The Multimodal Mind

Siddharth Choudhary • March 2026

The Thesis

The next frontier in AI is a single model that understands the physical world well enough to imagine its futures and act to shape them. The path runs through a natively multimodal foundation model—one that perceives, reasons, and generates across all modalities (image, video, text, speech, action)—organized around a three-stage cognitive loop: Perceive → Imagine → Act.

Perceive → Imagine → Act

Biological intelligence follows a simple loop: observe the world, mentally simulate what should happen, then intervene. Current AI breaks this into disconnected pipelines. I propose building it as one model.

1. **Perceive.** The model ingests the current state across all modalities—video, audio, proprioception, language—through a natively multimodal architecture. Perception is inference in a unified space, not a separate module.
2. **Imagine.** Before acting, the model thinks in multimodal space—not just text chain-of-thought, but mental video of the planned trajectory, predicted proprioception, anticipated contact forces. A model planning to pick up a glass imagines the hand approaching, fingers closing, the weight shifting. This multimodal imagination is the world model.
3. **Act.** The model generates actions in whatever output the task requires—motor commands for a robot, keyboard/mouse for a computer, speech for conversation—conditioned on the imagined plan, not reactive perception alone. The action changes the world, new percepts arrive, the loop repeats.

Training Recipe

The recipe mirrors LLMs but extends every stage into multimodal space. Pretraining: next-token prediction across text, images, video, audio, screen recordings, and robot trajectories at internet scale—learning the prior over how the multimodal world works. Post-training: supervised fine-tuning on curated Perceive → Imagine → Act demonstrations with explicit multimodal reasoning traces. RL: self-play in simulation with automatic reward signals.

The Physics Flywheel: Self-Play in Simulation

Self-play is where this becomes self-improving. In simulation, the model runs its full loop and receives automatic rewards: Did what I imagined actually happen? (imagination accuracy). Did I succeed? (task completion). Does my prediction violate physics? (consistency via engine constraints—"RL from physics feedback," RLPF). Each cycle improves the world model, which improves planning, which improves actions, which generates richer training data. This is the compound scaling law text LLMs cannot access. The rate limiter is simulation fidelity—which is why a learned world simulator, trained on real video and corrected by physics engines, is essential.

One Model, Many Problems

Every major AI capability challenge today is a special case of multimodal conditional generation—and this single architecture subsumes them all:

Controllable video generation = text + image → video

Computer use = video → action (keyboard/mouse)

Robotics = video + text → action (motor commands)

World simulation = video + action → video

Autonomous driving = video + map → action (steering/throttle)

Creative tools = text + image → image + video + audio

Today each of these is a separate model, dataset, and research community. The Perceive → Imagine → Act framework unifies them as different routes through one architecture—and improvements to the shared backbone benefit every task simultaneously.