

The logo for the 2024 embedded VISION SUMMIT is centered on the left side of the slide. It features a white octagonal background with a colorful, multi-layered border in shades of purple, blue, green, yellow, and orange. The text "2024" is at the top, "embedded" is below it, "VISION" is in large, bold, dark blue letters with a gradient, and "SUMMIT" is at the bottom in a smaller, dark blue font.

2024
embedded
VISION
SUMMIT®

How Large Language Models Are Impacting Computer Vision

Jacob Marks

Senior ML Engineer & Researcher

Voxel51



Voxel51

Lead maintainers of the open source FiftyOne toolset



Innovators

Spun out of the Computer Vision Lab at the University of Michigan



Our Founders

Jason Corso, PhD
Professor of Computer Vision at the University of Michigan

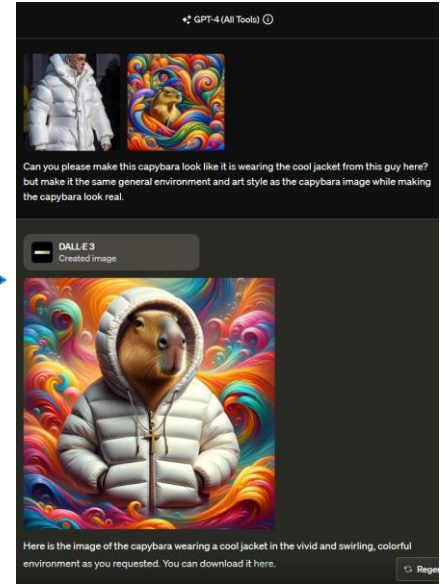
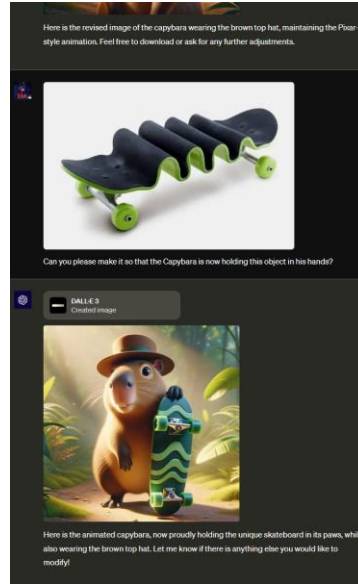
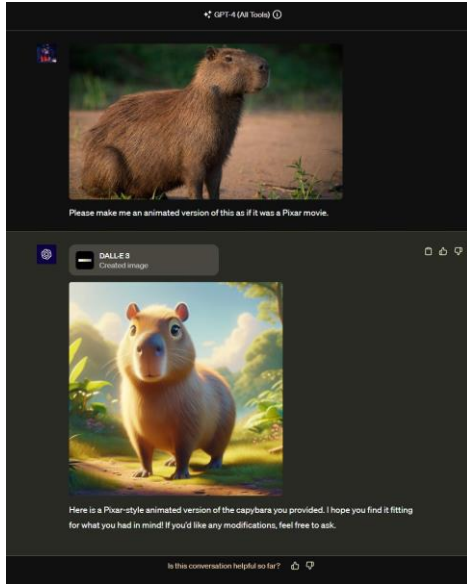
Brian Moore, PhD
Computer Vision Extraordinaire + Deep Expert in ML



