CIALLO: Chrome-based Interactive Agent for LLM-powered Learning Online

Kehao Zheng[1 [0009−0009−1316−5502]] and Yang Shi[1 [0000−0001−6486−4340]]

Utah State University, Logan, UT 84322, USA
zkh200310@gmail.com, yang.shi@usu.edu

Abstract. Video-based learning plays a central role in contemporary education; however, learners frequently experience diminished social presence and friction in retrieving instructional information. Essential contextual details are often embedded within video frames or conveyed through instructors' verbal explanations, making them difficult to access using conventional text-based tools. To address these challenges, we present CIALLO, a Chrome extension that augments YouTube-based learning through a dual-LLM pedagogical agent framework. CIALLO integrates an animated, emotionally expressive pedagogical agent capable of dynamically adjusting its appearance and motion based on affective signals. To balance instructional rigor and socio-emotional interaction, the system employs two specialized LLM components: one responsible for generating context-grounded, academically oriented explanations, and another dedicated to producing affect-aware dialogue and behavioral modulation. Embedded directly within the Chrome sidebar, CIALLO provides context-aware assistance grounded in video transcripts. This design aims to reduce information retrieval friction while potentially enhancing perceived social presence and learner engagement.

Keywords: Animated Pedagogical Agents · Generative AI · Video-based Learning · Large Language Model.

1 Introduction and Background

Video-based learning has emerged as a prominent pedagogical approach in contemporary education [18, 24], with platforms such as YouTube hosting a vast ecosystem of instructional content [29]. However, online learners frequently encounter two persistent challenges that hinder effective learning. The first challenge is the lack of social presence during online learning: the absence of instructor cues and peer interaction fosters isolation and disengagement [1, 12], particularly in lengthy technical "mega-tutorials" [14].

The second challenge is information retrieval friction: essential context is often embedded in video frames or verbal instruction, making it inaccessible to conventional text-based retrieval tools [29, 30]. For instance, when learners encounter conceptual uncertainties, they cannot readily extract relevant context for further inquiry. As these difficulties accumulate, they can impose significant cognitive burden and interrupt the continuity of learning [6].

Pedagogical agents (PAs) have been proposed to enhance social presence and interactive support in digital learning environments, with prior research demonstrating their ability to improve self-efficacy, emotional support, and engagement through mechanisms such as the persona effect, social agency theory, and embodied cognition [27, 21, 5]. However, most existing PA frameworks rely on pre-scripted interactions, resulting in largely static behavior that lacks the flexibility to adapt to diverse learning situations and may discourage sustained interaction [5, 22, 4].

Recent advances in LLMs present new opportunities to overcome the limitations of static PA models, enabling more dynamic, context-aware, and affectively responsive interactions [10, 4]. Moreover, LLMs have demonstrated strong proficiency in reasoning over extended textual inputs [7, 3], making video transcripts a natural contextual source for LLM-based instructional assistance [19]. For instance, Ray et al. [23] leverage YouTube transcripts to generate detailed responses to student inquiries, demonstrating the potential of transcript-grounded LLMs to reduce information-seeking friction in educational video contexts.

Recent studies have begun exploring the integration of LLMs with animated PAs. For instance, Chen et al. [4] integrate LLM-powered dialogue generation with an animated 3D VTuber avatar system. Equipped with real-time motion capture and anime-style visual rendering, these agents evolve from static assistant interfaces into dynamic, emotionally engaging interactive partners. However, such systems remain largely unexplored in authentic learning environments. Moreover, creating high-quality VTuber avatars requires specialized technical expertise, posing significant personalization barriers [13]. This limitation in personalization is particularly significant given the established pedagogical value of avatar customization, as research in virtual learning environments indicates that allowing learners to tailor their agent's visual identity fosters stronger psychological identification with the companion [28, 26].

