. Applied Inflated Inception-v1 model with sliding window method to detect facial actions.
- ***Multi-view Image Segmentation*** [PDF]
  - Proposed a multi-modal network using skeletons and handshapes as input to recognize individual signs and detect their boundaries in American Sign Language (ASL) videos

Explainable Recommendation System for Movies
Researcher, Computer Science Department of Fudan University

- Extracted features as tag preference and tag relevance from movie data.
- Utilized Explicit Factor Model based on features and rates of movies to do ***Explainable Recommendations***.

WORK HISTORY

Research Scientist/Engineer Intern— Adobe Inc. May.2024 – Aug.2024

- Designed novel generative AI for visual storytelling methods.

Research Scientist/Engineer Intern — Adobe Inc. May.2023 – Dec.2023

- Designed algorithms for improving diffusion models with human preference data

Intern— MISUMI (China) precision machinery trading co., LTD Nov.2018 – Feb.2019

- Predicted the loss of customers by applying LSTM neural network on customers' behavior data.
- Analysed the customers' chat context by applying the textrank model, word2vec embedding, and k-means clustering.

SKILLS

Programming Languages: Python, SQL, R
Frameworks: PyTorch, OpenCV
Academic Service: Reviewer for ECCV, CVPR