Focused

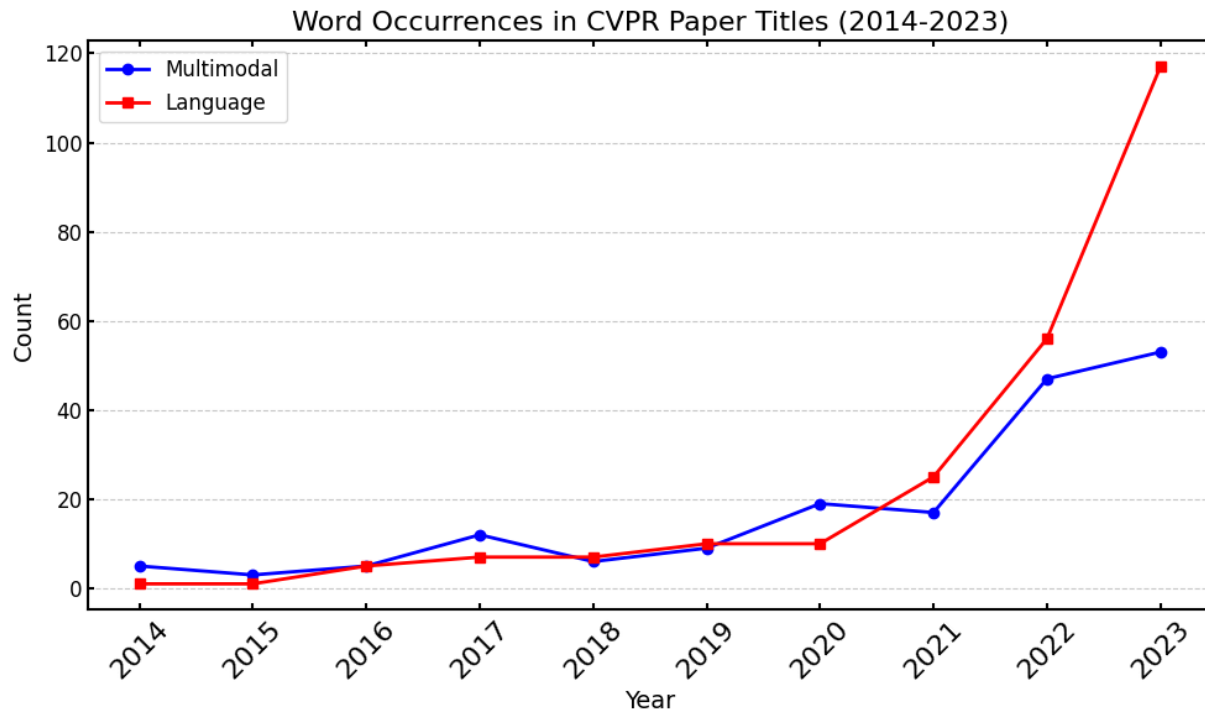
Team of 25+ driven, sharp, talented Computer Scientists and CV Experts supporting thousands of open source FiftyOne users and Fortune 100 companies

Multimodal Interactivity is Exciting



GPT-4 (All Tools)

“Vision” Increasingly Involves Language



What's Going On?

- ▶ Is this groundbreaking or hype?
- ▶ How are LLMs interfacing with vision?
- ▶ What can we expect moving forward?

