RESEARCH STATEMENT

AI That Thinks and Acts According to What It Senses : Sensory-Adaptive Intelligence

Daeun Lee

PhD Candidate at UNC Chapel Hill · daeun@cs.unc.edu

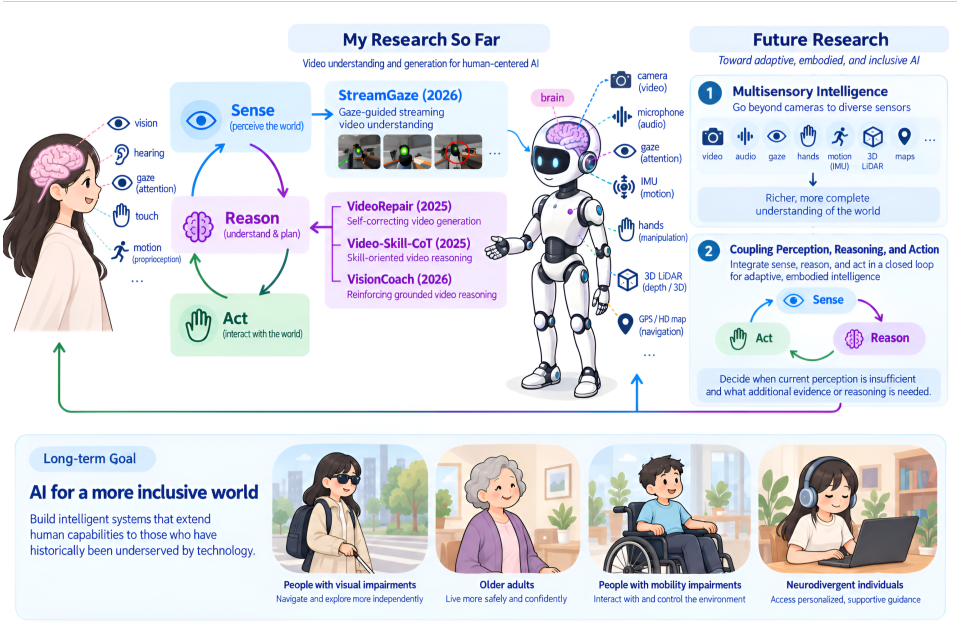
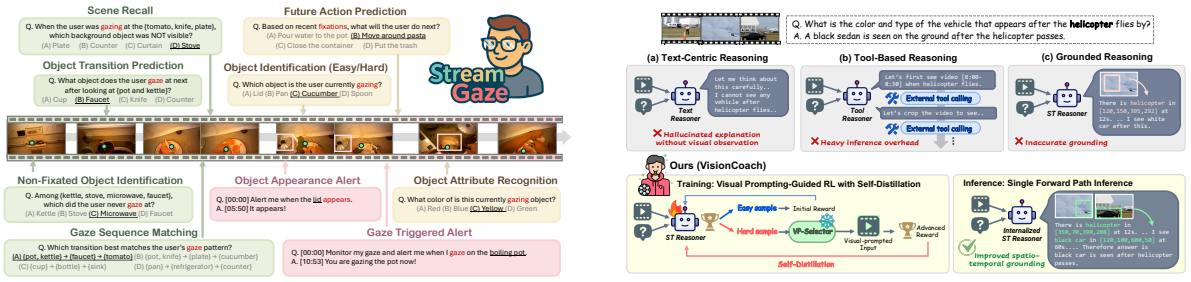


Figure 1. From human multisensory perception to embodied AI that **senses**, **reasons**, and **acts** to support diverse human needs.

Physical world does not reveal itself through a single modality, and intelligence cannot emerge from perception, reasoning, or action alone. Humans combine diverse **senses**, **adapt** their reasoning to the evidence available, and **act** in ways that reshape what they perceive next.

Despite remarkable progress, today's AI remains far from this form of intelligence in two fundamental ways. **First, current AI perceives only a fraction of the physical world.** Most foundation models rely primarily on vision and language, while humans and embodied agents draw on far richer sensory evidence, including gaze, depth, touch, motion, and 3D spatial context. Yet the physical world is inherently partially observable: no single sensory channel provides all the evidence needed to understand every situation. Robust physical intelligence therefore requires integrating complementary sensory signals to reveal what any single modality leaves unobserved. **Second, perception, reasoning, and action remain disconnected.** Most AI systems receive predefined inputs, reason over them, and produce an answer or action. While each capability has advanced rapidly in *isolation*, intelligence in the physical world requires a **closed loop**: what an agent perceives should shape how it reasons; reasoning should determine what additional evidence or action is needed; and those actions should, in turn, shape what the agent perceives next.



StreamGaze (Lee et al., 2026) **VisionCoach** (Lee et al., 2026)

Figure 2. Sensory-adaptive intelligence. Left: Expanding What AI Can Sense through richer sensory signals, as explored in StreamGaze. **Right: Adapting Intelligence to What It Senses** by changing how models reason based on perceptual evidence, as explored in VisionCoach.

