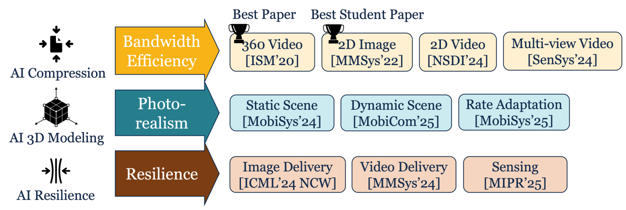

Figure 1: AI-System co-design for immersive computing.

Immersive computing is a technology that build end-to-end systems with networking, virtual reality (VR), augmented reality (AR), multimedia, and cloud computing technologies to make users feel immersed in various scenarios. Its application encompasses image, 2D video, and volumetric video streaming, and it shows great potential in the field of education, healthcare, and industrial production. The goal of immersive computing is to let users experience the virtual world just as the real world, where low latency, photo-realism, and reliability matter. However, it is challenging for existing immersive computing systems to achieve this goal, and the fundamental reasons are that they rely on handcrafted modules, explicit content representation, and explicit redundancy mechanisms. To overcome these obstacles, my research advocates AI-system co-design, an approach with two core principles: ❶ **AI-for-Immersion Insight**: ***Identifying the limitations of traditional immersive computing systems and employing advanced Artificial Intelligence (AI) techniques to tackle them.*** ❷ **AI-System Joint Design**: ***Jointly optimizing system designs and AI algorithms to meet practical system requirements (e.g., frame rate and latency).*** Guided by these principles, my research advances current immersive computing systems in terms of Bandwidth Efficiency, Photo-realism, and Reliability.

Research Impact: My research has culminated in over 20 publications, including in premier conferences such as USENIX NSDI, ACM MobiCom, ACM MobiSys, and ACM SenSys. It has been further distinguished by accolades such as the Best Student Paper Award at ACM MMSys, the Best Paper Award at IEEE ISM, one US Patent, and media recognition from the Siebel School News [19]. Beyond these achievements, my research has fostered extensive collaborations with academic institutions (*e.g.,* Princeton, UMASS, and GMU) and non-academic institutions (*e.g.,* US Army Research Laboratory, Illinois Fire Service Institute, and Boeing).

1. Towards Efficiency with AI-based Compression

Immersive computing requires connecting clients to vast virtual worlds at city, country, or global scales. While current technologies can handle smaller-scale data (*e.g.,* texts, images, 2D/3D videos), they fall short when faced with the immense data volumes over consumer-grade networks. AI-based compression, which employs deep neural networks for end-to-end encoding and decoding, offers better bandwidth efficiency but faces challenges in practical deployment. My research bridges this gap by addressing key system challenges:

▶**1.1 High Frame Rate.** AI-based video compression often processes frames sequentially, maximizing bandwidth efficiency but impacting frame rate. My work restructures frame dependencies into a binary tree, enabling parallel processing at the same tree level and improving the Quality of Experience (QoE) of state-of-the-art methods by up to 23.8% [12].

▶**1.2 Scalability and Generalizability.** AI-based compression systems often rely on symmetric, centralized, and static designs, which struggle with the demands of multi-camera systems. My research addresses these issues through: ❶ ***Asymmetric codec design:*** Distributed encoders and a centralized decoder based on the distributed coding theorem scale to any number of cameras, achieving over 64% bandwidth savings [4]. ❷ ***Decoupled training:*** Splitting training into intra-view (offline) and inter-view phases allows rapid adaptation to new scenes (*i.e.,* roughly 10 minutes) without extensive retraining.

▶**1.3 Ultra-low Bandwidth.** Under extremely low bandwidth, traditional compression methods produce artifacts (*e.g.,* blurriness and blockiness), hindering visual communication. My research addresses this with: ❶ ***Semantic compression:*** AI-based methods extract high-level image information during encoding and use it to generate smooth, visually coherent images during decoding, effectively conveying information [21]. ❷ ***Context-aware adaptation:*** Leveraging underwater light absorption properties, my work reduces transmission payloads by 7× by pruning the compression algorithm's codebook.

▶**1.4 Other Research on Efficiency.** Beyond AI-based compression, I have addressed the bandwidth and energy constraints of traditional methods: ❶ ***Compression efficiency:*** Profiling for visual analytics tasks identifies the complex correlation between compression parameters (*e.g.,* quantization settings) and performance, achieving up to 40% bandwidth reduction and a 2× speedup in computation offloading [2, 11]. This work earned the Best Student Paper Award at ACM MMSys 2022 and was featured in Siebel School

News [19]. Follow-up work with neural image codecs achieved an additional 20% reduction [9]. ❷ ***Energy efficiency:*** Video analytics is energy-intensive due to high frame rates and inefficient GPU utilization. My research addresses this through (i) frame filtering, reducing redundancy in 2D and 360-degree video analytics [7, 15, 14], and (ii) Dynamic Voltage and Frequency Scaling (DVFS), optimizing CPU and GPU energy use for edge [1] and cloud deployments [13].