Bridging the Modality Gap

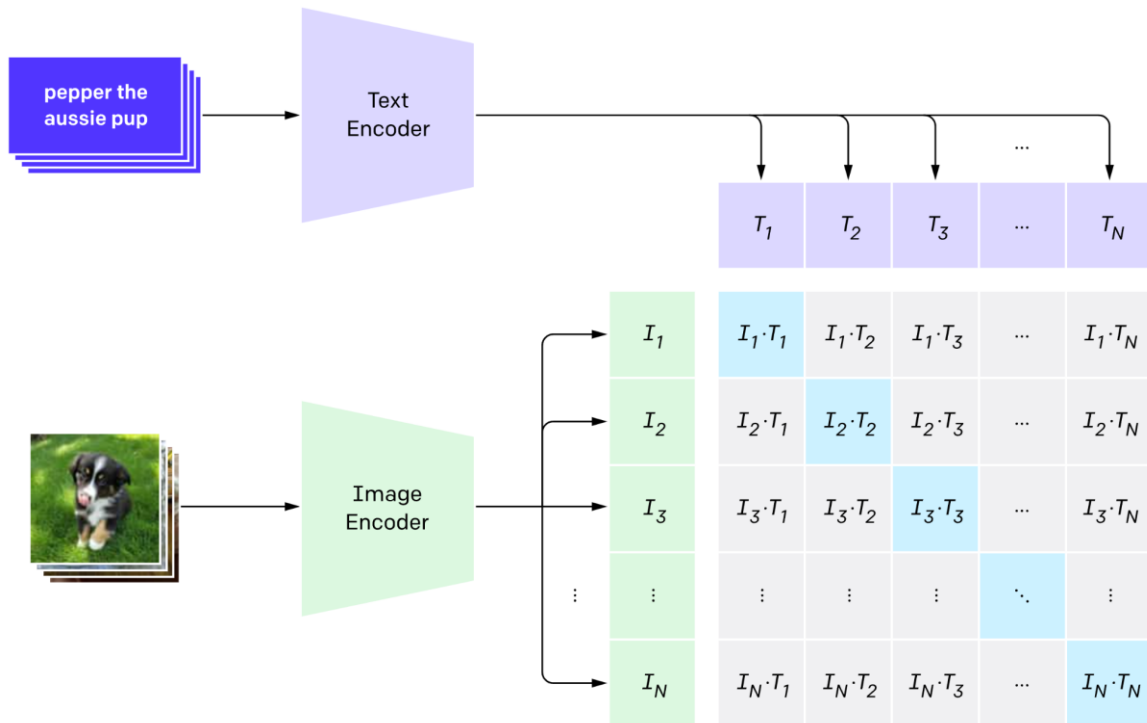


DALL-E3

Vision-Language Models

- ▶ Represent visual information in semantically meaningful way
- ▶ Train a multimodal model to ingest/interpret these representations

OpenAI CLIP Model



- ▶ Run vision models on visual data
- ▶ Pass results to LLM for analysis and conclusions

ViperGPT

Input Image:



Vision-Language Models

Where Vision-Language Models Thrive



LLaVA

- ▶ Open-world knowledge
