

---

# PLAICraft: Large-Scale Time-Aligned Vision-Speech-Action Dataset for Embodied AI

---

**Yingchen He**  
 University of British Columbia  
 mine01@cs.ubc.ca

**Christian D. Weilbach**  
 University of British Columbia  
 weilbach@cs.ubc.ca

**Martyna E. Wojciechowska**  
 University of British Columbia  
 martyna1@student.ubc.ca

**Yuxuan Zhang**  
 University of British Columbia  
 reacher@cs.ubc.ca

**Frank Wood**  
 University of British Columbia  
 fwood@cs.ubc.ca

## Abstract

Advances in deep generative modelling have made it increasingly plausible to train human-level embodied agents. Yet progress has been limited by the absence of large-scale, real-time, multi-modal, and socially interactive datasets that reflect the sensory-motor complexity of natural environments. To address this, we present PLAICraft, a novel data collection platform and dataset capturing multiplayer Minecraft interactions across five time-aligned modalities: video, game output audio, microphone input audio, mouse, and keyboard actions. Each modality is logged with millisecond time precision, enabling the study of synchronous, embodied behaviour in a rich, open-ended world. The dataset comprises over 10,000 hours of gameplay from more than 10,000 global participants.<sup>1</sup> Alongside the dataset, we provide an evaluation suite for benchmarking model capabilities in object recognition, spatial awareness, language grounding, and long-term memory. PLAICraft opens a path toward training and evaluating agents that act fluently and purposefully in real time, paving the way for truly embodied artificial intelligence.

## 1 Introduction

Embodied artificial intelligence (EAI) aims to build agents that perceive, act, and learn in dynamic environments through real-time sensory-motor interaction. As Brooks [1991] emphasized, such agents must “cope appropriately and in a timely fashion with changes in [their] dynamic environment” and “do something in [their] world; [they] should have some purpose in being.” These principles remain vital as research shifts toward agentic AI systems like Devin [AI, 2024] and Claude Code [Anthropic, 2024] that operate autonomously across extended tasks.

Yet most modern agents are *disembodied* in the sense that they interact with their environment indirectly, often through high-level text interfaces and asynchronous API calls. This limits their ability to engage in temporally grounded, perceptual-motor loops that characterize embodied intelligence. While simulated environments such as Habitat [Savva and et al., 2019], RoboTHOR [Deitke and

---

<sup>1</sup>We have done a privacy review for the public release of an initial 200-hour subset of the dataset, with plans to release most of the dataset over time.

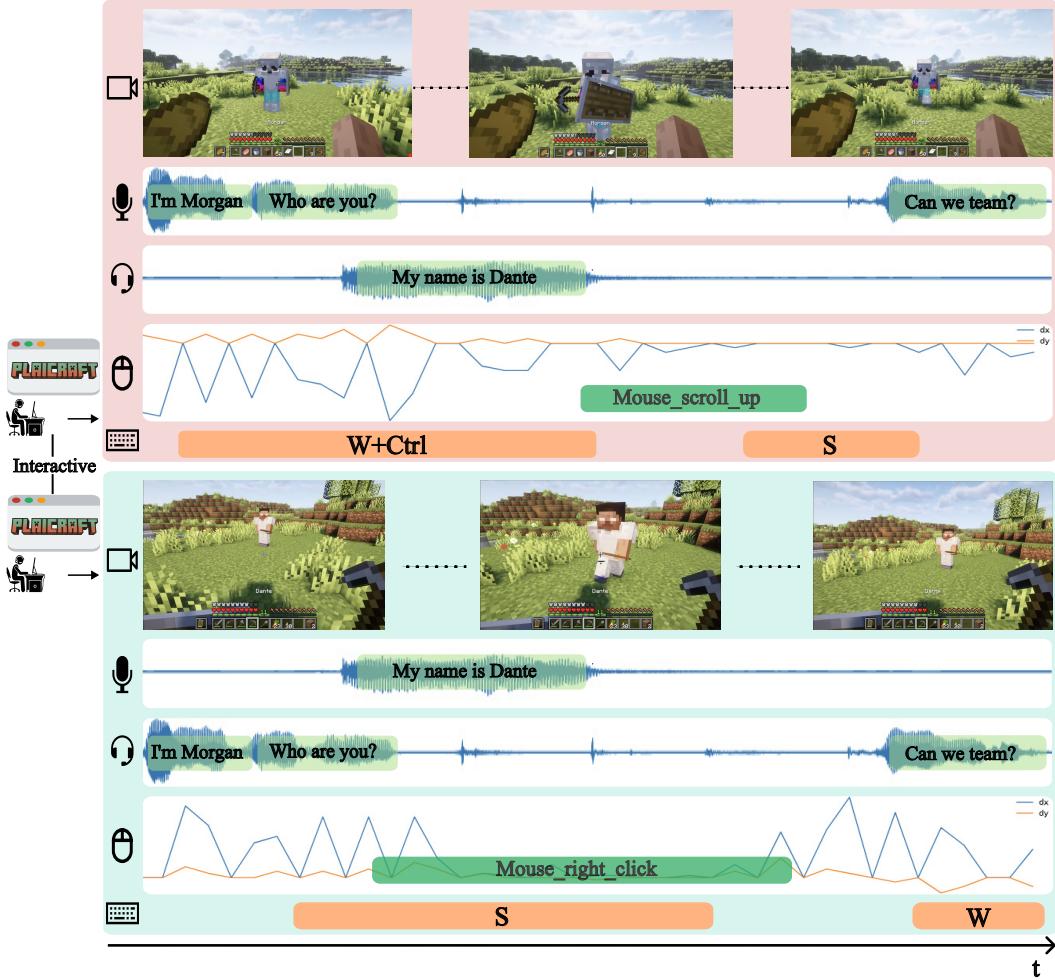


Figure 1: Illustration of [plaicraft.ai](#) dataset collection infrastructure with two, but often many, Minecraft players interacting with each other through our instrumented gameplay environment. Each player has five modalities: screen, microphone input audio, game output audio, mouse and keyboard. All interactions, including voice chat, are real-time and recorded with time alignment in millisecond precision between modalities. The players connect through a web browser to AWS EC2 Ubuntu instances running Minecraft with the instrumentation (Section 3).

et al., 2020], and ThreeDWorld [Gan and et al., 2021] support real-time, sensorimotor learning, they often lack socially interactive or linguistically rich settings. Within Minecraft, datasets like MineRL [Guss et al., 2019], OpenAI’s VPT [Baker and et al., 2022], and CraftAssist [Das and et al., 2019] have advanced perception-action and language learning. Most recently, MineDojo [Fan et al., 2022] introduced large-scale web-mined video-language data to support open-ended instruction following and knowledge transfer. However, these settings remain offline, non-interactive, heterogeneous, or lack real-time, social *grounding*—that is, the process of linking symbols to perceptual input, physical action, or shared environmental context.

We introduce **PLAICraft**, a dataset and platform designed for the study of real-time, socially situated, speech-interactive embodied agents. Built atop a modded multiplayer Minecraft server with a proximity-based voice chat plugin, PLAICraft records time-aligned streams of screen video, game output audio, player microphone input audio, keyboard, and mouse input with millisecond precision. Agents operate in persistent, open-ended worlds with other players where perception, speech, and action are tightly coupled in time. In its emphasis on rich, temporally grounded sensory-motor interaction, PLAICraft aligns with recent efforts to build datasets that support embodied learning in complex, naturalistic domains—such as autonomous driving benchmarks designed to capture diverse, real-world scenarios [Waymo, 2025, Zürn et al., 2024].

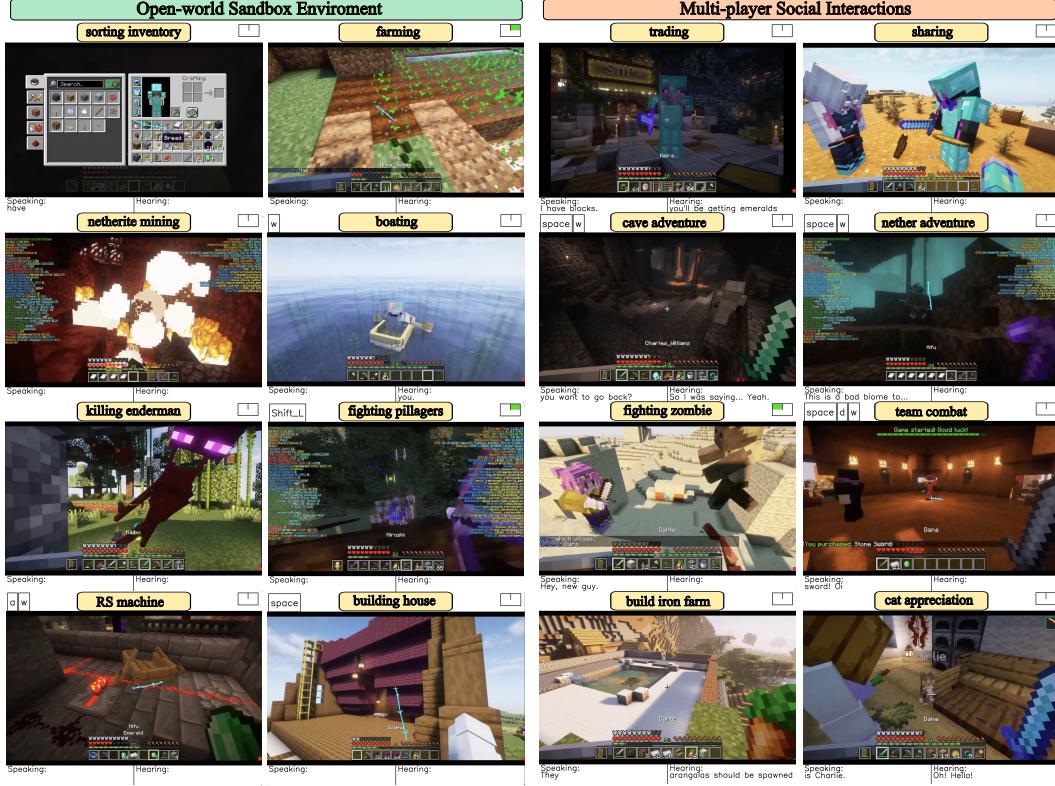


Figure 2: Visualization of the complex dynamics in the PLAICraft dataset. The frames here are a visualization that combines all modalities, not the actual data format, which stores each modality separately. Keyboard clicks are overlayed in the top left corner, and mouse clicks are overlayed in the top right corner. In the middle of the frame, the mouse movements are visualized by light blue arrows. To visualize speaking and hearing audio, we attached their corresponding transcript at the bottom of each frame.