To address the two aforementioned challenges, i.e., the lack of social presence in self-directed video learning and the friction in retrieving instructional information, we introduce CIALLO, a Chrome extension that augments YouTube-based learning through an integrated pedagogical agent framework. Unlike text-only chatbots, CIALLO embeds an animated, multimodal PA directly within the browser sidebar to support in-situ interaction. Compared to recent VTuber-based systems such as Chen et al. [4], which rely on complex modeling that poses significant barriers to learner-driven personalization, CIALLO employs AI-generated avatars that can be readily customized by students to create their own agent representations. Additionally, we provide an intuitive configuration interface that allows learners to map specific visual appearances to distinct emotion tags, enabling fine-grained affective embodiment without technical overhead. Crucially, rather than operating in isolated demonstration contexts, CIALLO is embedded directly within authentic educational workflows, where learners can interact with the agent while engaging with actual instructional content.

2 Tool Descriptions

2.1 System Overview

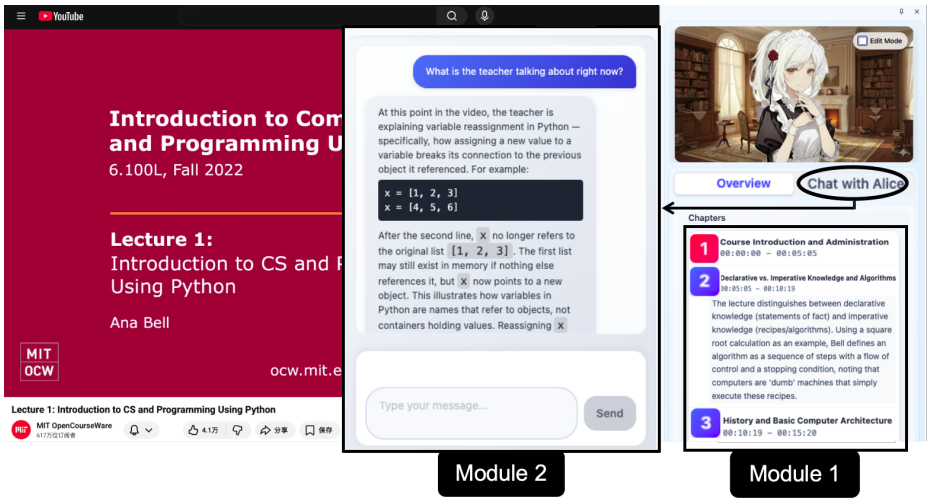

Fig. 1. Screenshot of CIALLO usage demo. CIALLO contains two modules: (1) the Overview module, which presents LLM-generated chapter segmentation and summaries synchronized with the video timeline, and (2) the Chat module, where learners interact with the pedagogical agent. The chat panel illustrates a context-grounded response generated based on the current playback position.

CIALLO is a Chrome extension designed to support learning on YouTube (see Fig. 1). The system detects the active video and retrieves its transcript using the YouTube-Transcript-API. The transcript is then processed by an LLM to generate chapter segmentation and summaries. These structured outputs are displayed in the Overview module (module 1 in Fig. 1) and synchronized with the video timeline. CIALLO continuously tracks the playback timestamp and highlights the corresponding chapter to support efficient navigation and contextual awareness.

The interface also includes a Chat module (module 2 in Fig. 1) that enables real-time interaction with the PA. The agent accesses both the transcript and playback position, allowing it to produce responses grounded in the instructional context currently being viewed, as illustrated in Fig. 1. This design aims to reduce information retrieval friction by enabling transcript-based assistance without leaving the video interface.

2.2 Multimodal Interaction

When a learner submits a query through the Chat module, the system invokes two LLM instances in parallel, each guided by distinct objectives. This dual-LLM