- ▶ Rich semantic/visual content

User
LLaVA

Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

Challenges for Vision-Language Models

▶ Hallucination

▶ On-device

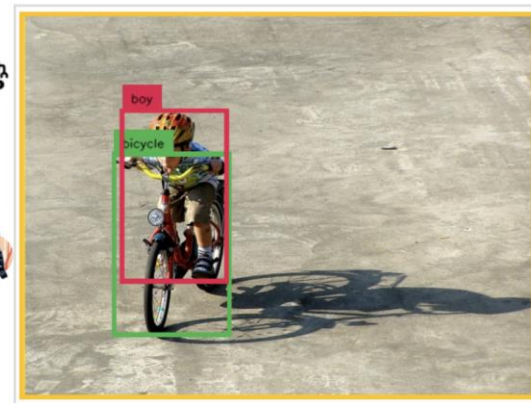
▶ Robust training



🗣️ Describe this image.

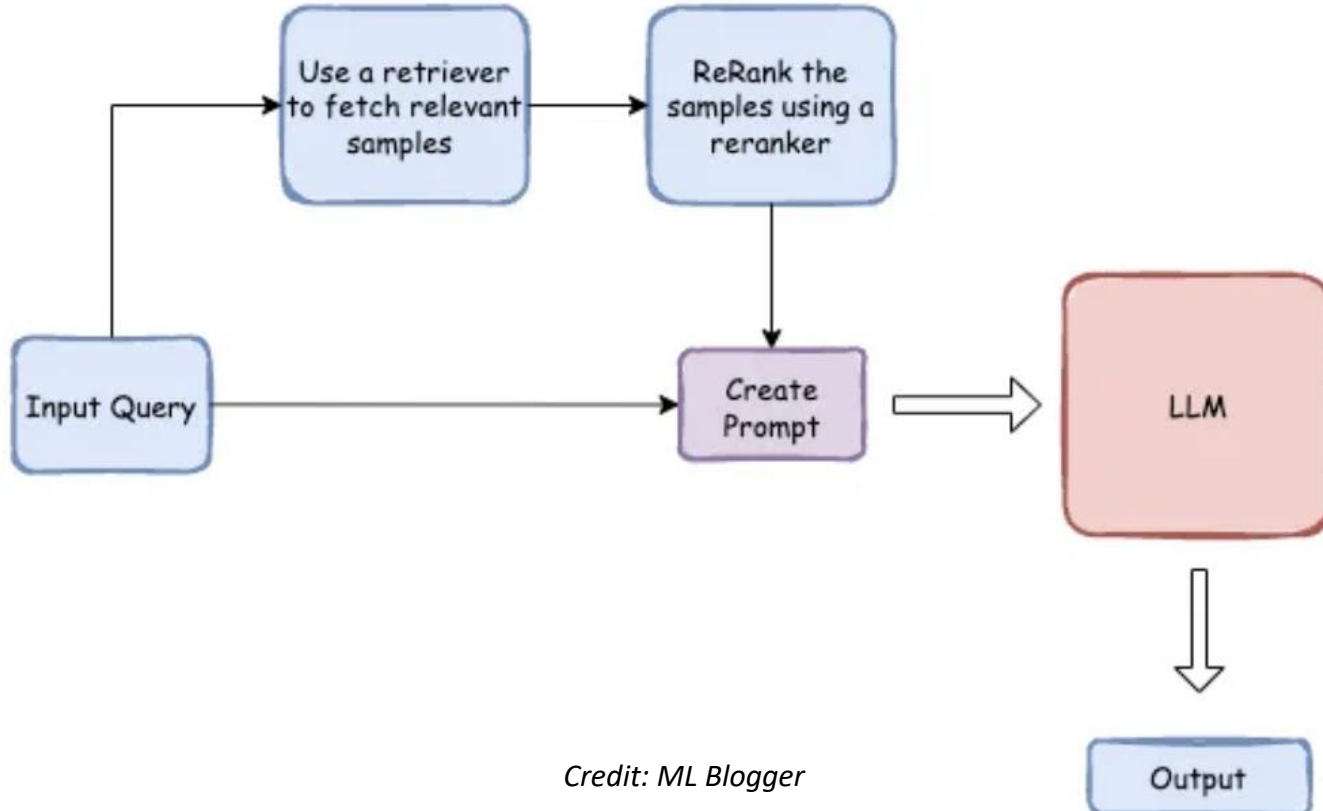
The image features a young boy riding a bicycle on a dirt road. There are several other people in the scene. A few of them are also riding bicycles, adding to the lively atmosphere.

