DialogLab: Authoring, Simulating, and Testing Dynamic Human-AI Group Conversations

Erzhen Hu[*]
University of Virginia
Charlottesville, VA, USA
eh2qs@virginia.edu

Yanhe Chen
Google XR Labs
Mountain View, CA, USA
yanhec@google.com

Mingyi Li
Northeastern University
Boston, MA, USA
li.mingyi2@northeastern.edu

Vrushank Phadnis
Google XR Labs
Mountain View, CA, USA
vrushank@google.com

Pingmei Xu
Google DeepMind
Mountain View, CA, USA
pingmeix@google.com

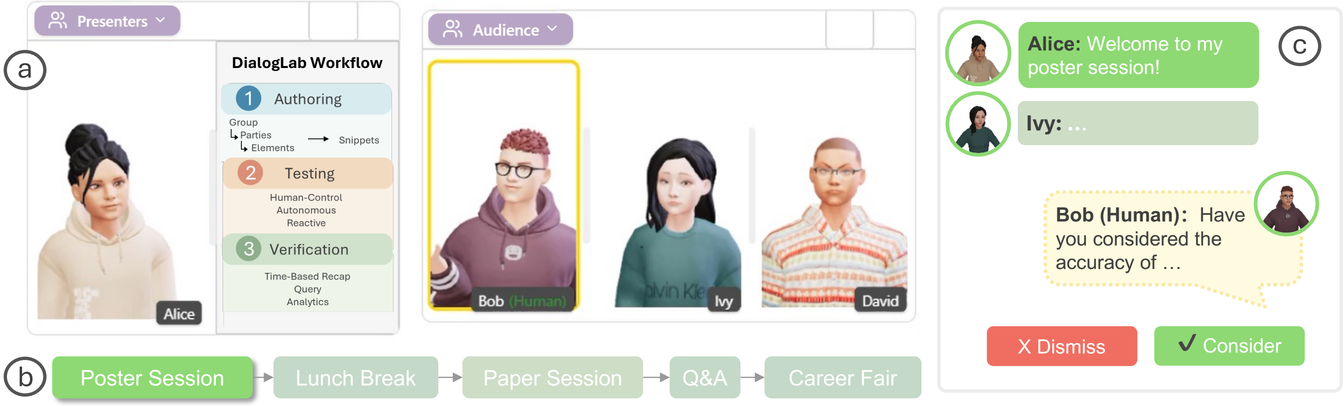
Xun Qian
Google XR Labs
Mountain View, CA, USA
xunqian@google.com

Alex Olwal
Google Research
Mountain View, CA, USA
olwal@acm.org

David Kim
Google XR Labs
Zurich, Switzerland
kidavid@google.com

Seongkook Heo
University of Virginia
Charlottesville, VA, USA
seongkook@virginia.edu

Ruofei Du[†]
Google XR Labs
San Francisco, CA, USA
me@duruofei.com



Alice: Welcome to my poster session!

Ivy: ...

Bob (Human): Have you considered the accuracy of ...

Figure 1: DialogLab supports the authoring, simulating, and testing of dynamic human-AI group conversations. (a) A designer sets up a human-AI group conversation scene where AI agents (Alice, Ivy, and David) follow a pre-scripted narrative. The scene involves a human participant (Bob) who can interject with unscripted questions during a poster session. (b) The interaction context is set to Poster Session, one of several event modes (e.g., Q&A, Career Fair) that modulate underlying conversation attributes such as turn-taking rules. (c) The *Preview Panel* visualizes a snippet of the ongoing conversation. While Alice's and Ivy's responses follow a scripted flow, Bob—marked as a human agent—asks an impromptu question that the conversation designer can review and either confirm or dismiss, or edit the message entirely. This showcases how DialogLab enables hybrid interactions that combine human input with pre-authored logic.

ABSTRACT

Designing compelling multi-party conversations involving both humans and AI agents presents significant challenges, particularly in balancing scripted structure with emergent, human-like interactions. We introduce DialogLab, a prototyping toolkit for authoring, simulating, and testing hybrid human-AI dialogues. DialogLab provides a unified interface to configure conversational scenes, define agent personas, manage group structures, specify turn-taking rules, and orchestrate transitions between scripted narratives and improvisation. Crucially, DialogLab allows designers to introduce controlled deviations from the script—through configurable agents

[*] This project was undertaken during the first author's internship at Google XR, with additional support from a Google PhD Fellowship awarded by Google Research.
[†] Corresponding author: Also contact: Erzhen Hu, eh2qs@virginia.edu

UIST '25, September 28-October 1, 2025, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09.
https://doi.org/10.1145/3746059.3747696

that emulate human unpredictability—to systematically probe how conversations adapt and recover. DialogLab facilitates rapid iteration and evaluation of complex, dynamic multi-party human-AI dialogues. An evaluation with both end users and domain experts demonstrates that DialogLab supports efficient iteration and structured verification, with applications in training, rehearsal, and research on social dynamics. Our findings show the value of integrating real-time, human-in-the-loop improvisation with structured scripting to support more realistic and adaptable multi-party conversation design.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

human-AI interaction, dialogues, multi-party conversation, real-time communication, conversation simulation, human-agent interaction, large language model

ACM Reference Format:
Erzhen Hu, Yanhe Chen, Mingyi Li, Vrushank Phadnis, Pingmei Xu, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2025. DialogLab: Authoring, Simulating, and Testing Dynamic Human-AI Group Conversations . In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3746059.3747696

1 INTRODUCTION

The simulation and creation of conversations with embodied agents have long been integral to virtual entertainment environments, especially for non-player characters (NPCs). Beyond gaming, these systems are increasingly deployed in diverse contexts such as education [60], healthcare [77], professional training [47], and retail [53]. Initially constrained by rule-based or scripted mechanisms and limitations in natural language understanding (NLU), the design of embodied agents have more recently leveraged Large Language Models (LLMs) to deliver more adaptive, context-aware dialogues.