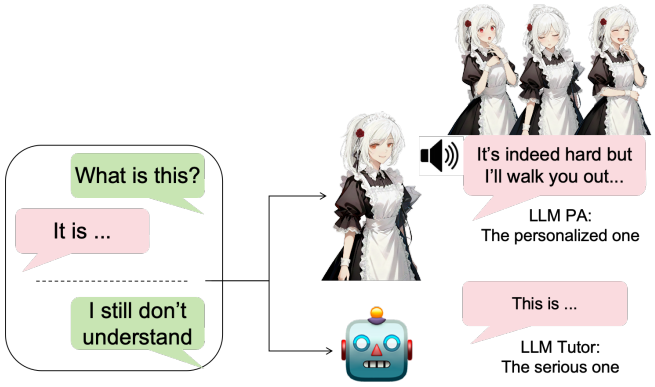

Fig. 2. Illustration of CIALLOs multimodal interaction pipeline. When a student submits a query, two LLM calls are invoked in parallel. $\text{LLM}_{\text{Tutor}}$ generates context-grounded instructional responses, while LLM_{PA} produces personalized, emotionally expressive dialogue. LLM_{PA} additionally outputs emotion tags that control the PAs appearance and animations, and can be synthesized into speech via a text-to-speech (TTS) module.

design is motivated by a practical limitation observed during system development: it is challenging for a single LLM to simultaneously maintain academically rigorous explanations while exhibiting rich socio-emotional expressiveness [17]. To balance this competing demand, we separate pedagogical reasoning from socio-emotional interaction.

$\text{LLM}_{\text{Tutor}}$ is responsible for generating academically rigorous, context-aware instructional responses grounded in the video transcript and conversation history. Functioning as a professional domain tutor, it prioritizes conceptual accuracy and clarity to support high-quality pedagogical guidance. LLM_{PA}, in contrast, is designed to enhance social presence and emotional engagement. Its design draws on prior research on PAs, social agency theory, and the persona effect, which collectively suggest that socially expressive agents can enhance learner motivation, perceived companionship, and engagement by fostering a sense of relational interaction [21, 5]. Rather than delivering technical explanations, it generates personalized, affect-aware dialogue that complements $\text{LLM}_{\text{Tutor}}$'s instructional response. By default, LLM_{PA} follows a predefined persona configuration; however, learners may customize its speaking style to better match their preferences. Beyond generating dialogue, LLM_{PA} produces structured emotion tags (e.g., happy, curious, surprised) that directly control the agents visual appearance and animations (e.g., subtle vertical motion or lateral shaking). Optionally, the generated dialogue is forwarded to a text-to-speech (TTS) module to synthesize spoken responses with natural, human-like prosody. In addition to a default voice configuration, learners may upload a preferred voice sample, enabling the TTS system to imitate its vocal characteristics during synthesis.

This personalized vocal embodiment is designed to further enhance the perceived relational connection between the learner and the PA.

2.3 Technical Implementation and Design Rationale

(1) LLM Integration and Prompt Engineering CIALLO employs a role-separated prompt engineering strategy to balance instructional rigor with socio-emotional engagement. A unified API interface supports flexible model provider selection, while distinct system prompts enforce strict behavioral boundaries and prevent role leakage.

LLM_{PA} is explicitly constrained to prioritize affective presence over technical explanation, yielding concise, emotion-tagged utterances that directly map to avatar animations. Example outputs include:

```
[curious] Hmm~ that seems kinda tough, doesn't it? {jump up and
down}
[smile] You're doing fine, don't rush yourself~ {sway left and
right}
```

In contrast, $\text{LLM}_{\text{Tutor}}$ is optimized for academic precision and transcript grounding, producing structured, timestamp-anchored responses:

```
"Regarding your question on dynamic memory allocation:
the segmentation fault occurs because the pointer references
deallocated memory. In this video segment (12:34), the
instructor demonstrates using valgrind to trace such
leaks by inspecting the call stack..."
```

When interfaced with commercial LLM APIs, streamed responses typically begin rendering within ∼1 second, supporting a fluid interactive experience.

(2) Text-to-Speech Module When enabled, CIALLO forwards LLM_{PA}'s generated dialogue to a TTS module for speech synthesis. We currently integrate CosyVoice v3 [8], an open-source model that supports easy voice cloning with minimal reference audio. This allows learners to upload a voice sample of their preferred character and synthesize emotionally expressive speech that matches the agent's affective state.

The TTS pipeline operates asynchronously: audio generation begins immediately after LLM_{PA} outputs text, while the UI displays the text response concurrently. Average synthesis latency is under 1.5 seconds when using commercial APIs. Users can also choose to disable TTS for text-only mode.