The image features a boy([0.196,0.255,0.4,0.694]) riding a bicycle([0.179,0.372,0.404,0.841]) on a road([0.002,0.002,0.997,0.997]). There are no other people or bicycles in the image.



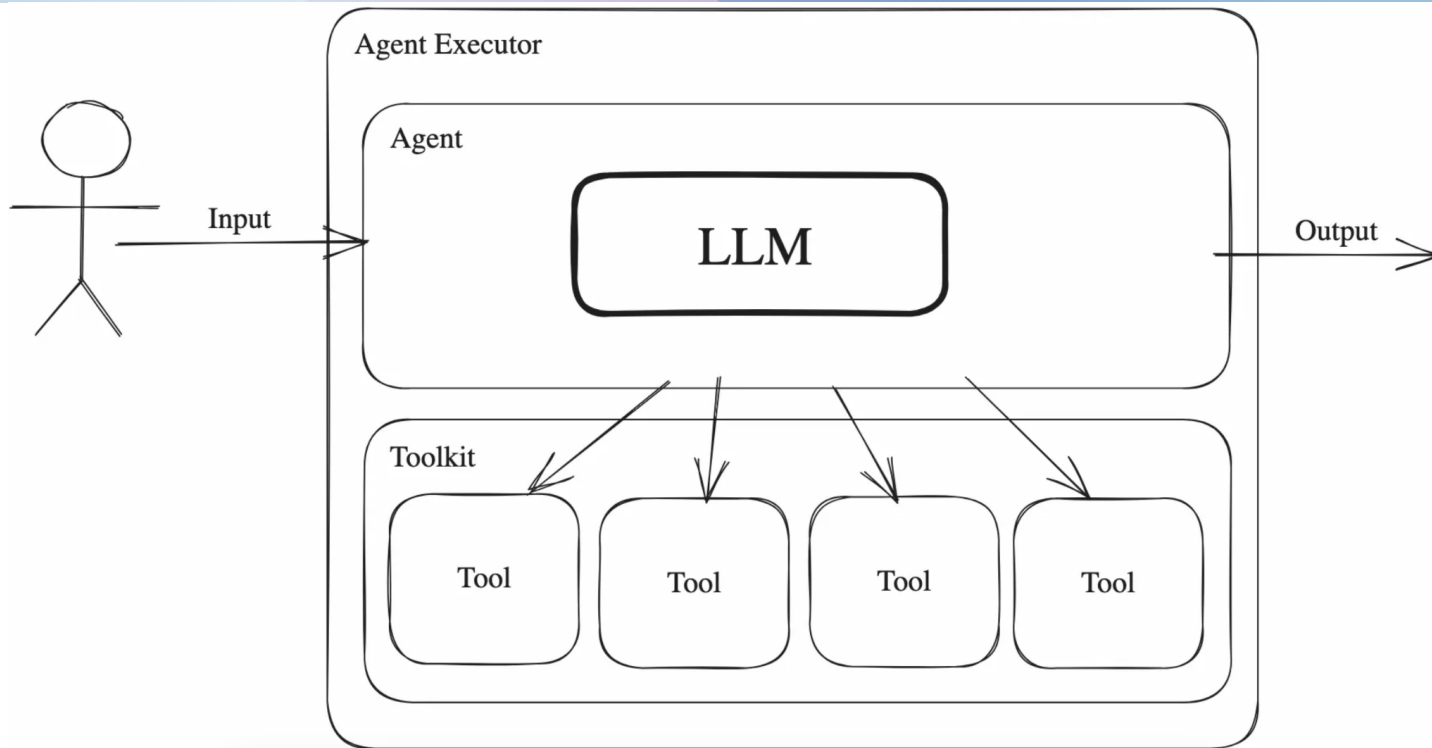
LLM-Aided Visual Reasoning

Basic LLM Information Flow



Credit: ML Blogger

Agentic Information Flow







Credit: Roger Oriol — GoPenAI

VisProg: Compositional Visual Reasoning

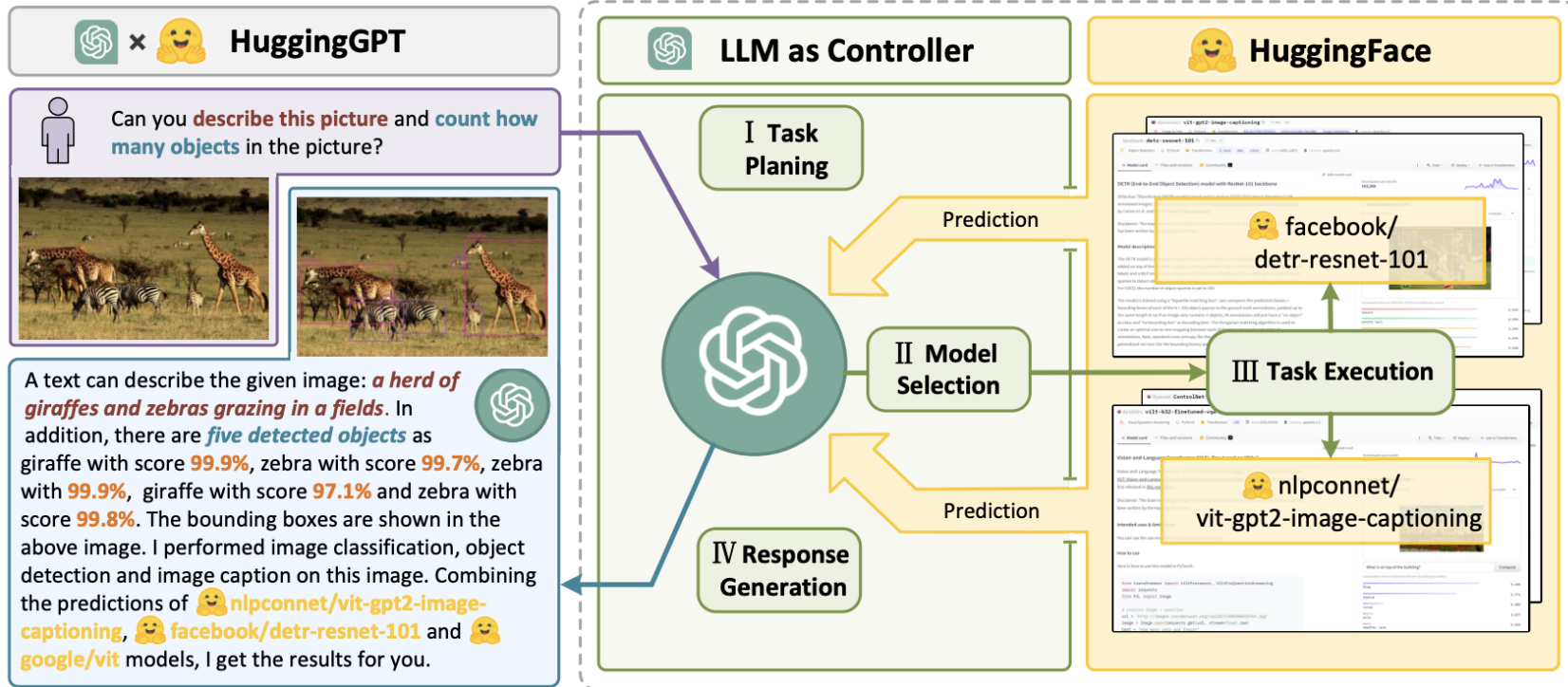