2. Advancing Immersiveness with AI-based 3D Modeling

The virtual world involved with immersive computing is built through 3D modeling. Traditional approaches like point clouds and meshes lack the immersive quality needed due to their discrete nature. Recently, AI-based 3D modeling (*i.e.,* representing 3D scenes with continuous and learned functions) has shown promise but struggles to deliver satisfactory QoE. My research addresses these challenges with the following techniques:

▶**2.1 Bandwidth Efficiency.** Emerging AI-based 3D modeling such as Neural Radiance Fields (NeRFs) involves multiple modalities such as neural networks, feature values, and polygon meshes. The complex relationship between these modalities and the implicit nature of feature values make them bandwidth-consuming in network delivery. My research reduces bandwidth demands through: ❶ ***Profiling:*** Modeling the complex relationships between compression configurations across modalities, reducing latency in serving neural representations by up to 66% without compromising quality [10]. ❷ ***Pruning:*** Quantifying the importance of feature values during rendering and pruning less critical ones, achieving up to 68.3% bandwidth savings [24]. ❸ ***Adaptation:*** Hybridizing compact, view-dependent NeRF data with less-compact, coarse-grained complementary data to ensure consistent quality for diverse viewing patterns, improving user Quality of Experience (QoE) by up to 41.2% [18].

▶**2.2 Computation Efficiency.** Rendering AI-based 3D models is computation-intensive, posing challenges for resource-constrained clients like wearable VR headsets. My research addresses this through: ❶ ***Caching:*** Accelerating neural rendering by caching and reusing intermediate computations, improving frame rate by 5.9× with minimal quality loss [6]. ❷ ***Quantization:*** Fine-tuning neural networks for varying quantization depths, improving rendering quality by up to 5 dB without additional client-side computation [10].

▶**2.3 Other Research on Immersiveness.** I have explored enhancing immersiveness through 360-degree video, enabling users to experience panoramic scenes by rotating their heads. My contributions include: ❶ Predicting user viewports with semantic information [17], ❷ Detecting objects using a pyramid neural network [3], ❸ Enhancing interaction with scene segmentation [23], and ❹ Providing guided viewing experiences through saliency analysis [25].

3. Chasing Reliability with AI-based Redundancy

Immersive computing requires accurately understanding user behaviors and surrounding environments, which involves querying powerful AI models. However, hazards such as packet loss and server failures can disrupt these queries. Existing approaches often consume significant system resources (*e.g.,* bandwidth and computation) to maintain performance under such conditions. My research addresses this issue through AI-based redundancy, which involves proactively training AI models or tailoring inference pipelines to tolerate hazards.

▶**3.1 Proactive Augmentation for Sensing.** Low-quality sensor inputs, such as distorted or occluded images, pose challenges for AI models. My research demonstrates that Large Language Vision Models (LLVMs) can assist the generation of high-quality images of analog gauge readings using a limited set of real images [20]. This approach simplifies autonomous gauge reading on an IoT camera into a template-matching process, achieving a 40× speed improvement over traditional methods without sacrificing accuracy.

▶**3.2 Reconstruction for Communication.** Data loss during communication, caused by quantization and packet loss, degrades AI model performance. ❶ My research restores lost precision by exploiting correlations between pre- and post-quantization content, improving average QoE by up to 31% compared to state-of-the-art methods [8, 16]. ❷ I extend this work to AI-based 3D modeling by reconstructing lost feature values using spatially and temporally adjacent content, improving SSIM from 0.759 to 0.902 when packet loss rate is above 50% [24].

▶**3.3 Collaborative Computation for Inference.** Reliable inference often involves replicating AI models or repeating inference pipelines, which is resource-intensive. My research enables communication between multiple AI models, allowing them to share and reuse intermediate results efficiently. This approach improves accuracy by up to 2.9% compared to replicated methods on ImageNet image classification tasks, while reducing resource demands [5].

▶**3.4 Other Research on Reliability.** Beyond AI models, I have investigated the reliability of chemical container tracking. RF signals attenuate quickly inside chemical containers. Since the placement of chemical containers is random, a tracking algorithm based on RF is rather unreliable. Instead, I designed a tracking system using permanent magnets, which maintain stable fields despite attenuation, achieving energy-efficient tracking (0.48 W) with high localization accuracy (2 cm error) [22].