Commercial tools like Character.AI[1] and Replika[2] illustrate this trend, enabling users to create custom AI personalities for one-on-one conversations. However, as human-AI interaction continues to evolve, user needs extend beyond single-agent dialogues. In many real-world scenarios, users participate in multi-party conversations, where multiple agents and humans interact under a blend of pre-scripted and emergent conditions. These settings introduce new complexities—*users must navigate shifting roles, manage turn-taking protocols, and negotiate control of the conversation as they move between reactive and proactive stances.* For instance, a human trainee may lead with structured questions in a medical simulation [75] but revert to a reactive role when questioned by an AI panelist. In multi-agent contexts, these transitions require sophisticated system support to maintain coherence across overlapping speaking turns, interruptions, and dynamic topic shifts.

Through a formative study, we sought to identify key challenges faced by conversation designers and developers: (1) Simulating how scripted agents balance rigidity and flexibility. (2) Testing interactions with human participants who deviate from expected behaviors. These challenges reveal two core design tensions: First, the **authoring and simulating** of multi-party hybrid human-AI dialogue systems remain complex, often demanding significant programming or prompt engineering [103], hindering rapid iteration by designers. Even with scripted guidance, balancing narrative structure reflecting diverse group dynamics with unpredictable human input requires extensive testing. Second, designers must navigate a trade-off between ***structured scripts***, which provide consistency and safety, and ***improvised human-led dialogue***, which fosters realism, user agency, and spontaneity [65, 104]. Current approaches often represent two extremes: fully scripted interactions, which ensure control but lack adaptability and spontaneity [55, 108]; and fully generative and automated systems, which allow open-ended dialogue but are difficult to steer toward specific goals [59, 106]. While structure is critical in domains like healthcare or education, rigid scripting can make users feel passive and limit adaptability. Generative systems, by contrast, allow open-ended interaction but lack control for task-oriented or training scenarios. These challenges become even more pronounced in *multi-party* and group settings due to the increased complexity of managing roles, turn-taking, and recovery from interruptions [38, 39, 93].

To address these challenges, we present DialogLab, a unified authoring tool for creating multi-party conversations with hybrid scripted/LLM-driven embodied agents. Our system enables designers to: (1) configure basic persona, scene, and conversation setups that help define diverse group dynamics (Fig. 1a-b); (2) embed LLM-driven agents during prototyping to simulate human unpredictability within pre-defined narratives (Fig. 1c); (3) incorporates features designed to facilitate verification and reflection of these complex group and conversational dynamics.

To evaluate DialogLab, we conducted a user study with five regular users and nine domain experts who used DialogLab to author and test multi-party conversation scenarios. Participants appreciated the diverse conversation flexibilities and agency during the authoring, and reported clear verification of group behaviors, and strong support for real-time improvisation and testing. These findings demonstrate DialogLab 's ability to support structured authoring, dynamic simulation, and reflective verification of complex hybrid human-AI conversations.

In summary, our contributions are:

- A flexible **framework** and **authoring paradigm** for *multi-party human-AI conversations*, enabling configuration of group dynamics, social protocols, and hybrid scripted/improvised interactions.
- **DialogLab**, an open-source system[3] implementing this framework, supporting the workflow of authoring, simulating, and interactively testing multi-party dialogues.
- **A human evaluation** comparing human agents with reactive agents *in group settings* and **a user study** of DialogLab with regular users and domain experts, demonstrating its effectiveness in building and simulating human-AI group conversations.

[1] character.AI: https://character.ai/
[2] Replika: https://replika.com/
[3] DialogLab: https://github.com/ecruhue/DialogLab

2 RELATED WORK

Our work builds upon a rich body of research in generative conversation and agents, and authoring tools.

2.1 Authoring Tools of Hybrid Human-AI and Human-Human Conversations

2.1.1 Scenario-Based and Open-Ended Conversational Systems. Prior work has explored scenario-based conversational agents, where dialogues follow predefined scripts or task flows and authors define possible user choices with pre-defined responses [78]. Schank and Abelson's Script Theory [83] laid the foundation by showing how people rely on cognitive scripts for routine interactions. For embodied agents, such scripts structure user responses and ensure systematic coverage of content [6]. Narrative-driven systems similarly guide users along story arcs, especially in education and training contexts [4, 26, 92], while Bogost's procedural rhetoric [7] emphasizes how rules shape user beliefs through structured interaction. In contrast, human-driven conversational agents emphasize spontaneity and user agency, drawing on sociological and HCI theories. Turn-taking theory [80] and Situated Action Theory [91] highlight how conversations unfold dynamically based on context, not pre-planned structure. These ideas inform flexible dialogue systems, such as chatbots and virtual assistants [101], that adapt to unpredictable user input. Frameworks like Activity Theory [20] and Common Ground Theory [13] further emphasize shared understanding and contextual adaptation [6]. Bridging these extremes, mixed-initiative systems combine scripted guidance with flexible user input. Horvitz's model of mixed-initiative interaction [37] supports seamless shifts in control between human and machine, allowing systems to guide or respond as appropriate—offering a foundation for hybrid dialogue design.