Figure 1 illustrates our approach: participants interact naturally while all modalities are recorded. Unlike most prior datasets, PLAICraft emphasizes dialogue, social context, and temporal embodiment, supporting the study of skills like object permanence, reactive speech, memory, and social reasoning. Agents must act with purpose amid uncertainty, social influence, and ambiguity.

Following Brooks’ philosophy, we focus not on reducing complexity but embracing it—starting with large-scale imitation learning as a goal. Standard reinforcement learning approaches, when used in isolation, struggle in this setting. PLAICraft, like the real world, lacks explicit, global reward signals and is not structured around predefined episodic tasks. While the game mechanics do allow for death and respawn cycles, these do not constitute clean episodic resets from the agent’s perspective: the agent persists through its avatar’s deaths, and such events may even form part of emergent strategies or social play.

We have collected over 10,000 hours of data (Section 2), and publicly release a privacy-reviewed 200-hour subset (Appendix A.4) at <https://blog.plaicraft.ai/2025/05/11/download-user-data/>. We plan to release further data incrementally. The data collection platform is described in Section 3. Section 4 covers an evaluation suite designed to probe reasoning, memory, and communicative competence. We conclude in Section 5.

## 2 Plaicraft Dataset

We have chosen Minecraft as our simulated environment, like many previous datasets [Shah et al., 2021, Fan et al., 2022, Baker and et al., 2022, Das and et al., 2019], as its environment is sufficiently rich such that lessons learned within it can be transferred to the real world. In vanilla Minecraft, players navigate procedurally generated 3D worlds abundant with diverse terrains, resources to gather,

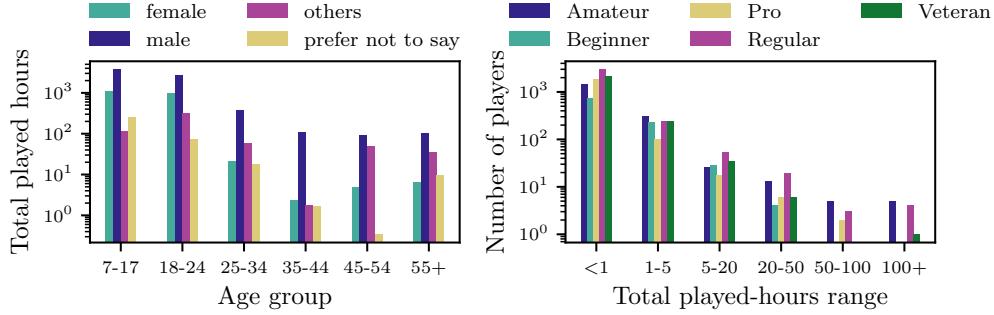


Figure 3: Players’ demographic distributions. Noted that all the players’ demographic information is provided voluntarily by themselves. **Left:** Played hours distribution over the gender and age groups. **Right:** Player count distribution by their played hours and experiences.

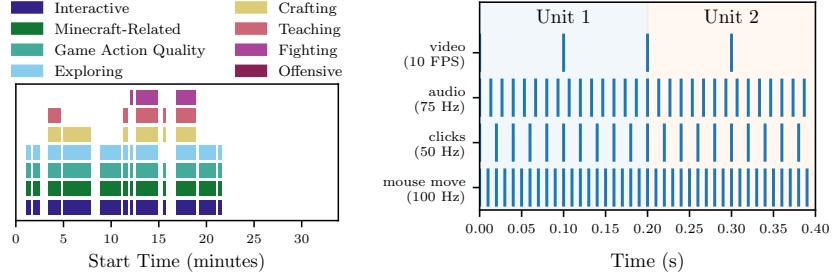


Figure 4: **Left:** Labelled segments within a session using an LLM on the audio transcripts (Appendix A.5). **Right:** Temporal alignment of the encoded data within a 400 ms window.

tools to create, structures to construct, and novel discoveries to encounter. The world players interact with is also boundary-free, allowing unrestricted exploration. Minecraft provides neither explicit goals nor predefined storylines, rendering it uniquely suitable as a foundation for open-ended embodied AI research. Unlike many prior Minecraft datasets, which are mostly restricted to single-player mode with limited sets of tasks [Shah et al., 2021], limited modalities or nonsystematic internet-fetched videos [Fan et al., 2022], our PLAICraft dataset is unique in the following ways.

**Unbounded Multiplayer Social Environment:** The open-ended nature of Minecraft inherently offers an unlimited spectrum of possible tasks. Our dataset further enhances this by employing a multiplayer environment where participants can socially interact with each other, and a proximity-based voice chat plugin for natural voice communication. This setup substantially enriches the complexity of achievable tasks, particularly enabling realistic social interactions. Figure 2 shows a wide range of solo and collaborative tasks players have performed on the server<sup>2</sup>.

Our participants comprise Minecraft enthusiasts globally, spanning diverse age groups, genders, and experience levels. Self-reported demographic statistics of participants are illustrated in Figure 3. During each gameplay session, participants receive no specific task assignments or restrictions regarding interaction areas. Instead, they are encouraged to play as naturally as they would, thinking of this not as a research project but just regular Minecraft, subject only to basic behavioural guidelines promoting mutual respect. The virtual microphone on the VM is forced to be enabled, and we encourage our players to continuously talk, especially when they play interactively with others. In each gameplay session, players have a maximum of 4 hours of play time, thus, they can perform a wide range of different tasks and social interactions. For example, as shown in the left figure of Figure 4, the player engaged in a wide range of interactive activities from minute 1 to minute 22, but no interactive audio was recorded afterwards.

Consequently, our dataset exhibits extensive dynamic variability. Solo player activities span a wide spectrum, beginning with basic actions such as moving with WASD keys or breaking a dirt block and progressing to moderately complex behaviours like mining, crafting, and combat. They extend further to highly advanced tasks, including defeating the Ender Dragon, constructing intricate redstone

<sup>2</sup>For video examples, please check our blog website: <https://blog.plaicraft.ai/>.

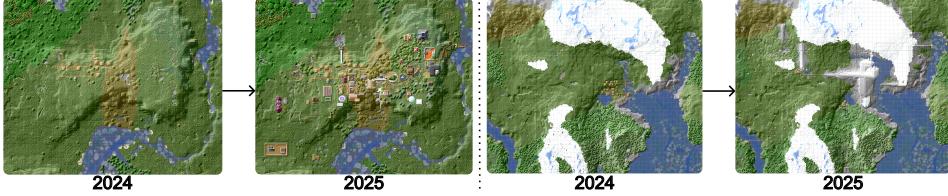


Figure 5: The birdview images showing the world state changes. **Left Two:** A 10-player base. **Right Two:** The world origin at  $x = 0, z = 0$ .

mechanisms, or even exploiting server bugs for personal benefit. Additionally, the dataset captures numerous social interactions, ranging from simple greetings and self-introductions to collaborative tasks involving team formation and cooperative gameplay. Notably, participants often form teams spontaneously, engaging in continuous interaction through collective base construction and various other tasks. Players between different teams interact with each other more cautiously and less frequently. Despite rigorous moderation efforts, occasional negative interactions such as conflicts, theft, and intentional property damage are present. Such behaviours are explicitly identified, and strong measures are implemented to prevent malicious participation.

**Millisecond-precision Time-aligned Multimodal Data:** The PLAICraft dataset comprises tens of thousands of gameplay session recordings, each representing one session of an individual player. For every session, we simultaneously capture multiple modalities, including screen recordings, game output audio, player microphone input audio, and mouse and keyboard interactions. These modalities are recorded with millisecond-precision timestamps, facilitating precise temporal alignment across data streams. Critically, the alignment between mouse and keyboard inputs and their corresponding visual and auditory outputs is strictly causal, ensuring input actions always precede or coincide with their resultant effects. During preprocessing, mouse and keyboard data are categorized into mouse & keyboard clicks and mouse movements. Click data includes discrete "PRESS" and "RELEASE" events, whereas mouse movement data captures continuous relative positional changes.

In addition to raw multimodal data, we provide encoded data representations generated by specialized autoencoders (detailed in Section 3). The right figure of Figure 4 illustrates the temporal alignment of these encoded formats. Given the differing sampling rates of video encodings (10 frames per second) and audio encodings (75 frames per second), each video frame corresponds to 7.5 audio frames. To resolve this fractional correspondence, we define the minimal encoding unit as two consecutive video frames (200 milliseconds), forming the basic processing unit used by our dataloader (see Appendix A.12).

**Continual global world state:** Throughout the entire duration of our project, the server has maintained a single, persistent world state that has evolved continuously since its creation a year ago, without resets. Consequently, player attributes—such as position, inventories, and experience level—carry over seamlessly between their gameplay sessions, unless they die in the game, then they will lose their inventories and experience if it is not picked up soon enough. Structures built by players endure indefinitely, and any environmental changes (for example, mined blocks) become permanent. Figure 5 illustrates the development of a communal player base constructed collaboratively by roughly ten participants over a one-year period. Thus, our dataset can also be considered as the continual history of the world’s evolution from the perspectives of more than 10,000 unique players.

