Research Statement

Daechul Ahn

Building AI Agents that Perceive, Reason, and Act as Humans Do

Updated: April 2026

My research goal is to build AI agents that *perceive*, *reason*, and *act* as humans do—agents that perceive **faithfully** without hallucinating what isn't there, reason **reflectively** through coordination or self-critique, and act **adaptively** in dynamically changing environments. Throughout my research, I have pursued this vision along three intertwined directions: (1) faithful multimodal perception, (2) reflective agentic reasoning, and (3) adaptive embodied decision-making.

(1) Faithful multimodal perception. My research in multimodal video-language understanding began with challenges in language-driven temporal grounding: enabling temporal localization without paired supervision [1], maintaining coherent context over long horizons [2], and leveraging fine-grained temporal cues to facilitate corpus-level video moment retrieval [3]. However, as video-language models grew more capable (*i.e.*, Video-LLMs), a deeper concern emerged: capability is *not* faithfulness. Standard *supervised fine-tuning* approach often leads models to produce plausible outputs that are poorly grounded in the visual evidence, *i.e.*, *hallucination*. To address this, I proposed AI feedback-driven reinforcement learning approach [4, 5]. I further observed that even well-aligned, high-performing Video-LLMs tend to rely on language-prior shortcuts, rather than grounding in the video's temporal structure, which motivated me to develop a temporal diagnostic benchmark and an explicit multi-event chain-of-thought approach for temporal understanding [6]. Together, these efforts point to a lesson: faithful multimodal perception demands both semantic and temporal grounding—not only in *what* models see, but in *when* things happen.

(2) Reflective agentic reasoning. Beyond perceiving faithfully, effective decision-making requires an agent to *reason*—to plan, evaluate, and revise. Yet a single pass of reasoning may not be enough: plans rest on unchecked assumptions, errors compound without inspection, and complex tasks overwhelm a single reasoning path forced to weigh many competing considerations at once. These limitations motivated me to pursue *reflective* agentic reasoning—reasoning that examines and corrects itself, moving beyond single-turn generation to plan over extended horizons, coordinate across roles, and revise its own decisions. I first explored reflective reasoning in real-time strategy (RTS) games—a testbed that demands long-horizon planning, real-time adaptation, and reasoning under partial observability. There, I found that decomposing reasoning into specialized roles and reflectively aggregating their judgments through multi-agent hierarchies enables better strategic reasoning than monolithic prompting [7]. Beyond coordinating reasoning across agents, I also study reflection as a mechanism for self-improvement. LLM evaluators typically assess each sample in isolation, unable to learn from their past judgments; I showed how they can self-improve at test time by learning from their own experience [8]. Similarly, I equipped RTSGameBench [9]—a large-scale RTS benchmark I built—with a self-evolving pipeline of reflective agents that continuously expand the benchmark while improving the pipeline itself across generation cycles. Across coordinated reasoning and continual self-improvement, reflective agency turns decision-making into an iterative process rather than a single forward pass.

(3) Adaptive embodied decision-making. Deploying AI agents to assist with everyday tasks in the real world demands more than faithful perception and reflective reasoning. It requires agents to act *adaptively* under varying circumstances—objects move, sub-goals evolve mid-task, and some

steps turn out harder than others. This motivated me to pursue adaptive decision-making for embodied agents under the real-time constraints of physical interaction. Reliable operation in such environments raises two distinct demands: agents must know *when* to revise their course of action, and *how confidently* to act at each step along the way. Addressing the first, I propose BINDER [10], which decouples deliberative planning from continuous execution-time monitoring, enabling open-vocabulary mobile manipulators to detect state changes in real time and trigger replanning in dynamic environments. Addressing the second, I propose SCALE [11], which exploits a vision-language-action model's own uncertainty to modulate visual attention and action decoding at each step—exploring when uncertain and committing when confident—without additional training or multiple forward passes. Together, these contributions frame adaptive embodied action as a matter of *when* to reconsider and *how firmly* to commit—two faces of the same competence.

Looking forward. The three directions connect through one question: how to close the gap between what a model can do and what it can be trusted to do. Perception must be grounded, reasoning process must examine itself, and action must adjust to what the moment demands. Going forward, I plan to deepen each direction in its own right while remaining attentive to how they inform one another–working toward AI agents that, like humans, earn trust through how faithfully they see, how carefully they think, and how adaptively they act under varying circumstances.

References

† denotes equal contribution.

[1] Jinwoo Nam†, **Daechul Ahn**†, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot Natural Language Video Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. **Oral presentation.**

[2] **Daechul Ahn**, Daneul Kim, Gwangmo Song, Seung Hwan Kim, Honglak Lee, Dongyeop Kang, and Jonghyun Choi. Story Visualization by Online Text Augmentation with Context Memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[3] Yura Choi†, **Daechul Ahn**†, and Jonghyun Choi. Moment-Aware Video Retrieval for Video Corpus Moment Retrieval. *IEEE Access*, 2025.

[4] **Daechul Ahn**, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. **Oral presentation.**

[5] **Daechul Ahn**†, Yura Choi†, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. ISR-DPO: Aligning Large Multimodal Models for Videos by Iterative Self-Retrospective DPO. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[6] **Daechul Ahn**†, Yura Choi†, Hyeonbeom Choi*, Seongwon Cho, San Kim, and Jonghyun Choi. What Happens When: Learning Temporal Orders of Events in Videos. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

[7] **Daechul Ahn**†, San Kim†, and Jonghyun Choi. Society of Mind Meets Real-Time Strategy: A Hierarchical Multi-Agent Framework for Strategic Reasoning. In *Conference on Language Modeling (COLM)*, 2025.

[8] Seungyeon Jwa, **Daechul Ahn**, Reokyoung Kim, Dongyeop Kang, and Jonghyun Choi. Becoming Experienced Judges: Selective Test-Time Learning for Evaluators. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026. Short paper, **Oral presentation**.

[9] San Kim†, **Daechul Ahn**†, Reokyoung Kim, Hyeonbeom Choi, Seungyeon Jwa, and Jonghyun Choi. RTSGameBench: An RTS Benchmark for Strategic Reasoning by Vision-Language Models. *under review*, 2026.

[10] Seongwon Cho†, **Daechul Ahn**†, Donghyun Shin, Hyeonbeom Choi, San Kim, and Jonghyun Choi. BINDER: Instantly Adaptive Mobile Manipulation with Open-Vocabulary Commands. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[11] Hyeonbeom Choi†, **Daechul Ahn**†, Youhan Lee, Taewook Kang, Seongwon Cho, and Jonghyun Choi. SCALE: Self-uncertainty Conditioned Adaptive Looking and Execution for Vision-Language-Action Models. *arXiv preprint*, 2026.