DialogLab provides a visual authoring paradigm using snippets to offer fine-grained control over the blending of scripted structure and human-driven improvisation specifically within complex multiparty conversational flows.

2.1.2 AI-Based Avatar Authoring Tools. AI-driven avatars are increasingly used in education, training, and collaborative settings to simulate human interaction [22]. However, many existing systems depend on predefined responses and require technical expertise for customization [76]. To improve accessibility, recent authoring tools allow non-programmers to configure avatar behaviors, language, and roles [96]. While some efforts have aimed to support flexible training goal modification, challenges remain—such as the complexity of dialog management interfaces reported in virtual agent tools (*e.g.*, [75]). With LLM integration, tools like GPTAvatar [76] and Fink's authoring system [21, 23] allow users to create realistic, voice-enabled avatars through text or graphical interfaces. Others, such as MAGI [107] and Convai [66], focus on embodiment and one-on-one interactions in educational or game environments. However, these systems do not support adaptive multi-agent dialogue or group conversation dynamics beyond dyadics.

In contrast, DialogLab is designed for multi-party conversation authoring and management. Rather than focusing on individual avatars, it enables configurable turn-taking behaviors, and real-time communication simulation across multiple participants. By supporting structured and emergent group dialogues, DialogLab offers greater flexibility than traditional avatar-based tools.

2.1.3 Authoring Tools for Real-Time Human-Human Communication Systems. Group dynamics are naturally evolved and formed in real-time human-human conversations, where multiple participants contribute to the conversation together synchronously. Tools such as Microsoft Teams (together mode), and gather.town that support real-time communication thus have allowed users to customize their own shared spaces, thereby redefining the boundaries of conversation. Some tools [27, 39] enabled live malleable mirrors of users' presence during meetings to help conversational dynamics and improve turn-taking. Additionally, other tools [40, 74] allowed for meeting background customization with text-to-image and image-to-image models to help construct personalized shared space.

For DialogLab, the focus is not how the spatial factors may effect or adapt to different the conversation dynamics via enabling user agency [27] or AI-mediated approaches [74], but how diverse group dynamics can be more seamlessly constructed, and designed to enable flexible prototyping of hybrid human-AI group conversations.

DialogLab shifts focus to human-AI communication—enabling users to define and explore different group dynamics in mixed human-AI settings with less focus on the environment setup. Furthermore, agency also plays a key role in the authoring process itself [33, 34], especially when LLMs take part in the conversation design. Given the open-ended nature of LLMs, co-creation emerges between creators and AI-generated contributions. The inherent unpredictability of LLMs [10, 52] introduces serendipity and improvisation, which can enrich the authoring process by inviting new, unexpected directions. WhatELSE [61], WhatIF [65], and Orchid [104] enabled visualization for authoring emergent narratives, and provide pivots in the outline space with LLMs, but focuses on the control of text-based story and narrative generation and storytelling, rather than dialogues between characters. DialogLab supports this creative interplay by offering authors tools to balance scripted structure with AI-driven spontaneity—enabling iterative refinement and real-time testing of complex, multi-party conversations.

2.2 Evolution of Conversational AI

2.2.1 Early Dialogue Systems. Early research in conversational AI, as surveyed by Gao et al. [24], established fundamental categorizations of dialogue systems into *question-answering agents*, *task-oriented dialogue systems*, and *social chatbots*. These early systems primarily focused on information retrieval, task completion, and open-domain engagement through single-agent architectures. This traditional approach laid the groundwork for more sophisticated multi-party interaction systems, though it was limited by its focus on dyadic conversations between a single AI agent and user.

2.2.2 LLM-based Role-Playing Conversational Agents. Recent advances have enabled the development of sophisticated role-playing conversational agents that can emulate specific characters or personas within predefined scenarios [87, 88, 97]. Systems such as Character.AI and OpenAI's ChatGPT demonstrate how large language models can maintain certain personalities, respond in character, and adapt dynamically to user inputs [11, 12, 32, 89]. However, these systems typically focus on **open-ended, freeform role-play** rather than **structured, goal-driven conversations**. While they excel at maintaining character consistency, they lack the conversation scaffolding and behavioral constraints necessary for simulating professional interactions. Character.AI offers "Room" options where multiple bots can be invited into a shared space. However, the system lacks structured control over conversational dynamics and requires the user to manually select which character should respond, or rely on a randomized selection (*e.g.*, a dice roll). Role-playing AI has been explored in medical training, leadership coaching, and customer service [44, 47, 53, 55, 60, 62, 77], but existing systems are predominantly single-agent and focus on dyadic interactions. Unlike these approaches, DialogLab explicitly models multi-party conversational dynamics, incorporating turn-taking, backchanneling, and interruptions to mirror real-world communication.

2.2.3 Multi-Agent Systems and Social Behaviors. The evolution of conversational AI has increasingly focused on multi-agent architectures and social dynamics. Frameworks such as AutoGen [102] and Camel [58] demonstrate how multiple AI agents can collaborate autonomously on complex tasks through defined functional roles (e.g., Planner, Critic, Executor). Similarly, systems like ChatCollab [49] enable peer-to-peer human-AI collaboration on specific work products like software teams. These systems have revealed emergent properties in AI-to-AI coordination, though they primarily focus on task completion rather than complex group dynamics.