Object Replacement

Replace the desert by sandy beach



	IMAGE
	OBJ0=Seg(image=IMAGE)
	OBJ1=Select(image=IMAGE, object=OBJ0, query=['desert'], category=None)
	IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='sandy beach')

HuggingGPT: Control and Orchestration



Should I Use LLMs to Visually Reason

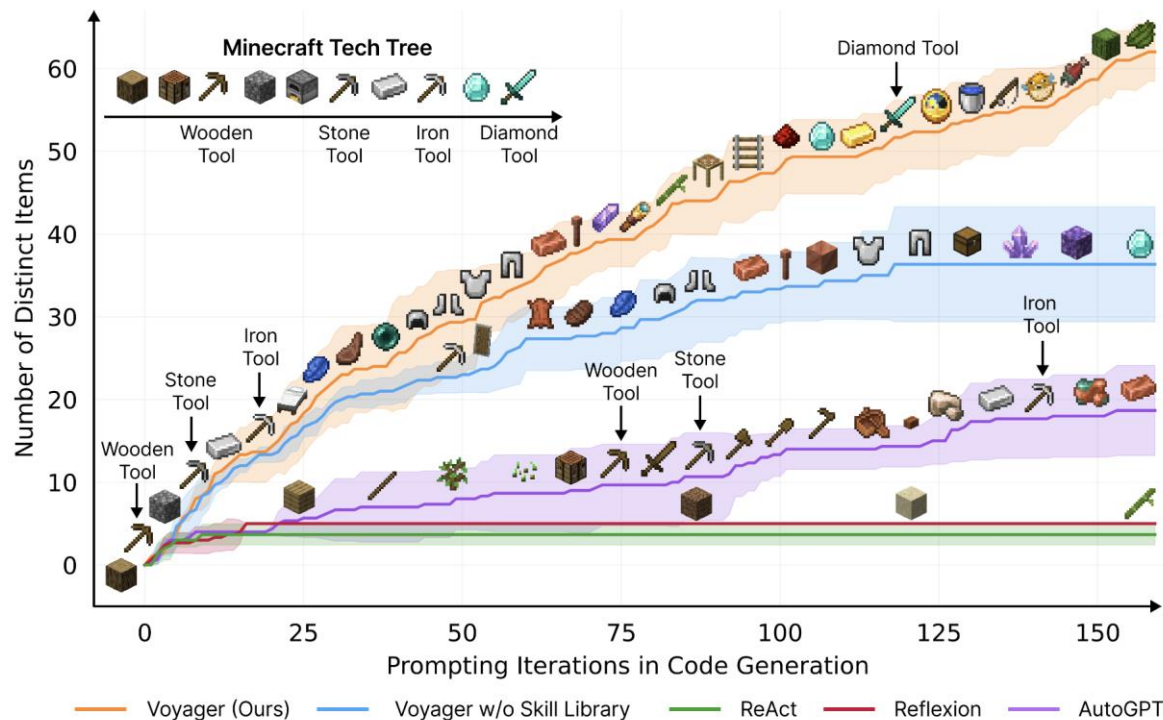
Pros

- ▶ Flexible & Modular
- ▶ Interpretable
- ▶ Training-free (using foundation models)