### 3 Data Collection

Gathering gameplay data often involves developing specialized platforms tailored to specific games, each requiring unique strategies for user engagement and data collection. MineRL [Guss et al., 2019], for instance, utilizes a server and plugin system that players can access for gameplay. However, this platform mainly captures data within a predefined set of actions and tasks. On the other hand, the data set provided by MineDojo [Fan et al., 2022] is crawled from disjoint, pre-existing internet platforms such as YouTube, Wikipedia, and Reddit. While it contains a wide array of gameplay styles and strategies, it also introduces significant variability in data quality and distribution. The PLAICraft data collection platform addresses these limitations by standardizing the data collection process, ensuring the capturing of time-aligned video, game output audio, microphone input audio, keyboard

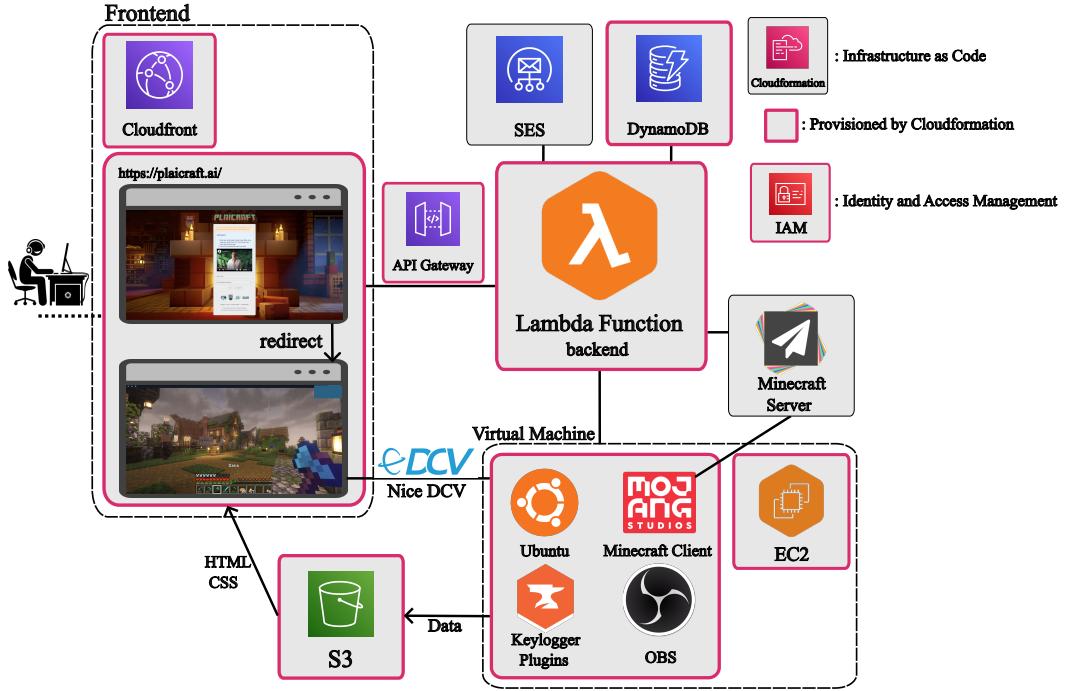


Figure 6: Illustration of [plaicraft.ai](https://plaicraft.ai/) data collection infrastructure. Participants can simply start a gameplay session like playing regular Minecraft, in any browser on any hardware device. Nice DCV will then connect users to a virtual machine that is running multiplayer Minecraft. On the EC2 instance, multiple pieces of recording software run in the background to record players’ video, audio and mouse & keyboard actions. Then the data will be saved to an S3 bucket. Our backend is centered around Lambda functions and multiple AWS resources.

and mouse action data while providing participants the freedom to naturally and collaboratively engage within a multi-player Minecraft setting.

### 3.1 Platform Setup

As illustrated in Figure 6, our platform is constructed mainly using Amazon Web Services (AWS). The central component is a remote-controlled virtual machine (VM) running as an AWS Ubuntu EC2 instance. Each session is hosted on a dedicated VM instantiated from an identical AMI image, ensuring secure, isolated, independent and uniform data recording, with automatic data upload upon session completion. The VM is equipped with [Amazon NICE DCV](#), a high-performance remote desktop and application streaming protocol, in conjunction with an NVIDIA T4 GPU. In combination, they provide a seamless Minecraft gaming experience featuring shader-enhanced visuals and high-quality, consistent and smooth video recordings. The platform is designed to immerse participants in gameplay upon connection to the EC2 instance, with the session concluding when the player stops gameplay. We ensure this by disabling any UI interaction players can perform with the desktop other than the Minecraft window. The Minecraft client launches in full-screen mode and cannot be minimized. Upon instance initialization, auxiliary monitoring scripts track both player connectivity and window focus; if they detect that the player has disconnected or shifted focus away from the Minecraft window, the scripts automatically terminate the EC2 instance and upload the recorded data to an Amazon S3 bucket.

Leveraging the Amazon DCV web frontend alongside AWS CloudFront, participants can effortlessly engage with the platform from any browser on any computer with a Chromium-based browser by logging into [plaicraft.ai](https://plaicraft.ai). We deployed our infrastructure at various AWS data center locations. Upon login, an instance is dynamically provisioned at the nearest available data center, ensuring stable connection speeds and low latency. Consequently, players experience consistently high-performance gameplay at a stable 60 FPS irrespective of their local hardware capabilities. Our platform offers a similar experience to popular cloud gaming services such as GeForce NOW and Xbox Cloud Gaming.

Moreover, as all recording processes occur within the secure environment of the virtual machine started from the same AMI image, participant privacy is fully protected, and the dimensionality and semantics of all modalities' data collected are uniform and fixed. Nonetheless, NICE DCV's constraints allow mismatched screen aspect ratios to appear as black borders in the recorded video, while the audio remains susceptible to background noise and the varying quality of participants' microphones.

### 3.2 Minecraft Configurations

To ensure game stability and compatibility with various modifications, we fixed the Minecraft version to 1.19.4 and retained most vanilla features. Players exclusively engage in survival mode with normal difficulty settings, and all three standard worlds (the Overworld, Nether, and the End) are fully accessible. To encourage social interactions, teleportation capabilities and a global chat interface are provided.

We incorporated multiple custom-developed and third-party plugins to enhance players' experience. Central to facilitating game play with over 10,000 participants is our AutoJoin plugin, which allows multiple participants to use a shared Minecraft license while preserving individual game states. Since providing unique licenses for all participants is impractical, this plugin enables us to maintain a minimal number of licenses while supporting an unlimited number of participants. The Autojoin plugin also automatically relocates first-time players to a safe position near the world spawn. To prevent localized overcrowding, we periodically shift the spawn point to new regions.

Another critical enhancement is the [Simple Voice Chat](#) plugin, a Spigot-based modification enabling immersive, proximity-based 3D voice communication within the game. Players can communicate naturally using microphones, with audio volume dynamically adjusting based on in-game distance and relative positioning. This feature substantially enriches player interactions, elevating the dataset's complexity and realism to a level comparable with real-world interaction datasets.

Finally, the inclusion of the [Complementary Shader Pack \(Unbound\)](#) significantly improves visual detail and complexity closer to real-world environments, enabling features like real-time shading, realistic clouds, water flow with reflections and a number of other high-quality effects. This add-on helped us attract a lot more participants and makes the dataset's visual space more realistic. A comprehensive list of applied plugins is provided in the Appendix [A.6](#).

### 3.3 Recording Software

**Video & Audio** Video recordings are captured using [Open Broadcaster Software \(OBS\)](#) running on the EC2 instance. Videos are recorded at a constant frame rate of 30 FPS with a resolution of 1280x720, encoded using the H.264 codec. OBS is specifically configured to capture only the Minecraft application window in full-screen mode, ensuring that all recorded frames pertain solely to gameplay. Audio streams are simultaneously recorded by OBS, separately capturing user microphone input and game audio output. Each audio stream is recorded at a constant sampling rate of 48 kHz in stereo format, encoded with the AAC codec.

**Mouse & Keyboard** Mouse and keyboard inputs are precisely captured using a custom-developed Forge mod integrated directly into the client-side Minecraft application. This method provides superior temporal accuracy, capturing server input event processing time compared to traditional system-level logging that is unaware of the Minecraft server. Input events for both mouse and keyboard are recorded on an event-driven basis with millisecond precision, maintaining a maximum polling rate of 100 Hz, resulting in a minimum interval of 10 ms between consecutive data points. Additionally, comprehensive system-level logging using [pyxhook](#) is implemented as a secondary data collection measure. These system-level logs, although less precise in timestamp accuracy, serve primarily to validate the integrity and completeness of the mod-generated input data.

### 3.4 Data Filtering and Pre-processing

We implemented comprehensive filtering and pre-processing strategies to maintain the integrity and quality of our dataset. Initially, real-time filtering was deployed on the Minecraft server through dedicated plugins for profanity detection and grief prevention, mitigating malicious behaviours and inappropriate content during gameplay. Post-recording, further pre-processing steps were conducted

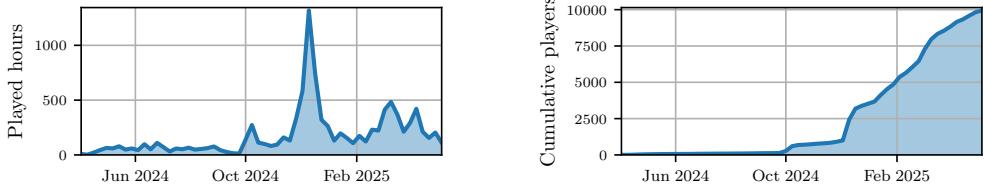


Figure 7: **Left:** Total hours collected per Week. **Right:** Cumulative contributors over time.

to ensure that all collected data pertained strictly to gameplay activities. All modalities underwent precise trimming aligned to gameplay start times, achieving synchronization at millisecond precision. For spoken audio streams, noise filtering using Voice Activity Detection (VAD) was employed alongside speaker diarization, ensuring audio recordings contained clear, isolated speech segments from a single speaker at any given time. Additionally, we also provide the transcription with word-level timestamps for both audio streams by running WhisperX [Bain et al., 2023] on the denoised version of the two audio tracks.