Another line of research explores the *sociality* of agents in group settings. For instance, Chen et al. [11] introduced a benchmark to assess agent social behaviors at individual and group levels, highlighting their role in group dynamics. At a broader scale, work on *agent societies*—such as Generative Agents [69] and Xi et al. [105]—has examined how autonomous agents interact in open-ended environments, exhibiting *emergent behaviors* like cooperation and norm formation. This focus on emergent social dynamics is also central to frameworks like AgentGroupChat [28], which uses debate scenarios to elicit and analyze collective behaviors shaped by language.

In contrast, DialogLab centers on *micro-level* multi-party human-AI conversations, with an emphasis on structured turn-taking, role-based interaction, and adaptive dynamics. Different from enabling autonomous social emergence, DialogLab supports controlled, user-configurable conversation creations involving both humans and AI agents.

3 FORMATIVE STUDY

To understand the challenges and design considerations in creating human-AI conversations, we conducted semi-structured interviews with domain experts.

3.1 Participants and Procedure

We recruited 7 participants (4 female, 3 male; ages 27–40, $\bar{x} = 30.6$) through targeted outreach. Participants included software/UX engineers, designers, scientists, and PhD students with experience in human-agent conversations across contexts such as sales, classrooms, meetings, and healthcare. Their demographics are attached in Appendix (Table 1).

Each remote interview lasted 30–45 minutes, followed by a brief survey collecting demographic information and summaries of participants' project goals, workflows, and challenges. Interviews began with an overview of participants' projects involving human-agent conversations, including their motivations, application contexts, and whether conversations were scripted or adaptive. Participants described a variety of scenarios, including AI tutors and healthcare companions. We then focused on their authoring process, exploring how they approached system building, conversation flow design, and refinement. Finally, we discussed challenges they faced during iteration and evaluation, and their experiences with existing authoring tools and workflows.

3.2 Findings

Two researchers employed affinity diagrams to analyze and categorize participants' responses. This analysis revealed three fundamental characteristics of human-agent conversations. Additionally, we identified three key challenges in authoring conversations **(F1-F4)**.

F1: Fragmented and Tedious Multi-Agent Authoring Workflows. Participants described the process of configuring conversations with multiple AI agents as time-consuming, manual, and technically fragmented. Many relied on patching together different platforms for scripting, voice synthesis, avatar animation, and scene composition. For example, P1 explained her workflow where she generated scripts using LLMs, converted them to speech via tools like Speechify, manually assigned distinct voices to avoid duplication, and synchronized these with avatar gestures and movements, - "*recording the entire thing is just like really tiresome... there is no system that supports this. So you have to do everything from scratch...* " In classroom simulations (P2), getting student agents to speak only in response to silences required meticulous control over timing and turn-taking: "*The most important part...is that the agent only engages when the silence happens.*"

This emphasizes the need for offering an integrated authoring environment combining conversation structure, behavior rules, and role-specific templates to support rapid iteration and design reuse.

F2: Rigidity vs. Flexibility in Scripted vs. Adaptive Agents. Participants often chose between two extremes: fully scripted systems that ensured consistency but limited engagement, and adaptive LLM-based systems that allowed for richer interaction but lacked reliability. Scenario-based designs were common in contexts like healthcare or education, where agents had to follow strict guidelines or learning objectives. As P1 noted, "*Fixed scripts provide consistent control variables for smaller samples.*" Yet over-scripting risked alienating users: "*I was worried that if agents are just talking to each other...people will disengage.*" Adaptive systems, used by P2 and P5, allowed for more naturalistic interactions by responding dynamically to user input, but could derail the conversation, especially in high-stakes scenarios. As P7 noted, "*A small kernel of uncertainty...could cause serious consequences.*"

Participants developed hybrid designs, such as P2's classroom prototype (scripted teacher, adaptive students) or P5's training system (keyword triggers, LLM follow-ups). However, they reported lacking tools to orchestrate transitions between scripted and adaptive phases, or to designate specific agents (e.g., authority figures) as fixed versus others (e.g., learners) as adaptive.

F3: Testing Interactions With Human Participants Who Deviate From Expected Behaviors. Participants found it difficult to test how their conversation systems would perform when interacting with real users, especially when those users behaved in unexpected or off-script ways. Many relied on text-based prototyping to save time, but this failed to surface critical interaction problems: "*We just use text input, which is quicker, but it doesn't really catch what happens when people talk out of order or go off topic.*" (P2), while others scripted targeted scenarios ("*engineers came up with examples...for the kinds of demos we run*", P4). However, these methods failed to capture the full spectrum of human improvisation, leading to critical gaps in testing coverage.

This indicates a need to prototype how systems recover from dynamic user behavior, such as someone changing the topic, expressing frustration, or misinterpreting an instruction.

F4: Subjective and Object Metrics For Verifying Systems with LLMs. Most participants (4/7) described ad-hoc refinement processes, such as "*does it work? Okay...now do the user study*" (P1), which made identifying issues in complex workflows difficult. As P3 noted, "*How do you know anything is broken?*" The lack of systematic testing frameworks of human-AI conversations was compounded by the subjectivity of conversational success metrics (*e.g.*, tone, coherence), which resisted automation and required labor-intensive human review. "*Automating metrics is not possible...some are too subjective*" (P5).

3.3 Design Goals

We highlight the following three design goals informed by our findings and related work.

