NOTE 1

Vision-Grounded GPT for Real-World Systems

Sentinel Agro Labs — Research Note

Abstract

Large Language Models (LLMs) such as GPT demonstrate strong reasoning and planning capabilities but remain fundamentally limited by their lack of direct grounding in the physical world. This note outlines a research direction focused on integrating GPT-style language models with vision-based representations to improve real-world understanding, consistency, and applicability in domains such as robotics, drones, and sensor-driven agriculture.

Background

Current GPT-style models operate primarily on symbolic text representations. While effective for abstraction and reasoning, this approach can lead to:

- hallucinations,
- weak spatial understanding,
- limited awareness of physical constraints.

In contrast, vision and video models encode structure, geometry, and dynamics of real environments. Recent research (e.g., JEPA-based architectures, self-supervised visual embeddings) suggests that compact latent representations of the world can support prediction and planning without relying on raw pixels.

Core Idea

We propose a **vision-grounded language architecture** in which:

- A **vision or video model** encodes the environment into a compact latent "world state".
- A **GPT-style language model** reasons *over this latent state*, rather than over raw sensory data.
- Language is used for:
  - planning,
  - explanation,
  - decision justification,
  - human interaction.

This separation allows perception and reasoning to specialize while remaining tightly coupled.

Conceptual Architecture

1. **Perception Layer**
   - Vision / video encoders (self-supervised).
   - Produces stable latent representations of scenes and dynamics.
2. **Reasoning Layer**
   - GPT-style LLM operates on:
     - latent state summaries,
     - structured descriptors,
     - retrieved domain knowledge.
3. **Grounding & Safety Layer**
   - Consistency checks.
   - Uncertainty estimation.
   - Human-in-the-loop review where required.

Potential Applications

- Robotics navigation and task planning
- Autonomous drones (inspection, monitoring, agriculture)
- Sensor-driven decision systems
- Explainable AI for environmental monitoring

Status

This work represents a **research direction** and conceptual framework. No deployed system is claimed at this stage. Development will proceed through staged prototypes and evaluation.

References (selected)

- GPT-4 Technical Report
- V-JEPA / VL-JEPA (Meta AI)
- DINOv2, SAM