To optimize data handling and reduce GPU resource demand during model training, we pre-encoded the collected modalities. Video data were encoded using the `sdxl-vae-fp16-fix`, a variant of the SDXL Variational Autoencoder (VAE) [Podell et al., 2023] optimized for 16-bit floating-point precision, selected for its superior ability to preserve textual details in the game’s graphical user interface. Audio data of speaking and hearing was individually encoded using the `Encodex 24 kHz model` [Défossez et al., 2022], a state-of-the-art causal neural audio codec model with residual vector quantizers pretrained on a variety of audio data. Additionally, keyboard and mouse click data were compressed through a custom-trained temporal convolutional autoencoder with GRU layers to achieve dimensionality reduction. Given the inherently low dimensionality of mouse movement data, we opted against encoding it, choosing instead to simply apply global normalization.

### 3.5 Promotion Effort & Data Collection History

Hiring expert gamers to provide large amounts of demonstration data is usually costly and infeasible [DeepMind, 2019, Fan et al., 2022]. However, being the most popular game in the world in terms of sales [Wikipedia, 2025], Minecraft has a huge player community globally. All we need to do is reach out to this community and attract them to our platform. With an extensive media outreach and promotion efforts, see Appendix A.7, we have gradually formed a large community ourselves with Minecraft enthusiasts and experts. Figure 7 demonstrates our data collection history.

### 3.6 Ethical Considerations

This research ensures the safety and privacy of participants through the guidance of UBC’s Office of Research Ethics<sup>3</sup>. Study procedures are reviewed routinely by the Behavioural Research Ethics Board (BREB); and have been found acceptable on ethical grounds for research involving human subjects, as well as deemed minimal risk to participants. Participants’ data is collected with their consent, or that of their guardian, and any reasonably identifiable aspects of this data are anonymized systematically. We intend to release access to our full dataset with time, once participants have been given the opportunity to withdraw.

## 4 Evaluation

Motivated by Animal-AI [Voudouris et al., 2022], which adopts testing paradigms from comparative psychology; BIB [Gandhi et al., 2022], which assesses machines’ abilities to reason about other agents’ intentions by observing their actions; and the taxonomy of cognitive abilities outlined in Cattell-Horn-Carroll (CHC) theory [Carroll, 1993], we publish an evaluation suite that probes how agents perceive, understand, reason, and act within PLAICraft. The idea behind this design is intuitive: if an agent truly understands what it sees and hears, it should be able to (i) talk about it, (ii) act on it, and (iii) do so within a time-bounded, social context.

<sup>3</sup><https://researchethics.ubc.ca/>

Test Type	Prompt	Response	Metric
audio to action binding	Press A.	A pressed.	$D_K$
spelling ability	Please press the letter that ends the word STOP, as in S-T-O. What letter was that?	P pressed.	$D_K$
object recognition	What is this? What am I fighting?	Zombie.	$Acc_R$
world-model understanding	Right now, will the spider attack us?	No.	$Acc_R$
arithmetic	If you add up the blocks in the wool and granite pile, how many are there?	Five.	$Acc_R$
long-term memory	Have you seen this chest before?	Yes.	$Acc_R$
object constancy	Jump one time. Can you see me?	Yes.	$Acc_R$
spatial understanding	How many blocks away is the box?	Hundred.	$Acc_R$

Table 1: Representative evaluation tasks.  $D_K$ : Hamming distance between expected and actual keyboard events;  $Acc_R$ : exact-match accuracy on transcribed verbal responses.

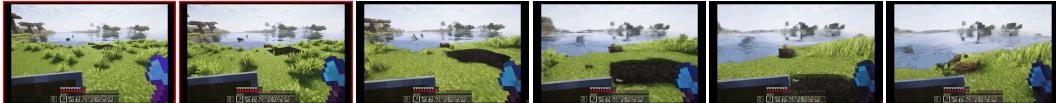


Figure 8: Sample from a trained model. It is conditioned on ground truth data for the first two frames (red), and then sampled the following four frames. Included with permission from Yoo et al. [2024].

Our evaluation suite contains carefully curated prompts. Each prompt maps to one of the ten broad cognitive abilities in CHC theory [Schneider and McGrew, 2018]: fluid intelligence, quantitative knowledge, crystallized intelligence, short-term memory, visual processing, auditory processing, long-term retrieval, processing speed, decision speed, and reading/writing. Each prompt elicits a keyboard/mouse action or a concise verbal answer, enabling evaluation via exact-match or semantic-similarity metrics.

During evaluation, the agent’s prompt consists of its entire lived life up to the current moment, including every observation and interaction so far, and is then followed by a spoken instruction that triggers the task. A correct response from the agent may be a mouse & keyboard action ("A pressed") or a brief verbal answer ("Zombie"). The test type categorizes the cognitive ability being assessed, while the metric is the measure used to judge responses’ correctness.

Table 1 selects eight representative prompts from the evaluation suite in Appendix B. The tasks span from visual object recognition ("What am I fighting?"), long-term memory ("Have you seen this chest before?"), to spatial estimation ("How far is the box?") and arithmetic challenges. These tasks require agents to integrate visual, audio, and action in a temporally grounded context. We hope this focus on semantically rich evaluation will encourage research toward EAI agents that reason and play like real people.

## 5 Conclusion & Discussion

In this paper, we present a novel data-collection platform and accompanying dataset that together create new opportunities for research on open-ended learning, reinforcement learning, complex social interaction, memory consolidation, and other related topics.

We do acknowledge that our work has some limitations, for example, the social and audio interaction portion of the dataset is still relatively small, as we give players freedom to not interact with other players if they don’t want to. We have limited deployment regions due to cost considerations and have to focus on English-speaking regions. The evaluation suite is still not comprehensive enough and needs further exploration. Despite the complexity of the dataset, it is still not a real-world dataset.

But we are actively developing to address these limitations, and this is just our first step. Looking ahead, we are leveraging this dataset to train human-like embodied agents in Minecraft that can converse and cooperate with other players in real time. One of the examples, as demonstrated in Figure 8, Yoo et al. [2024] leveraged a 50-hour subset of our dataset—collected from an anonymous player (“Alex”)—to illustrate that a video-diffusion model can be effectively trained from a continuous

gameplay stream in a continual learning setting [Wang et al., 2024]. Moving away from Minecraft, because the platform is flexible and modality-agnostic, it can also be repurposed with minimal effort to collect multimodal data from any computer workflow, for example, screen-reading digital assistants, and eventually towards robotics.

## References

- Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991.
- Cognition AI. Devin: The first ai software engineer. 2024. <https://www.cognition-labs.com/>.
- Anthropic. Claude code overview. <https://docs.anthropic.com/en/docs/clause-code-overview>, 2024. Accessed: 2025-05-15.
- Manolis Savva and et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019.
- Mitchell Deitke and et al. Robothor: An open simulation-to-real embodied ai platform. In *CVPR Workshops*, 2020.
- Chuang Gan and et al. Threedworld: A platform for interactive multi-modal physical simulation. In *NeurIPS*, 2021.
- William H. Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations, 2019.
- Bowen Baker and et al. Video pretraining (vpt): Learning to act by watching unlabeled online videos. <https://singularityhub.com/2022/06/26/openai-is-new-ai-learned-to-play-minecraft-by-watching-70000-hours-of-youtube/>, 2022.
- Abhishek Das and et al. Building modular and compositional agents with the craftassist framework. *Transactions of the ACL*, 2019.
- Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge, 2022.
- Waymo. 2025 waymo open dataset challenges: Vision-based end-to-end driving. <https://waymo.com/open/challenges/2025/e2e-driving/>, 2025. Accessed: 2025-05-15.
- Jannik Zürn, Paul Gladkov, Sofía Dudas, Fergal Cotter, Sofi Toteva, Jamie Shotton, Vasiliki Simaiaki, and Nikhil Mohan. Wayvescenes101: A dataset and benchmark for novel view synthesis in autonomous driving. <https://wayve.ai/science/wayvescenes101/>, 2024. Accessed: 2025-05-15.
- Rohin Shah, Cody Wild, Steven H. Wang, Neel Alex, Brandon Houghton, William Guss, Sharada Mohanty, Anssi Kanervisto, Stephanie Milani, Nicholay Topin, Pieter Abbeel, Stuart Russell, and Anca Dragan. The minerl basalt competition on learning from human feedback, 2021.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio, 2023. URL <https://arxiv.org/abs/2303.00747>.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL <https://arxiv.org/abs/2307.01952>.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL <https://arxiv.org/abs/2210.13438>.
- DeepMind. Alphastar: Mastering the real-time strategy game starcraft ii, January 2019. URL <https://deepmind.google/discover/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii/>. Accessed: 2025-05-09.

Wikipedia. Minecraft, May 2025. URL <https://en.wikipedia.org/wiki/Minecraft>. Accessed: 2025-05-09.

Konstantinos Voudouris, Matthew Crosby, Benjamin Beyret, José Hernández-Orallo, Murray Shanahan, Marta Halina, and Lucy G. Cheke. Direct human-ai comparison in the animal-ai environment. *Frontiers in Psychology*, 13, 2022. doi: 10.3389/fpsyg.2022.711821.

Kanishk Gandhi, Gala Stojnic, Brenden M. Lake, and Moira R. Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others, 2022. URL <https://arxiv.org/abs/2102.11938>.

John Bissell Carroll. *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge university press, 1993.

W Joel Schneider and Kevin S McGrew. The cattell-horn-carroll theory of cognitive abilities. *Contemporary intellectual assessment: Theories, tests, and issues*, 733:163, 2018.

Jason Yoo, Yingchen He, Saeid Naderiparizi, Dylan Green, Gido M. van de Ven, Geoff Pleiss, and Frank Wood. Lifelong learning of video diffusion models from a single video stream, 2024. URL <https://arxiv.org/abs/2406.04814>.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application, 2024. URL <https://arxiv.org/abs/2302.00487>.