DG1: Support Structured yet Flexible Authoring of Multi-Party Conversations. Our formative study revealed that authoring multi-party, human-AI conversations is a time-consuming and technically fragmented process **(F1)**. Therefore, a primary goal is to unify this fragmented workflow, providing designers with a cohesive environment to seamlessly manage agent personas, conversational structures, and interaction rules without requiring extensive technical overhead.

DG2: Bridge the Gap Between Scripted Control and Emergent Spontaneity. Participants described a core tension between the need for for scripted control to ensure consistency and the desire for unscripted spontaneity to foster realism and user agency (**F2**). Furthermore, they lacked effective methods for testing how their systems would handle unpredictable human behaviors like interruptions or digressions (*e.g.*, "targeted demos," **F3**). These gaps highlight the need for fluid transitions between structure and improvisation, and enable designers to configure spontaneous, emergent human behavior during both authoring and testing.

DG3: Enable Insightful, Multi-Run Verification for Designers and End-Users of the Conversation. Participants faced tension between unreliable automation (*e.g.*, LLM variance, role alignment) and labor-intensive human review **(F4)**. Users need better ways to diagnose issues related to turn-taking, sentiment, and coherence across multiple interaction runs. The goal is to move beyond ad-hoc evaluation, by providing tools for systematic and interpretable verification that make complex multi-party dynamics accessible and actionable for users."

4 DIALOGLAB FRAMEWORK

We propose DialogLab, a framework for authoring and testing hybrid human-AI multi-party conversations. DialogLab supports the configuration of dynamic social structures, turn-taking patterns, and content-sharing elements—enabling the rapid creation and testing of realistic, multi-party conversational scenes. It introduces a layered conceptual model that separates social dynamics from conversational flow, and integrates configurable agents that support design-time testing and runtime improvisation. Following the design goals we have discussed, we introduce a series of concepts that we adopt in DialogLab's design. Then, we briefly illustrate the novel workflow we propose to support conversation creators to create high-quality human-AI conversations.

4.1 Conceptual Dimensions of Hybrid Human-AI Multi-Party Conversations

The conceptual model is organized along two interrelated dimensions: **group dynamics**, governing social structure and participant roles, and **conversation flow dynamics**, describing how dialogue unfolds via turn-taking, topic transitions, and interaction styles. This layered model enables users to construct scenes that reflecting both social context and dynamic interaction **(DG1)**. To illustrate these concepts, we use a running example: designing a hybrid conversation simulating four individuals interacting during a social event at a research conference.

> *Alice, Bob, and Charlie gather together at a social event to attend David's poster presentation. Later, all four grab lunch together with casual chats about their research interests and ongoing projects. After that, they will attend a talk presented by Alice, and others will be the audience, with a moderated Q&A session.*

4.1.1 Group Dynamics. ***Group dynamics*** include ***groups***, ***parties***, and ***elements***. First, a ***group*** represents the highest-level unit of a human-AI multi-party conversation, typically organized around a high-level topic such as a social event in a research conference. *Groups* establish the overall conversational context and are informed by social-psychological theories of group dynamics [57].

Within each *group*, we define ***parties*** as sub-groups that reflect distinct roles, functions, or perspectives. This hierarchical structuring allows for modeling interactional asymmetries and social roles, which are central to understanding real-world conversations [30, 94]. In the above example, during David's poster presentation, one *party* includes David who presents the poster, while another *party* (Alice, Bob and Charlie) acts as the audience. *Parties* are not always required in less structured conversations. Specifically, during lunch, the four individuals chat freely without forming any *parties* [68, 99].

Lastly, ***elements*** refer to the specific participants and shared content involved. ***Participants*** are the individual actors within each *party*, designated as either ***AI-driven agents*** or ***real humans***; for instance, Alice, Bob, and Charlie could be AI agents while David is a real human. ***Content*** includes shared materials, like the poster David presents to structure or stimulate interaction [14, 51, 63].

4.1.2 Conversation Flow Dynamics. While *group dynamics* establish the social and structural context, ***conversation flow dynamics*** capture how interaction unfolds over time [80, 84]. In DialogLab, we define that a multi-party conversation composes multiple ***snippets***, where each *snippet* serves as a placeholder representing an interaction phase with (1) a subset of *participants* involved, (2) a sequence of turns addressed by the *participants*, and (3) associated interaction styles. For example, David's poster session might begin with a structured introductory *snippet*, followed by a Q&A *snippet* between the audience *party* (Alice, Bob, Charlie) and David. Subsequent lunch *snippets* would exhibit different flow characteristics—more interruptions and backchannels—while a moderated *snippet* during Alice's talk Q&A would enforce hand-raising for structured turn-taking.

By adopting the *group dynamics* and *conversation flow dynamics*, DialogLab enables nuanced control over both social structure and interaction flow—supporting the design of rich, realistic multi-party conversations. In the next section, we will introduce a unified workflow supporting creators to define and test such human-AI multi-party conversations on-the-fly.

4.2 Author-Test-Verify Workflow

We propose a three-stage workflow (Fig. 2) to address the design considerations of designing multi-party conversations: ***Authoring***, **Testing**, and **Verification**.

During the ***authoring stage***, a user defines the *groups*, *parties*, *participants*, and *contents* that comprise a multi-party conversation. The user assigns roles to each *participant*, distinguishing between *real humans* and *AI agents*. Meanwhile, the user defines *snippets* that consist of multiple turn-taking chats among the *participants*.

