

Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Swetha Sirnam

MS by Research, IIIT-Hyderabad, 2018

B.Tech., IIIT-Hyderabad, 2016

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

2 April 2026, 11:00 A.M.

SPRK (UCF Global) Building - Room 229

DISSERTATION COMMITTEE

Professor Mubarak Shah, *Chair*, *shah@crcv.ucf.edu*

Professor Yogesh Rawat, *yogesh@crcv.ucf.edu*

Professor Sernam Lim, *sernam@ucf.edu*

Professor Pradipkumar Ramanlal, *pramanlal@bus.ucf.edu*

DISSERTATION RESEARCH IMPACT

Multimodal intelligence is becoming a central capability for modern AI systems that must interpret information from images, videos, audio, and text. Applications such as digital assistants, visual search, robotics, surveillance, and content recommendation heavily rely on models that can understand visual environments and connect them with natural language. Despite recent progress, enabling reliable multimodal understanding remains challenging. Current systems often struggle to capture the structured complexity of real world visual data, reason about interactions and events unfolding over time, and adapt to the diverse ways humans express intent. Further, existing evaluation protocols frequently provide only a limited view of model capabilities, making it difficult to assess whether systems truly understand visual scenes.

This dissertation advances more reliable and scalable multimodal intelligence by developing methods that better preserve visual information, maintain structure across modalities, and enable richer reasoning over visual data. It introduces learning frameworks that reduce dependence on costly annotations, strengthen temporal understanding in long videos, improve alignment with human intent, and establish more realistic benchmarks for evaluating reasoning capabilities. Together, these contributions move systems toward more robust, and dependable understanding of complex real-world environments.

SELECTED PUBLICATIONS

1. **Unsupervised Discriminative Embedding for Action Learning in Complex Activities** *Sirnam Swetha*, Hilde Kuehne, Yogesh S Rawat, and Mubarak Shah. In IEEE International Conference on Image Processing (ICIP), 2021.
2. **Multi-SK: Preserving Modality Structure Improves Multi-Modal Learning.** *Sirnam Swetha*, Mamshad Nayeem Rizve, Nina Shvetsova, Hilde Kuehne, and Mubarak Shah. In International Conference on Computer Vision (ICCV), 2023.
3. **X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs.** *Sirnam Swetha*, Jinyu Yang, Tal Neiman, Mamshad Nayeem Rizve, Son Tran, Benjamin Yao, Trishul Chilimbi and Mubarak Shah. In European Conference on Computer Vision (ECCV), 2024.
4. **SMPRO: Self-Supervised Visual Preference Alignment via Differentiable Multi-Preference Multi-Group Ranking.** *Sirnam Swetha*, Rui Meng, Shwetha Ram, Tal Neiman, Son Tran, and Mubarak Shah. In Proceedings of the AAAI Conference on Artificial Intelligence, 2026.
5. **VRR-QA: Visual Relational Reasoning in Videos Beyond Explicit Cues.** *Sirnam Swetha*, Rohit Gupta, Parth Parag Kulkarni, David G Shatwell, Jeffrey A Chan Santiago, Nyle Siddiqui, Joseph Fioresiand Mubarak Shah. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2026.
6. **Safe-LLaVA: A Privacy-Preserving Vision Language Dataset and Benchmark for Biometric Safety.** Younggun Kim*, *Sirnam Swetha**, Fazil Kagdi, and Mubarak Shah. (* equal contribution). In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2026.
7. **GT-Loc: Unifying When and Where in Images Through a Joint Embedding Space.** David G Shatwell, Ishan Rajendrakumar Dave, *Sirnam Swetha*, Mubarak Shah. In International Conference on Computer Vision (ICCV), 2025.
8. **TIGeR: A Unified Framework for Time, Images and Geo-location Retrieval.** David G Shatwell, *Sirnam Swetha* and Mubarak Shah. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2026.

DISSERTATION

ADVANCING MULTI-MODAL LEARNING

Modern AI systems increasingly rely on multimodal learning to interpret information from multiple sources such as images, videos, text, and audio. These capabilities are essential for applications including visual search, digital assistants, visual question answer, and robotics. While recent multimodal large language models (MLLMs) have demonstrated impressive performance through large-scale pretraining, they still face key limitations when applied to real-world visual data. In particular, fine-grained visual details are often lost during multi-modal alignment, models frequently struggle to capture subtle human intent, and existing approaches remain limited in their ability to reason over long videos where events unfold over time. Moreover, many current evaluation benchmarks rely heavily on explicit cues, making it difficult to assess deeper relational reasoning in visual environments. This dissertation develops multimodal learning methods that address these challenges by improving visual perception, alignment with human preferences, temporal understanding in videos, structure-aware representation learning, and probing implicit visual reasoning.

In this dissertation, first we introduce X-Former, a dual-attention fusion module designed to improve fine-grained visual perception in MLLMs. By integrating signals from frozen encoders that capture both global context and local visual detail, X-Former produces detail-aware visual tokens that preserve spatial structure while retaining cross-modal alignment. The unified contrastive-reconstruction learning objective allows the model to learn from image-text pairs without curated instruction-tuning data, improving fine-grained performance on tasks such as counting, spatial reasoning, and relational understanding.

While X-Former enhances the fine-grained perception capabilities of MLLMs, effective deployment also requires models to understand user intent and generate helpful response. For this, we propose SMPRO, a self-supervised visual preference alignment framework that moves beyond traditional binary preference supervision. Also, the first self-supervised multi-preference alignment framework capable of handling ambiguous visual question answers beyond purely logical queries. SMPRO generates multi-preference signals and optimizes them through differentiable ranking, modeling multiple positive and negative preferences simultaneously without human annotation or external expert models.

With perception and alignment strengthened for images, we next turn to more challenging domain of video understanding, where many real-world applications require reasoning over long videos where actions unfold over time. Next, we develop UDL, an unsupervised framework for discovering temporal structure in videos through a Discriminative Latent Concept Learning objective. UDL learns visual-temporal embeddings that jointly enforces reconstruction and discriminative separation of latent concepts, enabling the discovery of fine-grained sub-actions without requiring manual annotations.

In addition to temporal modeling, effective multimodal learning also requires preserving the internal structure of each modality during alignment. We introduce Multi-SK, a semantic structure-preserving alignment framework based on our Multi- Assignment Sinkhorn-Knopp algorithm. Unlike traditional contrastive objectives that can collapse modality-specific topology, Multi-SK maintains intra-modal relationships while learning shared multi-modal representations. This improves out-of-domain generalization.

Finally, evaluating whether multimodal systems truly reason about visual events remains an open challenge. To address this, the dissertation introduces VRR-QA, a benchmark for visual relational reasoning in videos that emphasizes implicit relationships and contextual dependencies between events. Unlike traditional benchmarks focused on explicit recognition, VRR-QA evaluates implicit reasoning about interactions, spatio-temporal dependencies in videos. Empirical results reveal that current multimodal models struggle with these tasks, high-lighting important gaps in their reasoning capabilities and motivating future research in robust multimodal understanding. Together, these contributions advance multimodal learning by improving perception, alignment, temporal structure, representation learning, and evaluation. The proposed methods move multimodal systems toward more structured, interpretable, and human-aligned understanding of real-world environments.



Swetha Sirnam

2012-2016	B.Tech., IIIT-Hyderabad.
2016-2018	MS by Research, IIIT-Hyderabad
2018-2019	Goldman Sachs
2022	Research Intern, Amazon Prime Video
2023	Research Intern, Amazon Search Science and AI
2024	Research Intern, Amazon Stores Foundational AI
2025 -	Applied Scientist II, Amazon AWS