Shalev Lifshitz, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila McIlraith. Steve-1: A generative model for text-to-behavior in minecraft, 2024.

Enshen Zhou, Yiran Qin, Zhen fei Yin, Yuzhou Huang, Ruimao Zhang, Lu Sheng, Yu Qiao, and Jing Shao. Minedreamer: Learning to follow instructions via chain-of-imagination for simulated-world control. *ArXiv*, abs/2403.12037, 2024. URL <https://api.semanticscholar.org/CorpusID:268532481>.

Stephanie Milani, Anssi Kanervisto, Karolis Ramanauskas, Sander V Schulhoff, Brandon Houghton, and Rohin Shah. BEED: The mineRL BASALT evaluation and demonstrations dataset for training and benchmarking agents that solve fuzzy tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=D1M0K2t2t2>.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4246–4247. AAAI Press, 2016. ISBN 9781577357704.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people, 2016. URL <https://arxiv.org/abs/1604.00289>.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su,

Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025a. URL <https://arxiv.org/abs/2412.19437>.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shufeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025b. URL <https://arxiv.org/abs/2501.12948>.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers, 2023. URL <https://arxiv.org/abs/2304.09116>.

Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion, 2024. URL <https://arxiv.org/abs/2402.04825>.

## A Appendix

### A.1 Acknowledgment

We are very grateful to Alexander Liteplo, Alex Wu, Andrew Smith, Andy Huang, Brooke Dai, Chengyuan Yao, Christopher Tardy, Daniel Crookall, David Tianyi Yin, David Yang, Edward Liang, Geo Lee, Hanson Sun, Joey Xiang, Lucas Qin, Mantaj Dhillon, Mehdi Safaei, Naveed Ghassemi, Perry Zhu, Sean Chuah, and Suzette Sun for their help in the development of the platform. We thank Alice Xia, Hotslicer Media, and many others for their promotion efforts. We also appreciate the help with server moderation and community organization by David Yang, Piper, Delara, and Hiroshi (the last three are anonymous players on the server who spontaneously offered to help with moderation). In addition, we would like to send a special thanks to all the amazing players who have spent countless hours on our server, enjoying Minecraft, building a wholesome community and contributing to our research.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, Inverted AI, MITACS, and Google. This research was enabled in part by technical support and computational resources provided by the Digital Research Alliance of Canada Compute Canada (alliancecan.ca), the Advanced Research Computing at the University of British Columbia (arc.ubc.ca), and Amazon.

### A.2 Related Work

**Minecraft for Machine Learning** Minecraft serves as an important tool for machine learning research [Lifshitz et al., 2024, Zhou et al., 2024] because of its open-world nature and intricate environment, along with a customizable game engine. Consequently, numerous datasets and benchmarks [Guss et al., 2019, Milani et al., 2023, Fan et al., 2022] have been established on Minecraft to enhance AI abilities. Microsoft’s first release of MALMO [Johnson et al., 2016] platform garnered great research interest in Minecraft. Based on MALMO, MineRL [Guss et al., 2019] introduces a complex environment for reinforcement learning tasks using human demonstrations in Minecraft. MineDojo [Fan et al., 2022] extends the capabilities by aggregating gameplay data from a variety of internet sources and providing a more expansive and flexible framework that allows for the exploration of a wider range of tasks and environments within Minecraft. BASALT competition [Shah et al., 2021] applies human evaluation to 4 open-ended tasks, and build upon that, BEDD [Milani et al., 2023] introduces a benchmark environment that systematically measures agent performance across diverse, complex scenarios

**EAI Evaluation** Evaluating EAI agent performance in complex environments effectively requires direct comparisons between human and AI abilities. There is a growing literature on such benchmarks, including the Animal-AI Environment framework which has been used to compare human and AI performance on cognitive tasks inspired by animal psychology, providing insights into how agents reason and generalize [Voudouris et al., 2022]. Similarly, benchmarks such as the Baby Intuitions Benchmark (BIB) assess agents’ ability to infer goals, preferences, and the actions of others, highlighting gaps in AI’s understanding of intuitive physics and social reasoning compared to humans [Gandhi et al., 2022]. These approaches underscore the challenge of developing evaluation frameworks that align with human cognitive benchmarks [Lake et al., 2016].

### A.3 Societal Impacts

Combining embodied AI with immersive game environments has the potential to amplify both the benefits and risks of artificial intelligence. On the upside, interactive gameplay environments like Minecraft provide a uniquely rich substrate for training and evaluating agents with social, linguistic, and perceptual competencies. Agents trained in such environments could support educational gameplay, serve as in-game tutors, or scaffold more collaborative and engaging human experiences. As embodied agents become more fluent in social cues—especially through real-time speech and interaction—they may also help simulate realistic training scenarios or serve therapeutic and accessibility functions.

At the same time, advances in speech-enabled embodied agents may blur the line between artificial and human players. Players may not always be able to distinguish between synthetic and human teammates or adversaries. While this ambiguity can enhance immersion, it may also raise concerns

about deception, manipulation, or the erosion of trust. Agents that convincingly imitate social behaviours could influence players in subtle ways, especially younger users. There is a risk that such agents could be used to simulate companionship without reciprocity, or to manipulate attention and emotional responses for commercial ends.

The collection of large-scale, multimodal human gameplay data also raises important questions about consent, privacy, and downstream use. We have undergone an institutional privacy review and obtained consent for data release, and have designed our infrastructure to preserve anonymity and prevent the recording of persistent identity information. Nonetheless, the richness of multimodal recordings—particularly audio—makes full anonymization difficult. Developers and researchers using such data should exercise caution in avoiding reidentification risks and consider the broader implications of using data collected in seemingly informal or playful contexts to train powerful AI systems.

Finally, embodied AI in gaming environments may raise labour and cultural concerns. As agents become capable of performing in-game roles—e.g., assistants, performers, or economic participants—they could displace human workers in virtual economies, reshape online social norms, or saturate game environments with synthetic interactions. While our work is aimed at research and foundational exploration, we believe that careful governance, transparency, and participatory design will be essential as embodied agents become more integrated into human play and work environments.

#### A.4 Initial Release

Our initial release, pending the internal review process, is a 200-hour dataset. The rest of the dataset will be released over time, once they are fully vetted and participants have been given the opportunity to withdraw. To prove the existence of the full dataset, we also provide the croissant metadata for the full dataset, available to download via: <https://blog.plaicraft.ai/2025/05/11/download-user-data/>. Video examples from the full dataset are also given on the same blog website.

The 200-hour dataset contains the complete recordings of 3 players: Dante, Morgan and Xander. They played together as a team in the same player base throughout their time on the server, interacting constantly in most of their gameplay sessions. Thus, this subset is a perfect example of the rich social dynamics in our dataset. Dante self-identifies as a 49-year-old male player with "Regular" level experience, Morgan self-identifies as a 9-year-old male player with "Pro" level experience and Xander self-identifies as a 37-year-old female player with "Amateur" level experience.<sup>4</sup> Within their 200 hours of gameplay, they performed a wide range of activities, including various types of team collaborations like base construction, biomes exploration, combating each other, building iron farms, etc. Some data examples are shown in the Figure 9.

We released the croissant files for both the 200-hour dataset and the full 10,000-hour dataset, and released the raw and processed data of the 200-hour dataset at <https://blog.plaicraft.ai/2025/05/11/download-user-data/>.

#### A.5 Automatic Data Annotation

To identify subsets of highly interactive data, particularly relevant for Embodied AI, we developed a methodology leveraging audio transcribed by WhisperX [Bain et al., 2023], and the capabilities of large language models (LLMs). We first defined highly interactive data as continuous, game-related interactions between at least two players, within a specific time window. Notably, a key indicator of interactivity was the presence of rich, conversational audio. We used WhisperX-generated audio transcripts to create a dataset and identify these interactions. To preserve conversational coherence, we split the transcript into segments such that a new segment begins after a pause of at least 10 seconds between utterances. These segments represent potential areas of interactivity and Minecraft social interactions. Given the scale of the resulting dataset, manual annotation of segments was infeasible. Instead, we leveraged recent advances in large language models to perform automated annotation. A binary classification approach using a 4-shot prompt demonstrated high accuracy and efficiency on a validation subset. Based on this evaluation and overall cost considerations, we used

---

<sup>4</sup>Again, we do not guarantee the truthfulness of these information as players stay 100% anonymous during participation and information is provided completely voluntarily.

