



Adapting Visual LLMs for gameplay in Pokémon Fire Red

Some thoughts on our works.

By Issac Lee, Erica Wu, Leo Mai

Dataset Creation

Stuff we
have to do

Dataset Creation

- ✓ Pull Youtube videos
- ✓ Figure out format
- ✓ Figure out how to caption
- ❑ Actually caption

Toolset

What tools were used?

- ❑ Qwen3vl-8B-Thinking
- ❑ yt-dlp
- ❑ Whisper
- ❑ Docker

Diagram



```
  "video_id": "video_0001",
  "clip_index": 21,
  "clip_file": "processed_videos/video_0001/clips/clip_0021.mp4",
  "start_time": 132.42,
  "end_time": 140.98,
  "duration": 8.560000000000002,
  "transcription": "hey you have Pokemon come on let's battle them all righty here we go bring it so here we go taking",
  "ai_caption": "***Caption:** \nThis segment shows **the player initiating a battle with an NPC trainer (Bug Catcher Ric) in FireRed gameplay**. Key elements include: \n\n- **Frame 1**: Player (red-haired sprite) stands face-to-face with an NPC (yellow-haired trainer) in a grassy area with trees; UI displays empty \"BADGES\" slots and the \"GLD\" logo\u2014no badges earned, no active battles or menu interactions. A text box shows *\"Hey! Y\"*\u2014indicating dialogue initiation toward interacting with the NPC. \n- **Frame 2**: Text box updates to *\"Come on! Let's battle 'em!\"*\u2014signaling the NPC\u2019s request to start a battle; player and NPC remain stationary, preparing for engagement. \n- **Frame 3**: Black screen overlay (likely YouTube video editing) obscures gameplay but indicates transition between frames; no in-game action here. \n- **Frame 4**: Battle interface appears: Player faces Bug Catcher Ric on an outdoor field; health bars (empty slots) and Pok\u00e9mon icons show readiness for battle. UI still shows empty badges, confirming early-game progression. \n- **Frame 5**: Battle screen details: Text box identifies the NPC as *\"BUG CATCHER RIC\"*; player\u2019s side displays Pok\u00e9mon slots with egg icons (no active Pok\u00e9mon yet), while opponent has a bug-type trainer sprite\u2014indicating the start of a battle sequence against this trainer. \n\nThe narrator\u2019s line (*\"hey you have Pokemon come on let's battle them all righty here we go bring it so here we go taking\"*) emphasizes **initiating a battle with an NPC trainer** (specifically Bug Catcher Ric), which is a standard early-game encounter where the player engages in their first combat. This segment focuses on **NPC interaction leading to battle initiation**, with no menu interactions or story progression beyond dialogue setup\u2014marking the transition from exploration to gameplay action.",
  "url": "https://www.youtube.com/watch?v=pb-a97oh-TY"
}
```

Problems

Some problems we faced

- ❑ Models are too big / slow
- ❑ Auto generated captions are undownloadable
- ❑ Have 219 videos with 100s to 600s clips

Fine Tuning

Stuff we
have to do

Fine Tuning

- ❑ Actually finalize the dataset
- ✓ Setup Supercomputer
- ✓ Figure out trainer
- ❑ Train!

Toolset

What tools were used?

- ❑ Qwen3vl-8B-Thinking / Instruct
- ❑ Ray
- ❑ Unsloth
- ❑ Docker

Diagram



Problems

Some problems we faced

- ❑ ParkPoint is double NATed
- ❑ Ray introduced a bug while setting up
- ❑ Random packet drops
- ❑ Setting up Unsloth is a pain

Implementation

Stuff we
have to do

Implementation

- ❑ Integrate Letta with the Framework
- ❑ Prompt Engineering
- ✓ Memory Setup
- ❑ Swap to Fire Red?

Toolset

What tools were used?