Cons

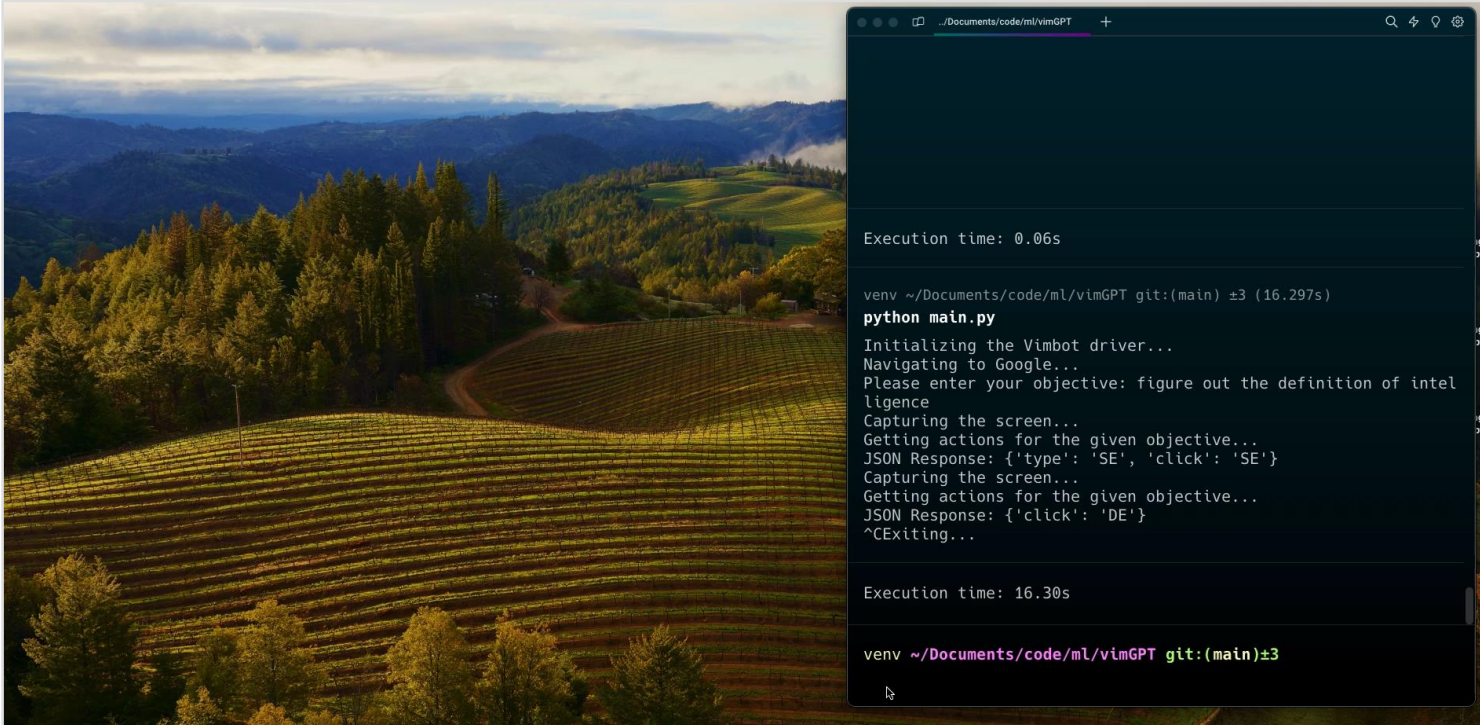
- ▶ Latency
- ▶ Reliance on Prompt Engineering
- ▶ Robustness & Reliability

Agents Can Acquire Skills



Voyager discovers new Minecraft items and skills continually by self-driven exploration, significantly outperforming the baselines.

Agents in the Field



VimGPT

- ▶ **Vision-Language** models promise to bring **open-world knowledge** to semantically rich visual tasks
- ▶ LLMs can visually reason by **orchestrating expert vision models**
- ▶ **Expect LLM-aided visual reasoning to proliferate** in the short term due to flexibility, modularity, and accessibility
- ▶ There is no silver bullet...yet

Projects Discussed

ViperGPT

<https://arxiv.org/abs/2303.08128>

VisProg

<https://prior.allenai.org/projects/visprog>

HuggingGPT

<https://arxiv.org/abs/2303.17580>

Voyager

<https://voyager.minedojo.org/>

Additional Reading

A Dive into Vision-Language Models

https://huggingface.co/blog/vision_language_pretraining

Multimodality and Large Multimodal Models

<https://huyenchip.com/2023/10/10/multimodal.html>

VoxelGPT

<https://github.com/voxel51/voxelgpt>

Thank You