🤖 Gemini Scout: The Embodied AI Rover

🔗 Project Links

- 📺 **YouTube Demo:** Watch the Video
- 💻 **GitHub Repository:** vikrantkulkarni07/Physical-World-Gemini-Scout

---

💡 Elevator Pitch

"We gave Gemini wheels." Gemini Scout is an autonomous rover powered by **Gemini 3 Flash**. It replaces traditional LIDAR and hard-coded logic with a single Multimodal LLM that "sees," reasons about physics, and navigates the real world using only a camera.

📖 The Story

Inspiration

We are living in the golden age of AI. Models like Gemini 3 can write poetry, debug complex kernels, and reason through scientific papers. Yet, there is one massive limitation: **AI is still trapped behind a glass screen.**

Most "Agentic" workflows today are just software talking to software. We wanted to build something different. We wanted to explore **Embodied AI**—giving a Large Multimodal Model (LMM) a physical body and the agency to move through the real world.

We asked a simple question: *Can we replace the entire navigation stack of a robot (LIDAR, SLAM, Object Detection) with a single call to the Gemini 3 API?*

What It Does

Gemini Scout is an autonomous rover that uses **Gemini 3 Flash** as its visual cortex and decision-making engine.

Unlike traditional robots that follow hard-coded paths or rely on specific object detection models (like YOLO), Scout "looks" at the world and "thinks" about it:

1. **It Sees:** Streams live video from an on-board camera.
2. **It Reasons:** Analyzes the scene for obstacles, terrain types (carpet vs. tile), and context (fragile objects vs. robust obstacles).

3. **It Acts:** Decides on movement commands (Forward, Left, Right) to navigate toward a goal or simply explore without crashing.

How We Built It

We built Scout using a "Hybrid Brain" architecture to balance cost and intelligence.

1. The Body (Hardware)

- **Microcontroller:** ESP32-CAM (Cheap, low-power, Wi-Fi enabled).
- **Actuators:** 4x DC Motors with an L298N Driver.
- **Chassis:** Standard 4WD Robotic Platform.
- **Power:** 2x 18650 Li-ion batteries.

2. The Brain (Software)

The ESP32 acts as a "dumb terminal," streaming raw MJPEG video over Wi-Fi. The heavy lifting is done by a Python control loop running on a host machine:

1. **Frame Capture:** We grab a frame from the video stream using OpenCV.
2. **The "Thought" Process:** We send the frame to **Gemini 3 Flash** with a specialized system prompt. We force the model to output structured **JSON** containing its reasoning and motor commands.
3. **Execution:** The Python script parses the JSON and sends HTTP GET requests back to the ESP32 to trigger the motors.

Challenges We Ran Into

- **The Latency Loop:** Real-time robotics usually requires millisecond-level reactions. Sending an image to the cloud and waiting for a response takes time (~500ms - 1s). To fix this, we implemented a "heartbeat" safety stop on the microcontroller—if it doesn't receive a new command within 200ms, it halts automatically.
- **Visual Hallucination:** Initially, the model would confidently say "Path Clear" while staring at a white wall because it looked like an open floor. We solved this by improving the system prompt to force "Chain of Thought" reasoning.

Accomplishments We're Proud Of

- **True Zero-Shot Navigation:** We didn't train any custom models. The rover can identify a "shoe" or a "cable" without us ever writing code for those specific objects.
- **Complex Reasoning:** Seeing the rover refuse to drive over a pile of wires because it "might get tangled" was a huge win. It showed the model understands physical consequences.

What We Learned

- **Gemini has "Physics Intuition":** We were surprised to find that Gemini 3 understands physical properties implicitly. It knows that glass is fragile and curtains are soft.
- **Multimodal is the Future:** Removing the need for specific sensor stacks is a game changer. Visual data + reasoning is enough for basic navigation.

What's Next for Gemini Scout

We plan to integrate **Gemini Live (Audio)** to create a fully interactive pet.