Future Directions

My prior work has been extended into two NSF proposals on (1) Resilient, Bandwidth-efficient, and Low-latency Immersive Video Streaming [10, 24, 6] and (2) Video Analytics at Scale via Collaborative AI [4, 5]. My future work focuses on AI-system co-designs to bring immersive computing that is interactive, inclusive, and ubiquitous. I plan to propose these ideas to venues such as **IIS: Human-Centered Computing** and **CNS: Networking Technology and Systems**. Specifically, my future work includes the following thrusts: ❶ **Interactivity with AI-based Real-time 3D Modeling:** Training and education are among the most promising applications of immersive computing. For example, photo-realistic 3D models can create immersive modules for surgeons to practice complex procedures. However, current methods require time-consuming pre-training 3D models before streaming, preventing interactive streaming sessions. I aim to explore techniques to reduce these delays, enabling interactive streaming of AI-based 3D models. For instance, shape similarities among objects (*e.g.,* human faces, cars, and tables) can be leveraged to group objects and pre-train generic models for them offline, which amortizes online training overheads. My previous research in AI-based 3D modeling [10, 24, 6] provides a strong foundation for advancing this direction. ❷ **Inclusiveness with AI-based Interface:** Immersive computing must serve not only those capable of standard interactions (*e.g.,* speech) but also individuals facing communication barriers, such as the 1–2% of the global population unable to speak. While sign language is an option, it is not universally understood, limiting accessibility. My research aims to break these barriers by leveraging LLVMs to interpret context and intent from multi-modal signals (*e.g.,* images, sounds) and convert them into speech. Properly designed prompts based on sign language documentation can guide LLVMs in generating meaningful outputs. One of my ongoing projects uses LLVMs to translate hand signals and environmental images into speech for divers, closely aligning with this vision [20]. ❸ **Ubiquity with AI-based Offloading:** LLVMs are poised to become indispensable agents for immersive computing, but their high computational overhead limits their feasibility on ubiquitous devices. Retrieval-Augmented Generation (RAG) is a technique that trades storage for computation time by storing relevant documents. Inspired by it, I aim to explore how caching relevant documents can make LLVMs more practical for more ubiquitous usage. My expertise in content streaming [8, 12, 10, 4, 17, 23, 25], analytics systems [3, 7, 2, 5], and LLVMs [20] equips me to tackle this challenge effectively.

Conclusion and Broader Impacts

I envision a future where humans can access immersive computing in an efficient, immersive, reliable, educational, inclusive, and ubiquitous way. My research focuses on employing AI-system co-designs to make this vision a reality. The proposed research will contribute to a variety of societal domains, as well as undergraduate and graduate education, and further promote Diversity, Equity, and Inclusion (DEI). ❶ This research will enhance diverse applications across domains such as training, healthcare, manufacturing, and entertainment. In these areas, the ability to perceive remote scenes and analyze the physical world is critical. Live streaming of AI-based 3D models and mobile agents with LLVM will help fulfill these demands, offering photo-realistic and interactive scene representations and real-time insights. ❷ This research will incorporate research findings into courses on immersive computing, mobile computing, and multimedia systems. I am also committed to fostering participation from students in traditionally underrepresented groups through mentorship programs and undergraduate research opportunities. ❸ This research will advance DEI by recognizing that people communicate in different ways and providing resources to help those with communication challenges overcome barriers. It will create inclusive systems and opportunities where everyone can fully engage in conversations, social interactions, and professional activities.

References

[1] Civjan Ben, **Bo Chen**, Ruixiao Zhang, and Klara Nahrstedt. Energy-efficient frame filtering at edge with dvfs. In *Submission to NOSSDAV 2025*.

[2] **[Best Student Paper] Bo Chen**, Zhisheng Yan, and Klara Nahrstedt. Context-aware image compression optimization for visual analytics offloading. In *Proceedings of the 13th ACM Multimedia Systems Conference (MMSys)*, pages 27–38, 2022.

[3] **Bo Chen**, Ahmed Ali-Eldin, Prashant Shenoy, and Klara Nahrstedt. Real-time spatio-temporal action localization in 360 videos. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 73–76. IEEE, 2020.

[4] **Bo Chen**, Hongpeng Guo, Mingyuan Wu, Zhe Yang, Zhisheng Yan, and Klara Nahrstedt. {ImmerScope}: Multi-view video aggregation at edge towards immersive content services. In *The 22nd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2024.

[5] **Bo Chen** and Klara Nahrstedt. {MARRS}: Accurate, resilient and secure machine learning analytics via intra-coded computation. In *Submission to SIGCOMM 2025*.

[6] **Bo Chen** and Klara Nahrstedt. {NeuroFlux}: Neural free-viewpoint video streaming with meta caching. In *Submission to SIGCOMM 2025*.

[7] **Bo Chen** and Klara Nahrstedt. {EScALation}: a framework for efficient and scalable spatio-temporal action localization. In *Proceedings of the 12th ACM Multimedia Systems Conference (MMSys)*, pages 146–158, 2021.