DialogLab's **testing stage** addresses **(DG2)**, recognizing that static authoring alone is insufficient for robust conversations. In practice, creators often rely on ad-hoc or random inputs to test their conversations, which fails to capture the complexity and variability of real-world interactions **(F3)**. To address this, we support runtime simulation that replays all authored *snippets* in sequence, allowing conversation creators to experience the entire conversation as it would unfold. Conversations will involve both *AI agents* and *real humans*. Handling the *real human* role presents challenges (**DG2**), so we offer two modes: (1) ***Real-human mode***, where the conversation creator may manually play the *real human* during testing. (2) Alternatively, ***simulated-human mode***, where creators can use a configurable AI **Human Agent** to represent the human for efficient iterative testing. These *human agents* are not limited to reenacting scripts. Instead, the creator can configure emergent and unscripted behavior [91], such as interruptions, emotional reactions, or conversational digressions to reflect the unpredictable real-human behaviors **(DG2)**. This also allows creators to stress-test the resilience of their designs without needing to recruit human participants at every iteration **(F3)**.

Last but not least, the ***verification stage*** facilitates structured analysis and reflection. It also provide an integrated environment for creators to interpret complex interaction data and understand multi-party dynamics **(DG3)**.

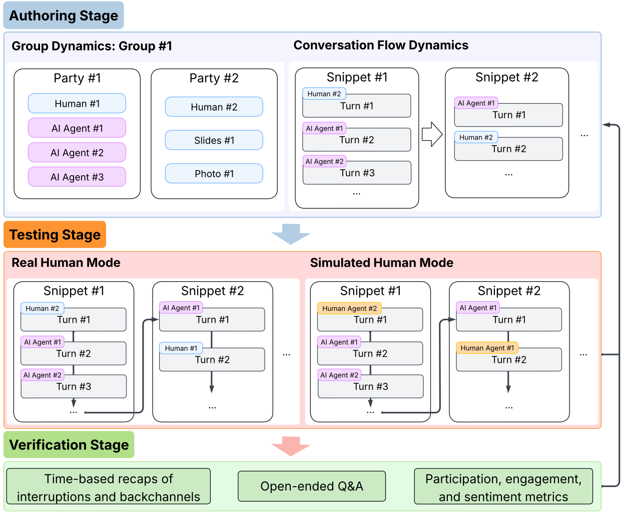


Figure 2: DialogLab workflow: (1) A conversation creator begins with the *authoring stage*, defining the *group dynamics* and *conversation flow dynamics*. (2) The *testing stage* allows testing the conversation either by involving a human user in the conversation (*real human mode*) or simulating a human user using an agent (*simulated human mode*). (3) In the *verification stage*, DialogLab generates structured analysis through recap, queries, and analytics. The creator may iterate on the authoring process based on insights from both testing and verification stages until satisfied.

Beyond a single-run inspection in the *testing stage*, the *verification stage* supports systematic, multi-run analysis with comparative insights across multiple conversation runs, visualizing patterns in turn-taking, topic coherence, and sentiments [17, 81]. This enables both iterative refinement of agent behaviors and snippet logic, and reflective learning about group dynamics, communication styles, and individual participation.

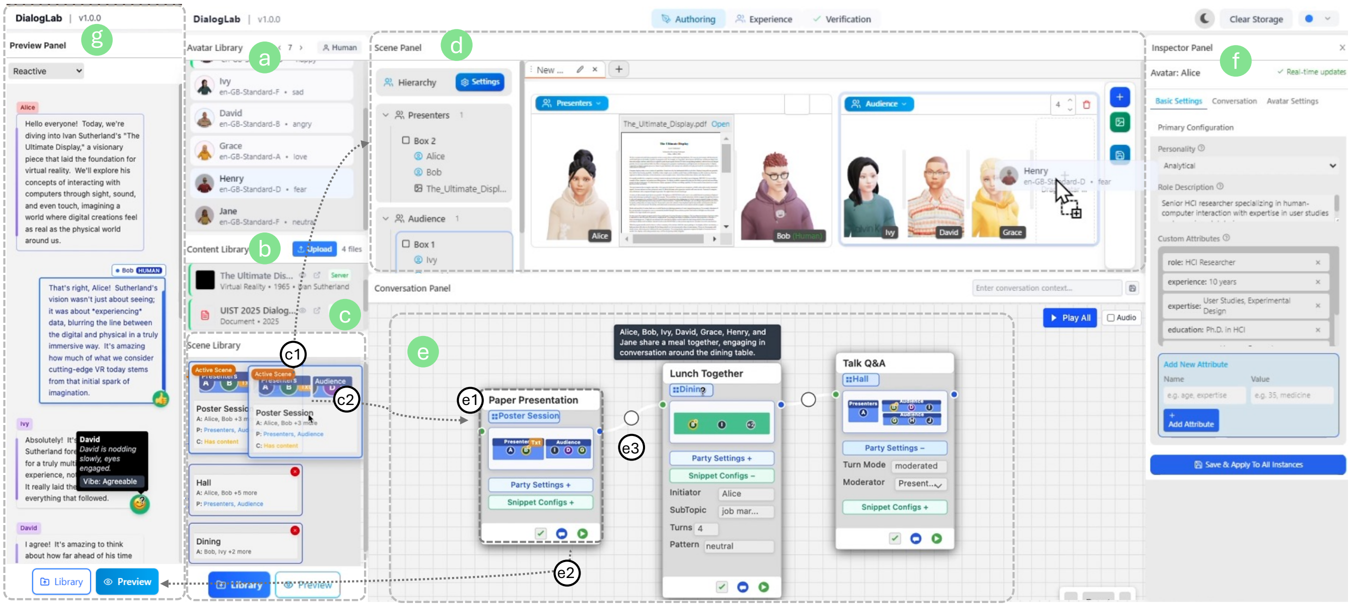
By referring to experience during the *testing stage* and the outputs from the *verification stage*, the conversation creator may go back to the *authoring stage* to modify the features and entities. After multiple iterations, the creator is satisfied with the human-AI multi-party conversations generated by DialogLab.

