Learning from Interaction: Reliable Online Adaptation in Multimodal AI

Jitesh Jain, Ph.D. Candidate, Georgia Tech

jitesh.jj2@gmail.com praeclarumjj3.github.io Google Scholar

A multimodal model that ships today is frozen. It answers one prompt well, then runs millions of interactions that teach it nothing. Worse, the systems that do try to learn from those interactions often degrade, because feedback arrives sparse and biased and one bad update propagates into everything that follows. I want to develop **multimodal models that improve from interaction without losing reliability**. My work has moved from getting perception right, to giving agents memory and control, toward how a model should improve itself on the go.

Perception & Reasoning. Adaptation starts from a foundational understanding of the task. *OneFormer* (CVPR 2023) [1], a universal modeling recipe that made semantic, instance, and panoptic segmentation one model trained once, is my most-cited work (**900+ citations**); I integrated it into Hugging Face Transformers, where it ships as a standard segmentation model. *VCoder* (CVPR 2024, **100+ citations**) [2] bridged perception and language, showing that multimodal LLMs still failed at counting and depth ordering, and repairing it with segmentation and depth maps. *VisPer-LM* (NeurIPS 2025) [3], with Microsoft Research, moved that fix inside the model, distilling expert visual features into the LLM's hidden states for **8.7%** better spatial reasoning with no extra encoders at inference.

Agents & Memory. A deployed system must also change its behavior based on environment feedback. *AUGUSTUS* (NeurIPS 2025 Workshop) [4] builds hierarchical memory for personal agents, summarizing interaction history into concept-indexed nodes retrieved by meaning rather than flat index, running **3.5× faster** than multimodal RAG at better conversational consistency. *SAGE* (CVPR 2026) [5], built at Allen AI alongside the open-weights video-language model *Molmo2* [6], is a recipe for any-horizon agents: an RL orchestrator that decides how much evidence to gather before answering, worth **8.2%** on videos beyond ten minutes.

Online Adaptation. I want to develop models that figure out how to improve on the go. This is **both an algorithmic and an infrastructure challenge**, and the environment is what makes it hard: feedback is scarce and biased, since users correct what annoys them rather than what is most wrong, and nothing marks which past update is costing accuracy now. At NVIDIA, I have been working on the tractable end of this: test-time learning for GPU kernel optimization, and automated research agents built on **>500B**-parameter MoE models, spanning both the algorithms and the infrastructure they run on. The open problem is everything without such a verifier. While a system is live, we have no way to tell whether an update helped or quietly cost something elsewhere. Forgetting, cross-user leakage, and irreversibility are what matter once a model is deployed, and none of them appear in how we benchmark agents today. I want to fix that measurement first, scoring a system on how much it regresses and not only on how much it improves, and then build the update policy that earns its keep against that measure.

I am looking for a research role where deployed interaction is the setting rather than an afterthought. That is where these questions stop being hypothetical, and where the algorithms and the infrastructure have to be built together.

References

[1] **Jitesh Jain**, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *CVPR*, 2023.

[2] **Jitesh Jain**, Jianwei Yang, and Humphrey Shi. VCoder: Versatile Vision Encoders for Multimodal Large Language Models. In *CVPR*, 2024.

[3] **Jitesh Jain**, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. Elevating Visual Perception in Multimodal LLMs with Visual Embedding Distillation. In *NeurIPS*, 2025.

[4] **Jitesh Jain**, Shubham Maheshwari, Ning Yu, Wen mei Hwu, and **Humphrey Shi**. AUGUSTUS: An LLM-driven multimodal agent system with contextualized user memory. In *LAW Workshop @ NeurIPS*, 2025.

[5] **Jitesh Jain**, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, and **Humphrey Shi**. SAGE: Training smart any-horizon agents for long video reasoning with reinforcement learning. In *CVPR*, 2026.

[6] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, **Jitesh Jain**, George Stoica, Winston Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. In *CVPR*, 2026.