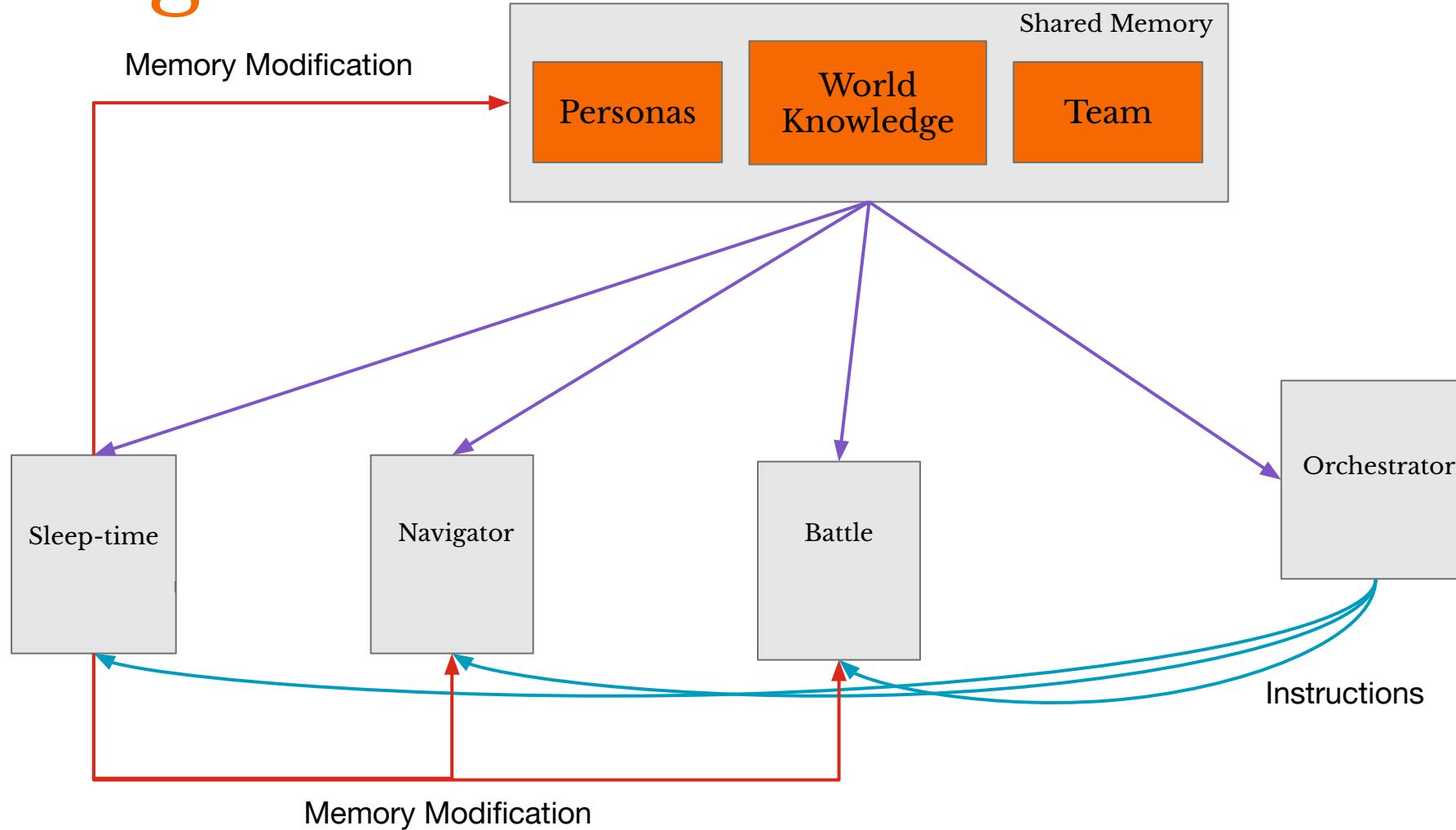
- ❑ Letta
- ❑ pokeagent-speedrun
- ❑ Qwen3-vl-8B-Thinking
- ❑ Qwen3vl_tools

Toolset

Our AI Agents

- ❑ Orchestrator Agent
- ❑ Navigator Agent
- ❑ Battle Agent
- ❑ Sleep-time Agent

Diagram



Examples



The title screen is showing copyright information. To begin playing, I need to advance past this sequence.

ACTION: A



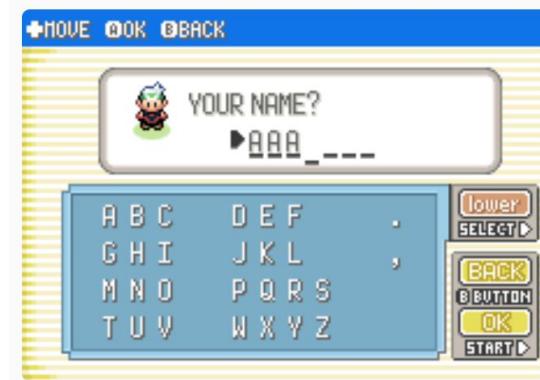
The screen displays 'Pokémon Emerald Version'. To move forward, I should interact with the logo.

ACTION: A



The game is now showing the main menu with 'NEW GAME' option visible. To start playing, I need to select this.

ACTION: A



The player has entered 'A' as their name but hasn't confirmed it yet.

To proceed with character creation, I need to press START to finalize the name selection.

ACTION: START

Getting here took 44.19k context!

Examples



10x speed
23min



200x speed 1hr
The big 30min to look left

Problems

Some Problems we've faced

- ❑ Getting stuck
- ❑ Too much context
- ❑ Input validation
- ❑ Not following prompts

Reminder of research questions

Research Questions

- ❑ How can a VLLM be adapted to play Pokémon Fire Red?
- ❑ Can fine-tuning on multimodal, game-specific data improve long-term planning and contextual reasoning?
- ❑ Are the learned strategies transferable to other interactive or sequential decision-making tasks?



Questions?