These two gaps motivate my research on **Sensory-Adaptive Intelligence**: building AI that perceives the physical world through richer sensory signals and adapts its reasoning and actions to the evidence they provide. My research asks: **How can we expand what AI can sense, and how should its intelligence adapt to what it senses?**

Expanding What AI Can Sense

My first research direction expands AI beyond conventional visual inputs toward richer sensory evidence. Across my work, I have studied whether AI can interpret and reason with LiDAR [1], HD maps [2], and human gaze [3]. In StreamGaze [3], I investigated whether streaming video understanding could move from passive visual observation toward the continuous, user-centered perception required by AR glasses. Gaze reveals more than where someone looks: it provides evidence about what captures their attention, what they are searching for, and what they may do next. developed a benchmark evaluating gaze-guided temporal reasoning, intention understanding, and proactive prediction. Humans achieved 82.7% accuracy, while even a leading closed-source multimodal model, GPT-4o, achieved only 53.5%, exposing a **substantial gap in current models' ability to reason from this sensory signal**. This led to a broader insight: giving AI another sensor does not automatically make it more perceptive. A multisensory system must understand what information each signal contributes, when that information matters, and how it should change subsequent reasoning and decisions.

Adapting Intelligence to What It Senses

Richer perception is useful only if what an agent perceives can change how it thinks and what it does next. My second research direction therefore investigates how intelligence can adapt to its **input** [4–7]. In Video-Skill-CoT [5], I studied whether video reasoning could adapt to the input domain rather than applying the same reasoning strategy to every video, by learning specialized reasoning skills and selecting among them based on the input. In VisionCoach [6], I extended adaptation to the learning process itself. VisionCoach identifies perceptually difficult questions during reinforcement learning and selectively provides visual prompts that direct the model toward relevant evidence. This adaptive training improved video question answering accuracy from 33.5% to 61.1%, demonstrating how recognizing when perception is insufficient can substantially improve reasoning. I further connected observation to action in VideoRepair [7], where a model inspects its generated video, diagnoses semantic

errors, and selectively regenerates erroneous regions while preserving correct content. Rather than producing an output once, the system uses what it observes about its own failure to determine what to change next. Together, these projects demonstrate early forms of input-adaptive intelligence, where observations can alter reasoning, learning, and generation. Yet this adaptation remains confined to predefined tasks, limited sensory inputs, and digital outputs. Real-world intelligence requires a more general capability: adapting even when an agent does not know in advance what sensory evidence, reasoning strategy, or action it will need.

Beyond Passive Perception: Toward Sensory-Adaptive Embodied AI

Building on these two directions, I will continue developing AI that can better understand and reason over diverse sensory inputs, while more tightly closing the loop among perception, reasoning, and action. Beyond this, my long-term vision is to build embodied AI that does not merely process available sensory inputs, but **actively constructs the perception it needs to help people in the physical world**. Rather than relying on fixed modalities or a predetermined reasoning pipeline, such systems should decide what to sense, what information is missing, and what action could reveal it.

Consider an assistive robot helping a person with limited vision find and retrieve medication in an unfamiliar kitchen. Vision alone may not reveal whether a bottle is occluded, whether it is the correct one, or whether it can be safely grasped. The robot could reposition itself for a clearer view, use depth to localize the bottle, inspect its label more closely, and use tactile feedback during manipulation. Each action would reveal new sensory evidence that updates the robot's understanding and determines what it should do next.

Ultimately, I hope to build AI whose perception is not fixed, but **actively shaped by human needs and the demands of the physical world**.

References

[1] Daeun Lee and Jinkyu Kim. "Resolving Class Imbalance for Lidar-based Object Detector by Dynamic Weight Average and Contextual Ground truth Sampling". In: *WACV 2023*. 2023, pp. 682–691.

[2] Daeun Lee, Minhyeok Heo, and Jiwon Kim. "HD Maps are Lane Detection Generalizers: A Novel Generative Framework for Single-Source Domain Generalization". In: *CVPR 2024 Data-Driven Autonomous Driving Simulation Workshop*. 2024.

[3] Daeun Lee, Subhojyoti Mukherjee, Branislav Kveton, Ryan A Rossi, Viet Dac Lai, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, and Mohit Bansal. "Streamgaze: Gaze-guided temporal reasoning and proactive understanding in streaming videos". In: *arXiv preprint arXiv:2512.01707* (2025).

[4] Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. "BECoTTA: Input-Dependent Online Blending of Experts for Continual Test-Time Adaptation". In: *International Conference on Machine Learning (ICML)*. 2024.

[5] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. "Video-Skill-CoT: Skill-based Chain-of-Thoughts for Domain-Adaptive Video Reasoning". In: *EMNLP 2025 Findings* (2025).

[6] Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. "VisionCoach: Reinforcing Grounded Video Reasoning via Visual-Perception Prompting". In: *European Conference on Computer Vision (ECCV)*. 2026.

[7] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. "Self-Correcting Text-to-Video Generation with Misalignment Detection and Localized Refinement". In: *Findings of the Association for Computational Linguistics: ACL 2026*. 2026, pp. 36464–36489.