Toward Multimodal Agents that Perceive, Reason, Act and Predict

Yan Ma

Research Agenda

I view multimodal agents as goal-directed systems that work in partially observed multimodal environments, where they must perceive what is happening, reason about it, act to gather useful information, and predict what will happen next. Such systems are needed to complete long-horizon multimodal tasks that unfold through interaction, such as multi-step computer use, decision-making from video or audio streams, and embodied tasks like robotics or driving. A central question motivating my research is how multimodal agents can not only perceive, reason, and act based on current observations, but also predict how the world will evolve after their actions. So far, my work has mainly focused on the first part of this question, using reinforcement learning (RL) to improve reasoning, perception, and tool use in multimodal foundation models. More recently, I have begun to extend this agenda toward predictive multimodal learning by studying how large-scale video data can support the anticipation of future observations and longer-term outcomes.

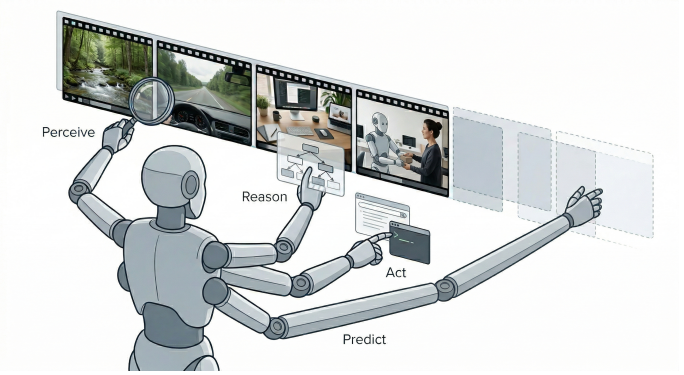

Figure 1: Research agenda toward multimodal agents that perceive, reason, act, and predict. Solid frames denote current observations and dashed frames denote anticipated future states; the goal is to integrate prediction into multimodal agent systems so that future observations and action outcomes improve perception, reasoning, and action in long-horizon multimodal tasks.

Past Work: Multimodal RL for Perception, Reasoning, and Tool Use

To study how multimodal agents perceive, reason, and act under current observations, I have mainly approached this question through RL and asked a more basic question: how does RL change multimodal models? My work shows that these changes are not only about higher benchmark scores. They also appear in training dynamics, in the joint improvement of reasoning and perception, and in what models actually learn in tool-use settings.

First, in reasoning-centric settings, I studied how RL changes training dynamics and generalization. The results show that RL changes more than final performance: it reshapes response length, reflection, and other training behaviors, and these changes are themselves sensitive to random seeds. On visual math tasks, RL also shows better generalization than high-quality SFT. I then asked a broader question: if RL mainly helps on reasoning-heavy tasks, can it really become a general training method for multimodal foundation models? To study this, we designed V-Triune, a unified framework for joint reasoning-and-perception training, and validated stable joint training on 8 reasoning and perception tasks. The results show RL can shape reasoning and perception together, with complementary gains across the two.

I then became interested in what RL learns when actions start to affect how the model gathers information. In vision tool-use RL, our analysis shows that the gains mostly come from intrinsic capability improvement and reduced tool-induced harm, rather than true tool mastery.

Taken together, these works form one basic line of my research on multimodal agents. I first study how models perceive and reason from current observations, and then move to more agentic settings where actions begin to change the information available to the model. These results also suggest that, for long-horizon multimodal tasks that unfold through ongoing interaction, acting only from current observations is not enough; models also need to anticipate how future states will evolve and what consequences their actions will bring.

Current and Future Work: Predictive Multimodal Learning from Video

To move toward predictive multimodal learning, I am currently studying its video data foundations. I think of predictive multimodal learning as modeling environment dynamics across multimodal tasks, where the target may range from future observations to more abstract internal states. Video is a particularly important training source in this setting because it provides scalable temporal experience. This leads to a basic question: how should video data be segmented, filtered, annotated, and organized for predictive learning?

In my current work, I use video generation models as a concrete testbed for this question, in which future states are expressed as future frames or clips. This setting allows me to study how data choices shape the temporal structure that models learn. Going forward, I want to study how predictive models can be integrated into multimodal agent systems, so that anticipating future observations and action outcomes can directly improve perception, reasoning, and action in long-horizon multimodal tasks. In the long run, I hope this line of work will help build multimodal agents that not only understand the present, but also act, plan, and adapt by anticipating what may happen next.