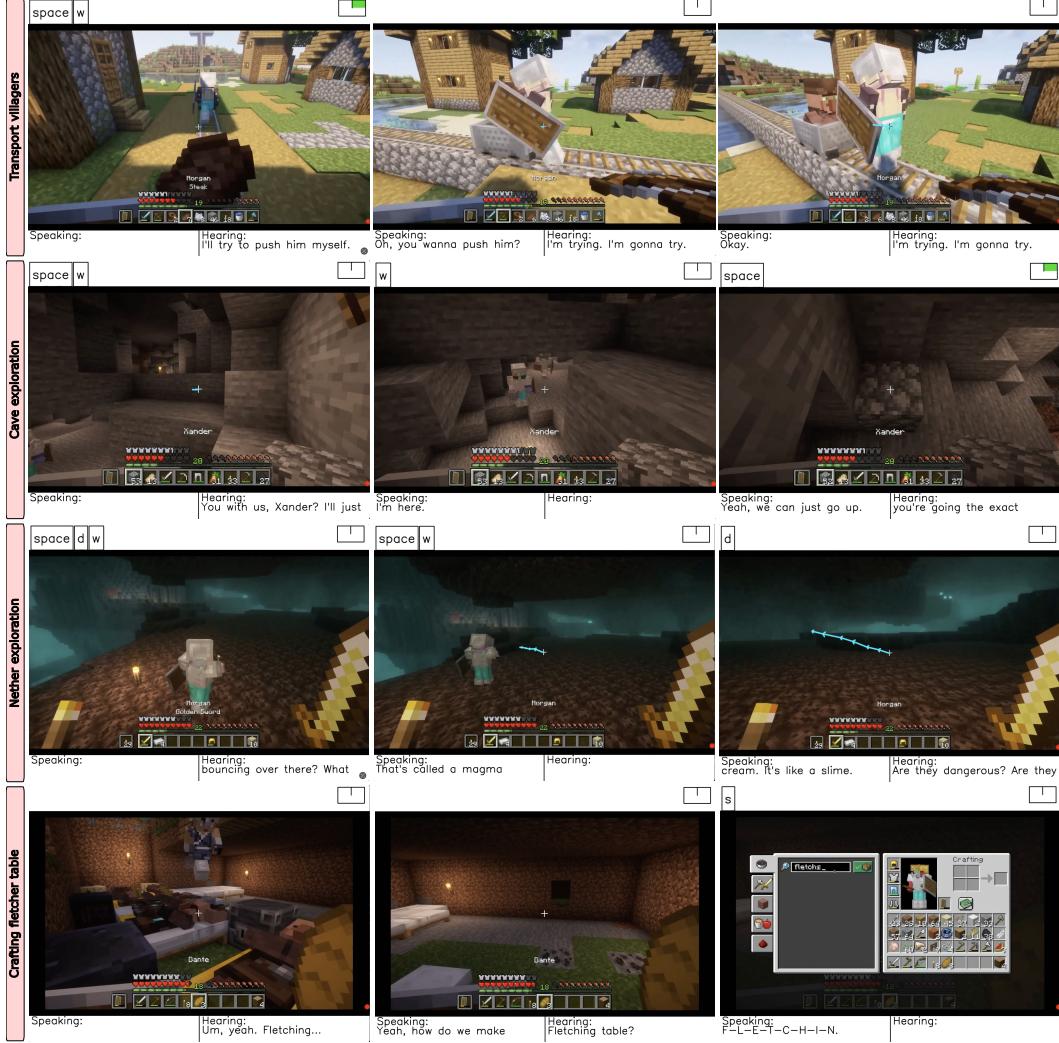


Figure 9: Examples of recordings played by Dante, Morgan, and Xander.

Label	Question
Interactive	Is this segment interactive?
Minecraft-related	Is this segment related to Minecraft gameplay?
Offensive	Does this segment contain offensive or inappropriate language?
Crafting	Are the players building or crafting something together?
Exploring	Are the players actively navigating the Minecraft world?
Teaching	Is one player helping another learn something about the game?
Fighting	Are the players involved in or discussing combat?
Game Action Quality	Is this segment a high-quality example of a gameplay interaction?

Table 2: Examples of Binary Labeling Questions

models such as DeepSeek-V3 [DeepSeek-AI et al., 2025a] and DeepSeek-R1 [DeepSeek-AI et al., 2025b] to label the full dataset with binary annotations. Table 2 demonstrates the labeling scheme. Segments assigned specific labels were then identified as highly interactive and selected for further analysis.

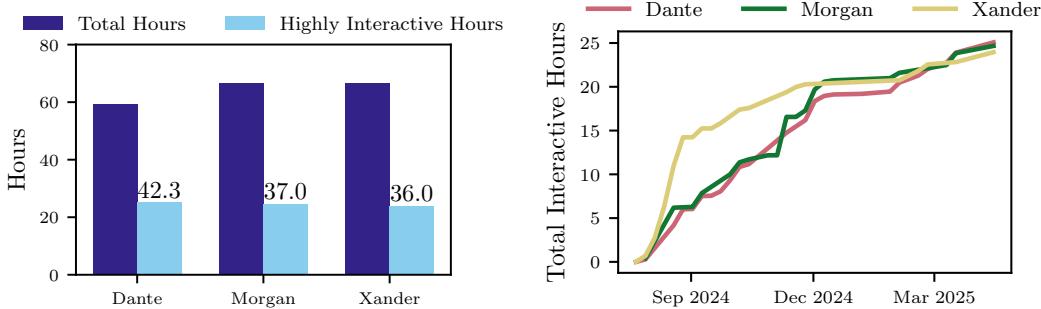


Figure 10: (Left) Total vs. highly interactive hours per player. (Right) Weekly accumulation of highly interactive hours per player.

## A.6 Minecraft Configurations

To ensure stability and mod compatibility, we standardized the Minecraft version to 1.19.4 while retaining most vanilla features. Similar to MineDojo, all three worlds (the Overworld, the Nether, and the End) are accessible to players. In addition to the standard Minecraft setup, we implemented several client-side enhancements:

1. We added several new worlds dedicated to minigames, isolated from the vanilla worlds. In these worlds, players can engage in a variety of minigames, such as Bedwars.
2. We restricted access to in-game settings to prevent players from exiting the game; should a player leave the Minecraft window, the instance is shut down, and the data recorded up to that point is uploaded, ensuring that all captured data consists solely of gameplay.
3. We implemented a mod to capture players’ keyboard and mouse actions with a more precise timestamp than system-level logging, bringing the recorded actions closer to the Minecraft server tick and thus better aligning cause and effect with rendered frames.
4. We incorporated a range of third-party client-side mods and plugins to enhance performance, visuals, and playability:

Name	Description
Complementary Shader Pack	Significantly enhances visual detail and overall graphical quality.
Simple Voice Chat	Proximity-based voice chat enabling real-time communication.
Canary	Performance optimization mod for general client-side speedups.
Clumps	Reduces item-entity lag by grouping dropped items into stacks.
EntityCulling	Skips rendering of off-screen entities to improve FPS.
Fastload-Reforged	Accelerates startup and world-loading times.
FerriteCore	Minimizes memory usage by stripping unused block states.
ModernFix	Addresses rendering performance issues and chunk-loading bugs.
Oculus	Optimizes block-update and rendering code paths for better FPS.
Redirector	Replaces generic implementations with faster runtime alternatives.
Rubidium	Speeds up core rendering operations via low-level optimizations.
Saturn	Bundles multiple performance tweaks for smoother gameplay.
SmoothBoot	Optimizes boot routines for faster game initialization.
Starlight	Overhauls the lighting engine for much faster chunk light updates.
BetterF3	Enhanced debug overlay with customizable performance stats.
Cloth-Config	Provides an in-game, user-friendly configuration GUI.
Entity-Model-Features	Adds extra model customization hooks for entities.
Entity-Texture-Features	Enables per-entity texture overrides and enhancements.
NotEnoughAnimation	Introduces additional character animations for more realism.
SkinLayer3D	Renders player skin layers in true 3D for added depth.

Table 3: Client-Side Mods & Plugins

On the server side, adjustments include:

1. A custom mod, AutoJoin, designed to manage player dynamics. We store players' state (inventory, level, location, etc.) in our own database. Every time a player joins, a random license will be selected and loaded. Upon joining, the plugin fetches the player's state, overwrites the license's state with it, and teleports the player to their last location. When a player leaves, their game state is saved back into the database, giving the player a seamless experience while allowing us to support more concurrent users.
2. To encourage social interactions, we added a few new worlds in addition to the default three worlds: A lobby world where players can access at any time by entering the command "/minigames", this is a containerized world with boundaries where people can safely meet up with other players. From the lobby world, players can choose to go to two other worlds: a world integrated with [BedWars](#), one of the most popular Minecraft minigames, and a world integrated with [Murder Mystery](#), another popular minigame. Both of these game modes are added in the hope of increasing the number of participants and players' interactions.
3. Various third-party server-side plugins are added:

Name	Description
<a href="#">AFK-Kick</a>	Prevents idle players by kicking after inactivity.
<a href="#">BetterStructures</a>	Adds new, interesting auto-generated structures.
<a href="#">Better-RTP</a>	Enhances random teleportation experience.
<a href="#">ChatFilter</a>	Filters and blocks offensive chat content.
<a href="#">Screaming Bedwars</a>	Implements Bedwars minigame mechanics.
<a href="#">Chunky</a>	Preloads chunks to reduce lag spikes.
<a href="#">CoreProtect</a>	Logs and rolls back griefing actions.
<a href="#">DecentHolograms</a>	Displays floating holograms without dependencies.
<a href="#">DiscordSRV</a>	Bridges Discord and Minecraft chat.
<a href="#">EssentialsX</a>	Provides core server management commands.
<a href="#">GravesX</a>	Spawns lootable death chests at player death.
<a href="#">HideNametag</a>	Removes player nametags for clean visuals.
<a href="#">InvisibleItemFrames</a>	Hides item frame visuals in-game.
<a href="#">Multiverse-Core</a>	Manages multiple worlds on one server.
<a href="#">PlaceholderAPI</a>	Provides a uniform placeholder system for other plugins.
<a href="#">ProtocolLib</a>	Allows manipulation of Minecraft network packets.
<a href="#">SuperVanish</a>	Enables admins to be invisible to players.

Table 4: Server-Side Plugins

Collectively, these modifications create a natural, intuitive, and enjoyable environment where participants can play Minecraft as they normally would, while also providing us with standardized, time-synchronized, multi-modal gameplay data.

#### A.7 Additional Participants Demographics

Our participants consist of Minecraft players across the world. To lower the overall latency players will experience, we deploy our EC2 instances in various locations across the world. Figure 11 lists out our deployment regions.

We do not hire specific participants to conduct our data collection process, as it is financially impractical, nor do we restrict anyone from participating. Instead, we focus on promoting our platform as a research project where participants can play free Minecraft with beautiful shaders on any device and make it available to everyone. We have done a series of promotions, including local poster distribution, Google Ads, YouTube and TikTok content creation. As a result, we get a constant stream of old and new participants joining our server every day. By the time of the publication, we had reached 10026 participants.