5 DIALOGLAB SYSTEM DESIGN

This section details the DialogLab framework's operation and design for creating and testing human-AI multi-party conversations.

5.1 Authoring Stage

The authoring stage consists of the core creative process of designing multi-party conversations. Users define participants, establish their relationships and spatial organization, and create the flow of dialogue through an intuitive, visual interface **(DG1)**.



Alice: Hello everyone! Today, we're diving into Ivan Sutherland's "The Ultimate Display," a visionary piece that laid the foundation for virtual reality. We'll explore his concepts of interacting with computers through sight, sound, and even touch, imagining a world where digital creations feel as real as the physical world around us.

Bob HUMAN: That's right, Alice! Sutherland's vision wasn't just about seeing; it was about *experiencing* data, blurring the line between the digital and physical in a truly immersive way. It's amazing how much of what we consider cutting-edge VR today stems from that initial spark of imagination.

Ivy: Absolutely! It's Sutherland fore for a truly mult experience, no It really laid the everything that followed.

David: *David is nodding slowly, eyes engaged.* Vibe: Agreeable

David: I agree! It's amazing to think about how far ahead of his time

Alice, Bob, Ivy, David, Grace, Henry, and Jane share a meal together, engaging in conversation around the dining table.

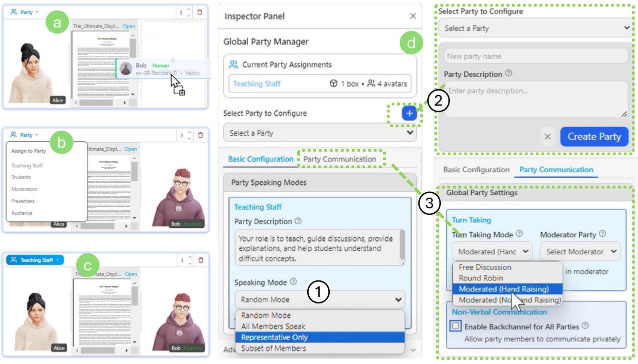
Role Description: Senior HCI researcher specializing in human-computer interaction with expertise in user studies

role: HCI Researcher
experience: 10 years
expertise: User Studies, Experimental Design
education: Ph.D. in HCI

Figure 3: DialogLab allows users to design, configure, and preview multi-party conversations. First, users author the *group dynamics* by creating a new *scene* in the (d) *Scene Panel* with *participants* selected from the (a) *Avatar Library* and *contents* from the (b) *Content Library*. Users can also (c1) drag an existing *scene* from the (c) *Scene Library* to the *Scene Panel*. Then, users author the *conversation flow dynamics* of each *scene* by (c2) dragging it to the (e) *Conversation Panel* as (e1) a *snippet*. Users can configure the sequential flows between any two *scenes* by (e3) connecting the *snippets*. The (f) *Inspector Panel* supports detailed configurations of the *group dynamics* and *conversation flow dynamics* when users select corresponding assets in the user interface. After authoring, users (e2) click to open the (g) *Preview Panel*, which simulates and tests the authored conversations in real-time.

5.1.1 Configuring Group Dynamics. *Scenes* serve as the key interface elements that define the *group dynamics* (§4.1.1). Users first author *scenes* that contain the key components of *group dynamics* via the *Scene Panel* (Fig. 3d). *Participants* are visually represented as avatars, which can be dragged to the center *Scene Panel* from the *Avatar Library* (Fig. 3a). Users can configure each avatar's name, voice, emotional tone, and backstory using the *Inspector Panel* (Fig. 3f) when selecting the avatar. While avatar configuration is not the system's core focus, DialogLab adopts conventions from recent LLM-driven frameworks [25, 70, 73] to support expressive agent behaviors for live, multi-party interactions. *Contents*, such as shared slides or documents, are pre-loaded in the *Content Library* (Fig. 3b), and can be added into *scenes* as shareable media. Users can further configure the *groups* and *parties* using the Scene Hierarchy (left). The toolbar on the right offers operational controls such as 'add' and 'save'.

Configuring Parties. As addressed in §4.1.1, *parties* serve as a key component of the flow in a group conversation. As users drag *participants* and *contents* into the *Scene Panel* (Fig. 4a), we allow users to further configure the formation of the *parties* via grouping box interaction. Box associations also indicate ownership and sharing of *contents*. When *contents* appears inside a box with one or more *participants* (*e.g.*, Fig. 4c), those *participants* are presenting it. *Contents* in standalone boxes without avatars is publicly visible but not attributed to any individual. DialogLab allows users to further author the role of each *party* with custom inputs or preset templates (Fig. 4d-2). Speaking Mode and Turn-Taking Mode control member participation during active turns. The *Speaking Mode* (Fig. 4d-1)for each *party* determines how members contribute when it's their party's turn: (1) *All* contribute collaboratively; (2) a *Representative* speaks for the party; (3) a defined *Subset* contributes; or (4) *Random*



Party Description: Your role is to teach, guide discussions, provide explanations, and help students understand difficult concepts.