(3) Animated Character The animated character shown in Fig. 1 is generated via Gemini Nano Banana 2. Each message from LLM_{PA} includes structured emotion tags (e.g., *happy*, *curious*, *surprised*), and learners may upload corresponding avatar appearances for each tag. As illustrated in Fig. 2, a single

character can be rendered in multiple expressive variants. Nano Banana 2 facilitates this process through its strong character-consistency preservation: learners generate a base avatar once, then request posture or stylistic adjustments without re-specifying identity features. This workflow lowers the technical barrier for personalized avatar creation.

Critically, this customization capability is not merely aesthetic. Prior work in virtual learning suggests that when learners shape their agent's visual identity, they develop stronger psychological identification with the companion [28, 26]. Such identification has been shown to enhance the emotional relevance of instructional interactions and improve retention of agent-mediated content [20, 11, 26]. Moreover, customizable avatars support learners' self-efficacy, which is consistently linked to increased motivation and deeper cognitive engagement [26, 2]. Another key rationale for adopting a stylized animated avatar is to mitigate the ethical and cognitive risks associated with AI anthropomorphism. Highly photorealistic agents can inadvertently foster misplaced trust or unrealistic expectations about AI capabilities, which poses significant concerns in educational contexts [9, 25]. Empirical studies indicate that stylized, non-photorealistic characters effectively reduce these risks by maintaining a clear "artificial" identity and evoking a more approachable, game-like aesthetic rather than mimicking human realism [15]. Employing 2D animated characters could also avoid the uncanny valley [4] while preserving the socio-emotional cues necessary to foster *social presence* and leverage the *persona effect*. As established in pedagogical agent research, the mere presence of a responsive, expressive companion, even when transparently non-human, can significantly enhance learners' perceived engagement, reduce isolation, and sustain motivation [16, 27]. By decoupling visual realism from emotional expressiveness, CIALLO ensures that learners benefit from these well-documented psychological effects without conflating the agent with human-level intentionality or instructional authority.

3 Discussion and Conclusion

We presented CIALLO, a Chrome extension that augments video-based learning through a dual-LLM PA framework. Our design is intended to address two key challenges in video-based learning: information retrieval friction and the lack of social presence. One limitation of the current work is the absence of user studies. Future work will focus on conducting systematic user studies to evaluate learners perceptions and learning outcomes.

Our work positions LLMs as collaborative learning partners in video-based environments, emphasizing human-AI co-agency and collaborative intelligence. The ultimate goal is to explore how to strengthen social presence and emotional support while maintaining instructional rigor, improving engagement without disrupting learners focus or learning continuity.

References

1. Akcaoglu, M., Lee, E.: Increasing social presence in online learning through small group discussions. The international review of research in open and distributed learning **17**(3) (2016)
2. Ames, C.: Classrooms: Goals, structures, and student motivation. Journal of educational psychology **84**(3), 261 (1992)
3. Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., et al.: Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3639–3664 (2025)
4. Chen, E., Lin, C., Huang, Y.K., Tang, X., Xi, A., Lin, J., Koedinger, K.: Vtutor: An animated pedagogical agent sdk that provide real time multi-model feedback. In: International Conference on Artificial Intelligence in Education. pp. 152–159. Springer (2025)
5. Chen, E., Lin, C., Tang, X., Xi, A., Wang, C., Lin, J., Koedinger, K.R.: Vtutor: An open-source sdk for generative ai-powered animated pedagogical agents with multi-media output. arXiv preprint arXiv:2502.04103 (2025)
6. Costley, J., Fanguy, M., Lange, C., Baldwin, M.: The effects of video lecture viewing strategies on cognitive load. Journal of Computing in Higher Education **33**(1), 19–38 (2021)
7. Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., Wei, F.: Longnet: Scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:2307.02486 (2023)
8. Du, Z., Gao, C., Wang, Y., Yu, F., Zhao, T., Wang, H., Lv, X., Wang, H., Ni, C., Shi, X., et al.: Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. arXiv preprint arXiv:2505.17589 (2025)
9. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychological review **114**(4), 864 (2007)
10. Feng, S., Sun, G., Lubis, N., Wu, W., Zhang, C., Gasic, M.: Affect recognition in conversations using large language models. In: Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 259–273 (2024)
11. Ganesh, S., van Schie, H.T., de Lange, F.P., Thompson, E., Wigboldus, D.H.: How the human brain goes virtual: Distinct cortical regions of the person-processing network are involved in self-identification with virtual agents. Cerebral cortex **22**(7), 1577–1585 (2012)
12. Holly, M., Hildebrandt, J., Pirker, J.: A computer-supported collaborative learning environment for computer science education. In: International Conference on Immersive Learning. pp. 287–301. Springer (2024)
13. Kim, D., Lee, S., Jun, Y., Shin, Y., Lee, J.: Vtuber's atelier: The design space, challenges, and opportunities for vtubing. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–23 (2025)
14. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C.: Understanding in-video dropouts and interaction peaks inonline lecture videos. In: Proceedings of the first ACM conference on Learning@ scale conference. pp. 31–40 (2014)
15. Korre, D.: Comparing photorealistic and animated embodied conversational agents in serious games: An empirical study on user experience. In: International Conference on Human-Computer Interaction. pp. 317–335. Springer (2023)

16. Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., Bhogal, R.S.: The persona effect: affective impact of animated pedagogical agents. In: Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. pp. 359–366 (1997)
17. Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: Camel: Communicative agents for" mind" exploration of large language model society. Advances in neural information processing systems **36**, 51991–52008 (2023)
18. Li, W., Pea, R., Haber, N., Subramonyam, H.: Tutorly: Turning programming videos into apprenticeship learning environments with llms. arXiv preprint arXiv:2405.12946 (2024)
19. Liu, J., Wang, Y., Lyu, Y., Su, Y., Niu, S., Xu, X.O., Zhang, Y.: Harnessing llms for automated video content analysis: An exploratory workflow of short videos on depression. In: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing. pp. 190–196 (2024)
20. Mantovani, F., Castelnuovo, G., et al.: The sense of presence in virtual training: enhancing skills acquisition and transfer of knowledge through learning experience in virtual environments. In: Being there: Concepts, effects and measurement of user presence in synthetic environments, pp. 167–182. Ios Press (2003)
21. Martha, A.S.D., Santoso, H.B.: The design and impact of the pedagogical agent: A systematic literature review. Journal of educators Online **16**(1), n1 (2019)
22. Okado, Y., D. Nye, B., Aguirre, A., Swartout, W.: How can virtual agents scale up mentoring?: Insights from college students experiences using the careerfair. ai platform at an american hispanic-serving institution. International Journal of Artificial Intelligence in Education **35**(4), 2596–2615 (2025)
23. Ray, S., Sharma, S., Aditya, S., Goyal, P.: Eduvidqa: Generating and evaluating long-form answers to student questions based on lecture videos. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 34689–34715 (2025)
24. Sablić, M., Mirosavljević, A., Škugor, A.: Video-based learning (vbl)past, present and future: An overview of the research published from 2008 to 2019. Technology, Knowledge and Learning **26**(4), 1061–1077 (2021)
25. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. International journal of human-computer studies **146**, 102551 (2021)
26. Turkay, S., Kinzer, C.K.: The effects of avatar-based customization on player identification. In: Gamification: Concepts, methodologies, tools, and applications, pp. 247–272. IGI Global Scientific Publishing (2015)
27. Veletsianos, G., Russell, G.S.: Pedagogical agents. In: Handbook of research on educational communications and technology, pp. 759–769. Springer (2013)
28. Weinreich, P., Saunderson, W.: Analysing identity. Routledge (2005)
29. Yadid, S., Yahav, E.: Extracting code from programming tutorial videos. In: Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software. pp. 98–111 (2016)
30. Zhao, D., Xing, Z., Xia, X., Ye, D., Xu, X., Zhu, L.: Seehow: Workflow extraction from programming screencasts through action-aware video analytics. In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). pp. 1946–1957. IEEE (2023)