Our players come from three main sources: (1) **Minecraft fans reached through our outreach**. We performed an extensive media outreach and promotions, including TikTok, YouTube content creation, collaborating with famous Minecraft YouTubers, poster distributions and developed a [blog](#)

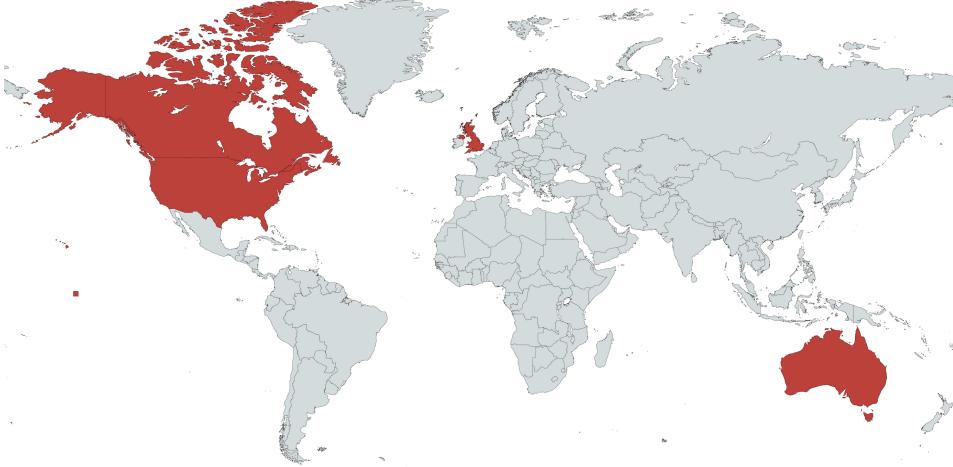


Figure 11: Our platform’s deployment regions. Participants are not restricted to these countries as countries/regions close to these areas can still get a low latency when playing on our server.

**website**. Early growth of this category was slow, but once our promotions reached some communities, such as a school or Discord group, sign-ups surged. These players usually log the most hours and contribute the most advanced and valuable data. (2) **Members of our development team**. A few of our teammates who love Minecraft have added several hundred hours of play. They often help organize groups and build shared bases to boost social interactions on the server. Their contribution rate is significantly smaller than the other two groups. (3) **Undergraduate students earning extra credit**. In one semester we partnered with one course that offered credit for playing a set number of hours. The participation rate was high at one point, though most of these students were Minecraft beginners and did not play for long. Since the program ended, the majority of this group has stopped contributing as well.

#### A.8 Content Moderation

To protect dataset integrity and community health, an active moderation team oversees the game server around the clock. Supported by automated plugins and help from veteran players, moderators can promptly detect, kick, and ban misbehaving users on sight. Because every gameplay action and chat message is recorded and logged, any incidents missed in real time can be traced and addressed retrospectively. These safeguards preserve a wholesome community atmosphere, sustain growth, and limit contamination of the collected data.

#### A.9 Data Pre-processing

We conducted a systematic and rigorous pre-processing pipeline on the raw data, transforming it into a structured and accessible format for public release.

##### A.9.1 Video

Video data were extracted from the original MKV format and converted into MP4 files with a standardized frame rate of 30 FPS. We trim the start of the video to align with the player’s first keyboard clicks, and we trim off the end to match when the player quits the game. This way, we make sure the content in the video is mostly related to the actual gameplay.

##### A.9.2 Audio

Audio data comprised two distinct tracks, each independently extracted and stored in WAV format. The first track, representing audio captured through player microphones (referred to as `audio_speak`), underwent a preprocessing step using the Silero Voice Activity Detector (VAD) to isolate valid speech segments from background noise. Furthermore, speaker diarization was implemented utilizing the

Pyannote model, assigning unique speaker labels to each speech segment. These speaker annotations were subsequently integrated with text transcripts generated using the Whisperx model. The second audio track (`audio_hear`), containing ambient audio, was processed similarly but without explicit noise filtering prior to transcription. Both audio datasets, along with their respective speaker labels, were systematically stored in a database inclusive of word-level timestamps in millisecond precision.

### A.9.3 Mouse & Keyboard

Mouse data preprocessing involved segregating mouse click events from mouse movement trajectories. Mouse movement data were aggregated into fixed temporal bins of 100 ms, each represented by arrays of dimensions (2, 10), corresponding to relative x and y movements across timestamps aligned to a maximum polling frequency of 100 Hz. Each data point was assigned to the nearest bin, and intervals lacking data were zero-padded. Mouse click events, analogous in structure to keyboard event data, were similarly binned into arrays of shape (79, 10), where 79 dimensions represent 76 distinct keyboard keys in addition to the left, right, and middle mouse buttons. This consistent binning procedure facilitated unified storage of both mouse movement and combined mouse-keyboard event data in an organized database, enabling flexible data querying and analysis. See Figure 12 for details.

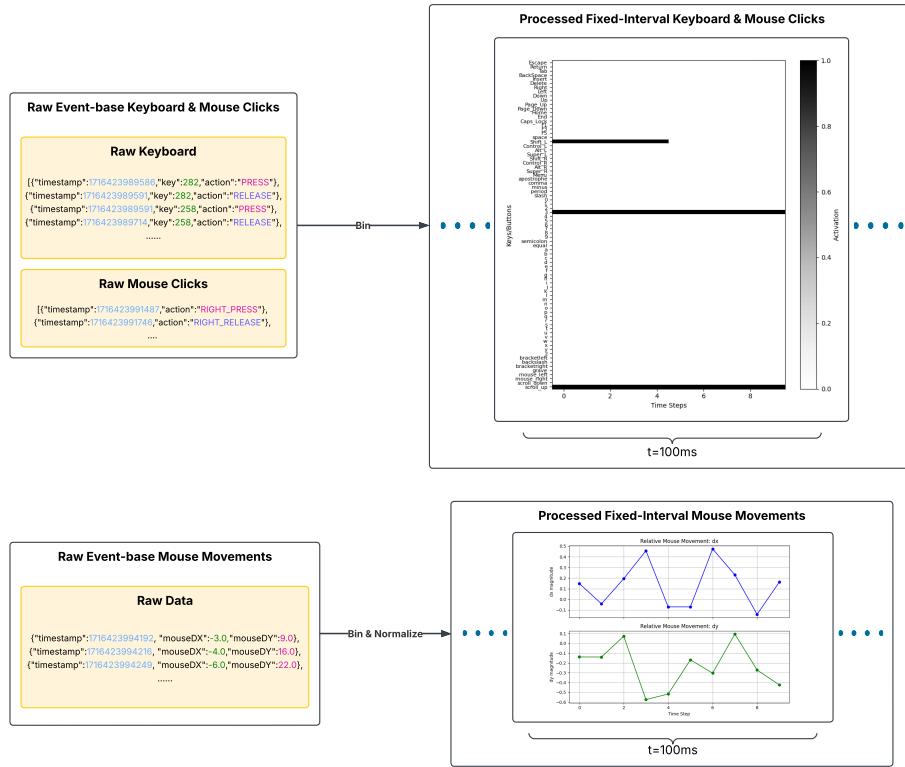


Figure 12: Illustration of how the mouse and keyboard data are pre-processed.

### A.9.4 Data Encoding

To reduce computational load on GPUs during future training, all data are pre-encoded into lower-dimensional neural representations. Specifically, video data are encoded using the SDXL-VAE-FP16-Fix model, an adapted variant of the SDXL Variational Autoencoder (VAE) optimized for 16-bit floating-point precision. This encoding generates latent representations sampled at 10 FPS with dimensions of ( $B, C = 4, W = 160, H = 96$ ).

Audio signals, encompassing both  $\text{audio}_{\text{speak}}$  and  $\text{audio}_{\text{hear}}$ , are encoded utilizing the EnCodec 24 kHz model. Following prior approaches [Shen et al., 2023, Evans et al., 2024], we employ continuous

latent vectors produced by this model, characterized by dimensions ( $B, D = 128$ ) with a frequency of 75Hz.

Additionally, we trained a temporal convolutional autoencoder with GRU layers to compress keyboard and mouse click event data, originally structured in arrays of shape (79, 10) per 100 ms interval (corresponding to one video latent frame). This autoencoder effectively compresses these inputs into a compact latent representation of (16, 5). Conversely, mouse movement data, already compactly represented with dimensions (2, 10) per 100 ms interval, is retained without additional encoding.

### A.10 Dataset Structure

The dataset is organized into multiple subfolders, each dedicated to data corresponding to an individual player. Within these player-specific directories, further subfolders contain both raw and encoded data from distinct gameplay sessions. Raw data includes video recordings in MP4 format captured at 30 FPS with a resolution of 1280x720, as well as two distinct WAV files with 48 kHz frequency representing  $audio_{speak}$  (player microphone recordings) and  $audio_{hear}$  (ambient recordings). An SQLite3 database contains detailed keyboard and mouse interaction data, optimized for efficient querying. The database also maintains comprehensive metadata for each session, including session identifiers, start timestamps, duration, and additional session-specific attributes.

Encoded data representations of video sessions are segmented into batches of 100 frames each and stored in PyTorch-compatible ‘.pt’ files. Encoded audio data for both audio tracks is stored in HDF5 files, facilitating efficient access patterns, as they permit partial data retrieval without necessitating the loading of entire datasets into memory.

Additionally, a global dataset containing metadata is provided, encapsulating comprehensive details of the dataset, including session identifiers, modality availability, UNIX timestamps corresponding to recording times, and associated player metadata. This structured metadata design supports efficient and precise queries by the data loader and related software components, thereby streamlining targeted data extraction from the extensive dataset.

