UCF | Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Shehreen Azad

M.S., University of Central Florida, 2026
B.S., Islamic University of Technology, Bangladesh, 2022

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

23 October 2026, 11 A.M.
SPRK (Formerly UCF Global), Room 229

DISSERTATION COMMITTEE

Professor Yogesh S Rawat, Chair	*yogesh@crcv.ucf.edu*
Professor Mubarak Shah	*shah@crcv.ucf.edu*
Professor Chen Chen	*chen.chen@crcv.ucf.edu*
Professor Soheil Sabri	*soheil.sabri@ucf.edu*

DISSERTATION RESEARCH IMPACT

Every day, cameras capture an enormous amount of video, from sports broadcasts and movies to security footage, hospital monitoring, traffic cameras, and personal recordings. Far more video is recorded than any person could ever watch, so we increasingly rely on AI systems to watch it for us: to find the moment a game was won, alert a caregiver when a patient falls, help a driver react to the road, or answer questions about what happened in a recording. Large multimodal models that combine vision and language have made this possible, but they still fall short of how people understand video. A person can follow an hour-long story and remember who did what, yet these models can only take in a limited amount of video at once and lose important details when longer videos are cut down to fit. A person watching a live event knows when they have seen enough to answer a question, while these models either guess too early or respond too late, which matters when a timely answer can prevent an accident or catch a critical event. And a person can often recognize someone they know from the way they move, even without seeing their face, whereas current systems rely on faces, clothing, or walking patterns that fail when a face is hidden, clothing changes, or someone is simply going about their day. Thus, closing these gaps is essential for building video systems that people can trust in the real world.

This dissertation addresses these gaps across three questions: what is happening, when to answer, and who is involved. We first present a method that processes long videos piece by piece, focusing on details relevant to the task while remembering both recent and earlier events, so it can handle videos of any length at a fixed cost. We then introduce a way for models watching live video to judge when they have seen enough evidence to answer, along with a new score and benchmark that reward answers that are both correct and on time. Finally, we introduce the task of recognizing people from how they perform daily activities, and present methods that separate a person's identifying traits from their appearance and movement, making identification more reliable in real-world settings. Together, these contributions move video models from passively watching toward actively reasoning about what they see, when to speak, and whom to recognize.

SELECTED PUBLICATIONS

1. **StreamReady: Learning *What* to Answer and *When* in Long Streaming Videos.** S Azad, V Vineet, and Y Rawat. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.
2. **HierarQ: Task-Aware Hierarchical QFormer for Enhanced Video Understanding.** S Azad, V Vineet, and Y Rawat. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
3. **DisenQ: Disentangling QFormer for Activity Biometrics.** S Azad, and Y Rawat. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025. (Highlight)
4. **Activity Biometrics: Person Identification from Daily Activities.** S Azad, and Y Rawat. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

DISSERTATION

FROM PERCEPTION TO REASONING: TOWARDS ADAPTIVE VIDEO INTELLIGENCE

Multimodal large language models, which combine vision and language, are now the starting point for most video understanding systems, powering applications such as video search, question answering, live monitoring, and assistive agents. Although these models perform well on images and short clips, they fall short on real-world video, which is long, arrives continuously, and involves people doing a wide range of activities. To understand such video, a system must answer three questions: what is happening, when there is enough evidence to answer, and who is involved. Current models struggle with all three. They lose fine-grained detail on long videos, ignore when an answer should be given, and identify people using cues that fail outside controlled settings. This dissertation argues that reliable video understanding depends on reasoning that adapts to the task, to the evidence available at each moment, and to the people in the scene, rather than on processing more input alone.

Our first contribution is HierarQ, a framework for understanding what is happening in long videos. Existing methods fit long videos into the model's limited context by sampling frames or compressing them, which discards important details and treats every frame as equally relevant. HierarQ instead processes the video frame by frame and uses the question to decide what to focus on: one stream captures fine-grained details about objects over short spans, while another tracks how the scene evolves over longer periods, each supported by its own short- or long-term memory. This keeps compute cost constant regardless of video length, and HierarQ achieves state-of-the-art results across ten benchmarks, showing that task awareness, not greater input capacity, is key to understanding long videos.

While HierarQ understands long videos well, it assumes the full video is available and that the question is known in advance. In live settings, frames arrive continuously and a question may be asked before the evidence needed to answer it has appeared. Next, we introduce StreamReady, which learns not only what to answer but when. A memory tree summarizes the incoming stream without knowing future questions, and a lightweight learned readiness signal lets the model judge from its own memory whether it has seen enough to respond. To measure this, we propose the Answer Readiness Score, which penalizes answers that come too early or too late, and ProReady-QA, an hour-long video benchmark with annotated evidence windows. StreamReady improves both accuracy and timing across nine benchmarks, establishing answer timing as a goal that can be learned directly rather than handled by external rules.

The methods above reason about events, but understanding a video also requires knowing who is involved. Existing person identification relies on faces, clothing, or walking patterns, which fail when faces are hidden, clothing changes, or people are doing everyday activities rather than walking. Finally, we introduce activity-biometrics, the task of identifying people from how they perform daily activities. ABNet establishes this task by separating identity from appearance through distillation from a silhouette-based teacher, and DisenQ removes the need for silhouettes by using language descriptions to separate identity, appearance, and motion, giving more weight to the activities that best reveal identity. These methods outperform prior approaches and generalize to traditional video-based identification.

Together, these contributions give one account of adaptive video intelligence: HierarQ adapts what a model attends to, StreamReady adapts when it responds, and activity-biometrics adapts how it recognizes people, moving video models from passively watching toward reasoning about what they see, when to speak, and whom to recognize.



Shehreen Azad

1999	Born in Dhaka, Bangladesh
2018-2022	B.S., Islamic University of Technology, Bangladesh
2022-2023	M.S., University of Central Florida
2023-2026	PhD, University of Central Florida
2026	Applied Scientist Intern, Ring, Amazon
2026 -	Applied Scientist II, Ring, Amazon