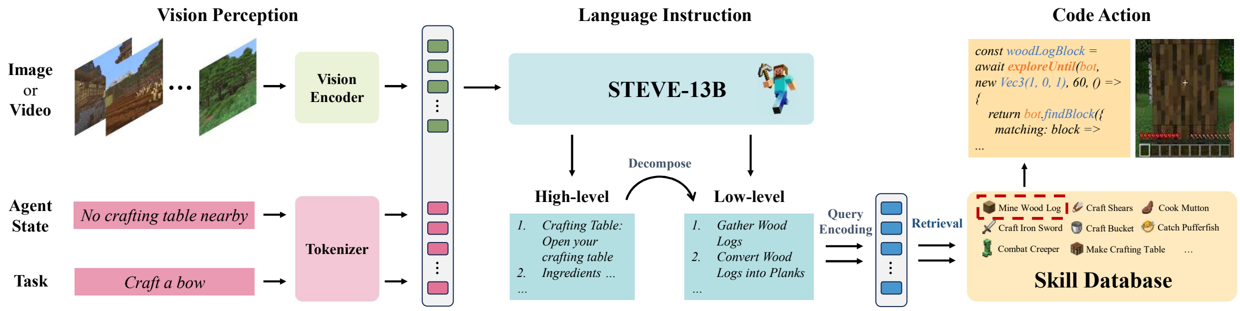
STEVE

See and Think: Embodied Agent in Virtual Environment

Past Work

We propose STEVE, a comprehensive and visionary embodied agent in the Minecraft virtual environment. STEVE consists of three key components: vision perception, language instruction, and code action. Vision perception involves the interpretation of visual information in the environment, which is then integrated into the LLMs component with agent state and task instruction. Language instruction is responsible for iterative reasoning and decomposing complex tasks into manageable guidelines. Code action generates executable skill actions based on retrieval in skill database, enabling the agent to interact effectively within the Minecraft environment. We also collect STEVE-21K dataset, which includes 600+ vision-environment pairs, 20K knowledge question-answering pairs, and 200+ skill-code pairs.

Related links. paper, code, website, demo video



```
const woodLogBlock =
await exploreUntil(bot,
new Vec3(1, 0, 1), 60, () =>
{
   return bot.findBlock({
     matching: block =>
...
```

Figure 1: **STEVE framework.** The Vision Perception part takes images or videos, encodes them into tokens, and combines them with the tokens of Agent State and Task as input. The STEVE-13B in the Language Instruction is for automatic reasoning and task decomposition, and it calls the Skill Database in the form of the Query to output code as action.

Future Work

We are currently adapting our framework to MetaGPT, which is a famous multi-agent codebase. After that, we will continue extend our work by follows. Each sub-direction will need **at least two** students involved.

Multi-agent. Build *Competition, Collaboration, Cooperation* relationship among agents.

Lifelong learning. Achieving long-term planning via efficient memory design.

Environment Feedback. Self-learning via visual and state feedback.

Complex tasks. Including creative and combination tasks, path optimization.