### A.11 Data Versioning

There are 4 versions of the dataset, version 5 to version 8. You can find the version of each session in the info.txt under each session directory. Versions before 5 were used for testing purposes and thus not included in the dataset (the world is reset after the test). Versions 5-7 are versions of the data we recorded during the test launch period, which count towards less than 5% of the entire dataset. They have minor issues in their audio modalities and we do not recommend using them for audio training. These issues are fixed after version 8. Here is the detailed changelog of the versioning:

Version	Change History	Known Issues
5	Initial test launch with all modalities	Audio has unexpected artifacts; keys “enter”, “esc”, and “backspace” not recorded in chat window (fixed in released version)
6	Fixed the keylogger issue in version 5	Audio has unexpected artifacts; bad FPS occasionally
7	Added F1, F3, F5 keys; removed key binds for P, R, K (related to shaders config); added optimization mods that boost FPS	Audio has unexpected artifacts
8	Fixed the audio-artifacts issue	None

Table 5: Dataset version history

### A.12 DataLoader and Batch Sampler Design.

Our dataloader functions with a global metadata database, which contains game sessions metadata like who played the session, what modalities are available in these recordings, the start time of each gameplay session, etc. The database download is also available at <https://blog.plaicraft.ai/>

[2025/05/11/download-user-data/](#). The dataloader can then load the data from any specified sessions or players or from the entire database. As mentioned previously, due to the frequency discrepancy between video and audio, we treat 2 video latent frames, i.e. a 200ms window, as our minimal unit. Thus the dataloader will return a dict object of the following content:

```

Metadata: [B]
Video tensor shape: torch.Size([B, T, 2, 4, 96, 160])
audio_speak shape: torch.Size([B, T, 15, 128])
audio_hear shape: torch.Size([B, T, 15, 128])
Mouse movement shape: torch.Size([B, T, 2, 10, 2])
Key press shape: torch.Size([B, T, 2, 5, 16])
Transcript_in lengths: [B]
Transcript_out lengths: [B]
```

## B Evaluation Tasks

### B.1 Complete Evaluation Prompt Table

Test Type	Prompt	Response	Metric	CHC	Video	P <sub>start</sub> (ms)	P <sub>end</sub> (ms)	R <sub>start</sub> (ms)	R <sub>end</sub> (ms)
audio to action binding	Please press Space.	Space pressed.	D <sub>K</sub>	G <sub>a</sub>	1	65184	67045	67970	68120
audio to action binding	Please press Shift.	Shift pressed.	D <sub>K</sub>	G <sub>a</sub>	1	70366	71666	73120	73650
audio to action binding	Please press T.	T pressed.	D <sub>K</sub>	G <sub>a</sub>	1	86852	87812	89240	89410
spelling ability	Please press the letter that ends the word STOP, as in S-T-O... What letter was that?	P pressed.	D <sub>K</sub>	G <sub>rw</sub>	1	104752	113036	115950	116070
spelling ability	Please press the letter that ends big, as in B-I. What letter did you press? Did you press the letter G?	G pressed.	D <sub>K</sub>	G <sub>rw</sub>	1	117057	117697	119133	119373
audio to action binding	Please press the letter Q.	Q pressed.	D <sub>K</sub>	G <sub>a</sub>	1	121790	122531	123195	123515
action recognition	What did you just close?	Door.	Acc <sub>R</sub>	G <sub>v</sub>	1	145813	146594	151430	152860
internal state recognition	How much? What is that green? What is that green dot?	XP.	Acc <sub>R</sub>	G <sub>f</sub>	1	474250	475731	476517	476837
world model understanding	I'm going to shoot you with an arrow, okay? And I want you to tell me how many hearts of damage are done, okay?	Two.	Acc <sub>R</sub>	G <sub>f</sub>	1	508154	508954	510590	511010
object recognition	What do you have in your right hand?	Carrot.	Acc <sub>R</sub>	G <sub>v</sub>	1	501410	502551	548097	548557
object recognition	What do you have in your left hand?	Torch.	Acc <sub>R</sub>	G <sub>v</sub>	1	554727	557228	559065	559325
color understanding	What color sheep am I standing next to?	Black.	Acc <sub>R</sub>	G <sub>v</sub>	1	545784	547284	567047	567367
color understanding	What color sheep am I standing next to?	Gray.	Acc <sub>R</sub>	G <sub>v</sub>	1	545784	547284	548097	548557
color understanding	What color flower am I standing next to and looking at?	White.	Acc <sub>R</sub>	G <sub>v</sub>	1	554727	557228	559065	559325
color understanding	What color of flower am I looking at?	Red.	Acc <sub>R</sub>	G <sub>v</sub>	1	564107	565929	567047	567367
color understanding	What color of flower am I looking at?	Yellow.	Acc <sub>R</sub>	G <sub>v</sub>	1	564107	565929	572471	572892
type recognition	What is the stack of stuff that I'm standing next to made out of?	wool.	Acc <sub>R</sub>	G <sub>v</sub>	1	862534	863555	1111207	1111607

*Continued on next page*

Test Type	Prompt	Response	Metric	CHC	Video	P <sub>start</sub> (ms)	P <sub>end</sub> (ms)	R <sub>start</sub> (ms)	R <sub>end</sub> (ms)
arithmetic	If you add up the number of blocks in the wool pile and the number of blocks in the granite pile, how many blocks total are there?	Five.	Acc <sub>R</sub>	Gq	1	1142820	1144281	1146213	1146573
arithmetic	Now, if you add up the number of blocks in the granite pile and the wool pile, how many blocks are there total?	Four.	Acc <sub>R</sub>	Gq	1	1180221	1185165	1243900	1244180
arithmetic	I am going to place a block in the granite pile. How many blocks are in the granite pile now?	Four.	Acc <sub>R</sub>	Gq	1	1233210	1234591	1243900	1244180
object constancy	Now, jump up. Can you see me?	yes.	Acc <sub>R</sub>	Gv	2	602432	604614	605865	606185
object constancy	Jump up. Did I move?	no.	Acc <sub>R</sub>	Gv	2	613679	614439	615528	615788
audio to action binding	Press W.	W pressed.	D <sub>K</sub>	Ga	2	677774	678475	679650	679740
audio to action binding	Press S.	S pressed.	D <sub>K</sub>	Ga	2	682398	682998	684010	684390
audio to action binding	Press A.	A pressed.	D <sub>K</sub>	Ga	2	687001	687641	689700	689850
audio to action binding	Press D.	D pressed.	D <sub>K</sub>	Ga	2	692265	692625	692750	692940
audio to action binding	Press S.	S pressed.	D <sub>K</sub>	Ga	2	682398	682998	697070	697150
audio to action binding	Do it again. Press S.	S pressed.	D <sub>K</sub>	Ga	2	707405	708345	709500	709700
audio to action binding	Press W.	W pressed.	D <sub>K</sub>	Ga	2	677774	678475	713210	721000
working memory, object constancy	Have you been here before?	Yes.	Acc <sub>R</sub>	Gsm	2	1452930	1453691	1482922	1483343
working memory, object constancy	Will we be someplace we've been before or will we be someplace new? So say old or new.	old.	Acc <sub>R</sub>	Gsm	2	1786469	1792373	1797021	1797241
object recognition	Can you see a spider?	Yes.	Acc <sub>R</sub>	Gv	2	1850053	1850994	223010	223250
object recognition	Can you see any sheep?	Yes.	Acc <sub>R</sub>	Gv	2	1854695	1855656	223010	223250
type recognition	What kind of block am I standing on?	Sand.	Acc <sub>R</sub>	Gv	2	1923506	1924907	1926608	1926968
object recognition	What am I in right now?	Water.	Acc <sub>R</sub>	Gv	2	1938866	1939747	1940424	1940731
world model understanding	Is the sheep dead?	no.	Acc <sub>R</sub>	Gf	3	101422	102222	103176	103456
world model understanding	Is the sheep dead?	yes.	Acc <sub>R</sub>	Gf	3	101422	102222	110801	111221
symbol recognition, numerosity	how many pieces of wool are in the box?	Four.	Acc <sub>R</sub>	Gv	3	316272	318352	319976	320256
symbol recognition, numerosity	How many different kinds of things are in the box?	Two.	Acc <sub>R</sub>	Gv	3	322233	324093	324998	325158
symbol recognition, numerosity	How many pieces of wood are in the box?	six.	Acc <sub>R</sub>	Gv	3	291773	293074	369764	370124
type recognition	What is the first thing that I put in the box?	Wood.	Acc <sub>R</sub>	Gv	3	483807	485687	488025	488445
spatial understanding	Approximately how many blocks away is the box?	hundred.	Acc <sub>R</sub>	Gv	3	578622	580584	583714	584054
object recognition	Morgan, in front of me, what is that?	zombie.	Acc <sub>R</sub>	Gv	3	846473	848094	849320	849741
object recognition	Morgan, are those torches?	Yeah.	Acc <sub>R</sub>	Gv	3	891198	892679	895606	896186
symbol recognition, numerosity	How many pieces of wood?	Six.	Acc <sub>R</sub>	Gv	3	1668117	1669038	1670080	1670321
Reading	What is the name of the last person who left the game?	Brent.	Acc <sub>R</sub>	Grw	3	1902286	1904908	1908101	1908621
Reading	Okay, what is the name of the person who joined just after me?	Chloe.	Acc <sub>R</sub>	Grw	3	1936849	1946756	1948190	1948630
long term memory	What is the middle item in the chest?	Carrot.	Acc <sub>R</sub>	Glr	3	2171282	2172942	2174829	2175269

Table 6: Prompt-response evaluation pairs. The CHC column refers to the categories in Schneider and McGrew [2018]. All three videos are from Morgan, video 1 corresponds to session d868fc041c231673, video 2 corresponds to session b164ada80b3e43a0, video 3 corresponds to session d4e57df96d532e93. You can find these in the validation global metadata database.

## B.2 Evaluation Prompt Visual Examples

We pick a few prompts and show their corresponding frames for visualization in Table 7 and 8.

Frame Prompt	Prompt	Response
	How many sheep do you see?	Two.
	Can you see the sun?	Yes.
	Can you see the sun now?	No.
	Am I facing the direction of the box?	No.
	Am I facing the direction of the box?	Yes.

Table 7: Prompt visualisation set 1

Frame Prompt	Prompt	Response
	What animal do you see?	Sheep.
	What do you have in your right hand?	Carrot.
	What color sheep am I standing next to?	Gray.
	Jump one time. Can you see me?	No.
	Jump one time. Can you see me?	Yes.

Table 8: Prompt visualization set 2