[8] **Bo Chen**, Mingyuan Wu, Hongpeng Guo, Zhisheng Yan, and Klara Nahrstedt. {Vesper}: Learning to manage uncertainty in video streaming. In *Proceedings of the 15th ACM Multimedia Systems Conference (MMSys)*, pages 166–177, 2024.

[9] **Bo Chen**, Zhisheng Yan, Hongpeng Guo, Zhe Yang, Ahmed Ali-Eldin, Prashant Shenoy, and Klara Nahrstedt. Deep contextualized compressive offloading for images. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 467–473, 2021.

[10] **Bo Chen**, Zhisheng Yan, Bo Han, and Klara Nahrstedt. {NeRFHub}: A context-aware nerf serving framework for mobile immersive applications. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*, pages 85–98, 2024.

[11] **Bo Chen**, Zhisheng Yan, and Klara Nahrstedt. Context-aware optimization for bandwidth-efficient image analytics offloading. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[12] **Bo Chen**, Zhisheng Yan, Yinjie Zhang, Zhe Yang, and Klara Nahrstedt. {LiFteR}: Unleash learned codecs in video streaming with loose frame referencing. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pages 533–548, 2024.

[13] Li Jiaxi, **Bo Chen**, Ruixiao Zhang, and Klara Nahrstedt. Energy-efficient video analytics with efficient image filter. In *Submission to ATC 2025*.

[14] Jiaxi Li, Jingwei Liao, **Bo Chen**, Anh Nguyen, Aditi Tiwari, Qian Zhou, Zhisheng Yan, and Klara Nahrstedt. St-360: Spatial–temporal filtering-based low-latency 360-degree video analytics framework. *ACM Transactions on Multimedia Computing, Communications and Applications*.

[15] Jiaxi Li, Jingwei Liao, **Bo Chen**, Anh Nguyen, Aditi Tiwari, Qian Zhou, Zhisheng Yan, and Klara Nahrstedt. Latency-aware 360-degree video analytics framework for first responders situational awareness. In *Proceedings of the 33rd Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, pages 8–14, 2023.

[16] Wei Luo and **Bo Chen**. Neural image compression with quantization rectifier. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023.

[17] **[Best Paper Award]** Jounsup Park, Mingyuan Wu, Kuan-Ying Lee, **Bo Chen**, Klara Nahrstedt, Michael Zink, and Ramesh Sitaraman. Seaware: Semantic aware view prediction system for 360-degree video streaming. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 57–64. IEEE, 2020.

[18] **Chen Bo** Rui-xiao Zhang, Tianchi Huang and Klara Nahrstedt. {NeRFlow}: Towards adaptive streaming for nerf videos. In *Proceedings of the 23nd Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2025.

[19] Aaron Seidlitz. Alumnus bo chen wins best student paper award from acm multimedia systems 2022, 2022. `https://siebelschool.illinois.edu/news/alumnus-bo-chen-wins-best-student-paper-award-from-acm-multimedia-systems-2022`.

[20] Beitong Tian, Mingyuan Wu, Ruixiao Zhang, Haozhen Zheng Zheng, **Bo Chen**, Yaohui Wang, Shiv Trivedi, Shanbo Zhang, Robert Kaufman, Leah Espenhahn, Gianni Pezzarossi, Mauro Sardela, John Dallesasse, and Klara Nahrstedt. {GaugeTracker}: Ai-powered cost-effective analog gauge monitoring system. In *Proceedings of IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2024.

[21] Beitong Tian, Lingzhi Zhao, **Bo Chen**, Haozhen Zheng, Mingyuan Wu, and Klara Nahrstedt. {AquaScope}: Reliable underwater image transmission using mobile devices. In *Submission to SIGCOMM 2025*.

[22] Yaohui Wang, **Bo Chen**, Beitong Tian, and Klara Nahrstedt. {EcoMat}: Cost-effective magnet-based asset tracking. In *Submission to IMWUT 2024*.

[23] Mingyuan Wu, Yuhan Lu, Shiv Trivedi, **Bo Chen**, Qian Zhou, Lingdong Wang, Simran Singh, Michael Zink, Ramesh Sitaraman, Jacob Chakareski, and Klara Nahrstedt. Interactive scene analysis for teleconferencing. In *2023 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2023.

[24] Nan Wu, **Bo Chen**, Klara Nahrstedt, and Bo Han. {NeVo}: Advancing volumetric video streaming with neural content representation. In *The 31st Annual International Conference On Mobile Computing And Networking (MobiCom 2025)*.

[25] Yinjie Zhang, Mingyuan Wu, Beitong Tian, Jiaxi Li, **Bo Chen**, Qian Zhou, and Klara Nahrstedt. {SAVG360}: Saliency-aware viewport-guidance-enabled 360-degree video streaming system. In *2023 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2023.