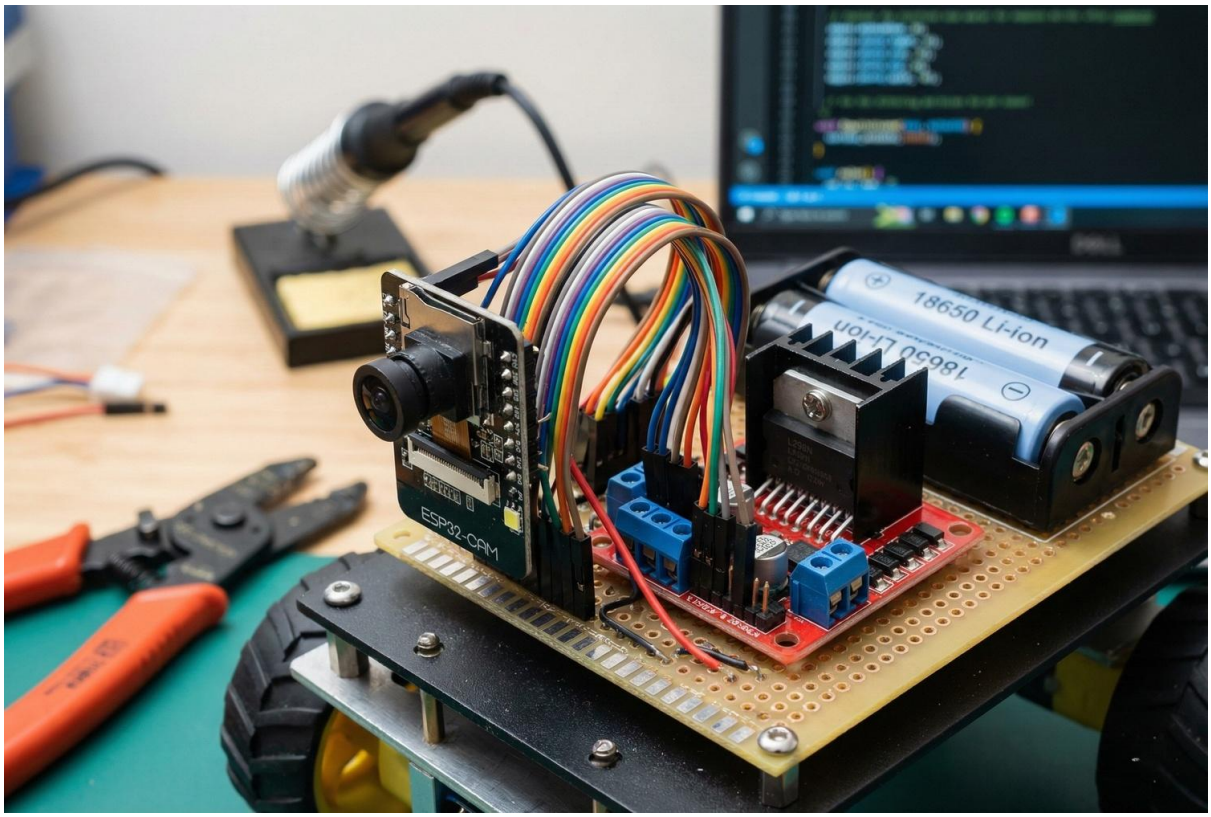
- *"Scout, come here!"* (Audio processing)
- *"Find my keys."* (Visual search)

🛠 Tech Stack

- **AI Model:** Gemini 3 Flash
- **SDK:** Google GenAI SDK
- **Language:** Python (Control Logic), C++ (Firmware)
- **Hardware:** ESP32-CAM, L298N Motor Driver, 4WD Chassis
- **Libraries:** OpenCV, Requests, NumPy

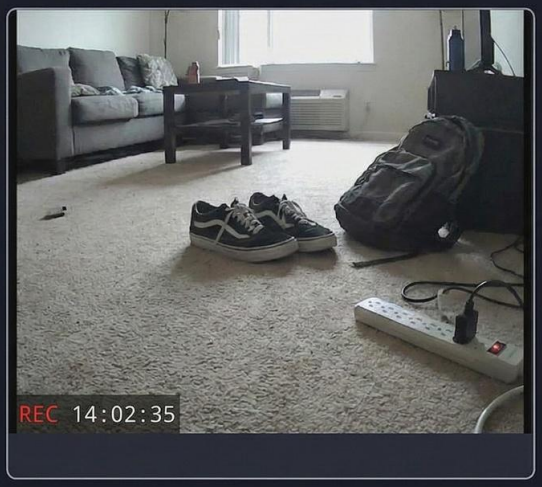
Snapshots


GEMINI SCOUT
Embodied AI | Gemini 3 API
OBSTACLE: Cable Cluster (Avoid)
PATH: Clear (Confidence 95%)


ESP32-CAM
18650 Li-ion

GEMINI 3: MULTIMODAL REASONING PIPELINE

RAW VISION INPUT
(ESP32-CAM)



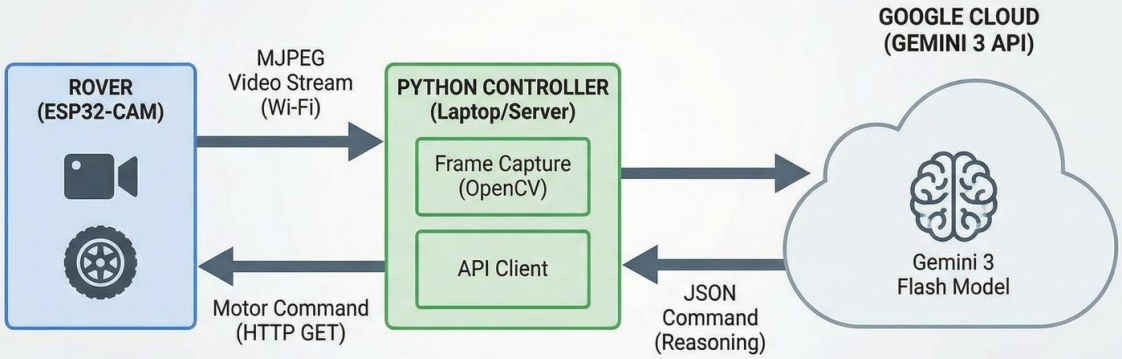
AI REASONING OUTPUT
(GEMINI 3 FLASH)

```
Analyzing Frame... [OK]

Detected: 'Sneaker' (Obstacle),
'Backpack' (Obstacle), 'Carpet' (Travers

Reasoning: Path ahead is blocked by
sneakers. Clear space detected to the
left. I will turn left to avoid
collision.

OUTPUT JSON:
{
  'command': 'LEFT',
  'speed': 180
}
```

RAW VISION INPUT
(ESP32-CAM)

AI REASONING OUTPUT
(GEMINI 3 FLASH)

SYSTEM ARCHITECTURE: HYBRID EDGE-CLOUD THINKING




OBSTACLE:
Shoes (Avoid)