Figure 4: *Group dynamics* authoring. In the *Scene Panel*, users can (a) drag *participants* and/or *contents*; (b-c) configure the formation of *parties* via the interactive boxes; (d) tune the speech and turn-taking behaviors of the *parties* via the *Inspector Panel*.

selection among these options adds variability. The *Turn-Taking Mode* (Fig. 4d-3) governs how members take their turns [100]: (1) *Free:* Members speak immediately. (2) *Moderated - Hand-Raising:* Members signal readiness; a moderator assigns turns.

After these authoring steps, users successfully configure the *group dynamics* of this conversation.

5.1.2 Configuring Conversation Flow Dynamics. After establishing the *group dynamics*, users configure the *conversation flow dynamics* using the *Conversation Panel* (Fig. 3e). As described in §4.1.2, conversations are divided into *snippets*, each representing a distinct sub-phase of dialogue (*e.g.*, transitioning from a talk to a Q&A exchange, or from a debate to a decision-making moment).

As shown in Fig. 3 e1, as users drag a *scene* into the *Conversation Panel*, it becomes a *snippet* represented as a node. Each *snippet* node includes (1) *Scene Configs* authored in the previous step; (2) *Snippet Configs*, where users define how the interaction unfolds; (3) *Transition Logic* (Fig. 3 e3), allowing *snippets* to be chained to simulate real-time conversational flows.

Configuring Snippets. To address the complexity of real-world group interactions, we distill key attributes of multi-party conversations from prior work. Users can first configure the *Snippet Details* from the *Inspector Panel* (Fig. 5c).

- **Snippet Initiation and Initiator** that define who initiates the *snippet* and how it begins—mimicking leadership behaviors and entry points observed in natural dialogue [84].
- **Turn Numbers** that specify the number of back-and-forth exchanges to control the length and pacing of the snippet [14].
- **Interaction Patterns** that control the overarching tone and effective posture of the *snippet* (*e.g.*, neutral, agreement, disagreement). These patterns support a range of conversational dynamics and interpersonal relationships, reflecting context-dependent scenarios [29, 31, 90]. For example, a debate scenario might favor disagreement patterns, while a decision-making meeting could emphasize consensus-building.

Further, to simulate the fluidity and unpredictability of human conversation, DialogLab also supports two additional attributes for *Advanced Settings* (Fig. 5e).

- **Interruptions** that define spontaneous overlaps and their frequency, initiators, emotional tone, and targets [15, 56, 86].
- **Backchannels** configure listener responses (*e.g.*, *"okay"*, *"hmm"*) and nonverbal cues (*e.g.*, nodding) to indicate understanding or engagement. These behaviors are known to shape interpretation and flow in group conversations [3, 15, 19, 45, 50, 98].

After the configuration, DialogLab updates the *scene* description (Fig. 5b) as well as the conversation prompt (Fig. 5d). Users are free to edit or regenerate the prompts to fine-tune the desired tone or emphasis for the later testing stage.

5.2 Testing Stage

After completing the authoring, users can test their conversation designs through real-time simulation. This phase is crucial for validating designs before deployment, ensuring they work not only in ideal conditions but also when faced with unexpected inputs, or behaviors (*e.g.*, interruptions).

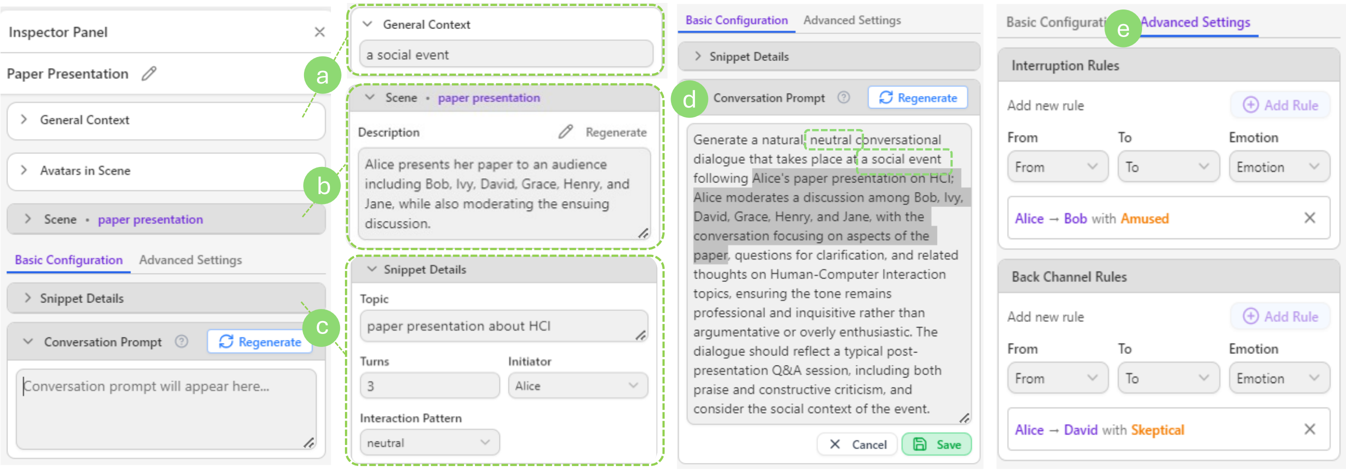
The *Preview Panel* (Fig. 3g) enables fast, lightweight interactive testing. It displays real-time interactions, including: (1) AI (Fig. 6a-1) and human (Fig. 6a-2) *speaking turns*; (2) *backchannel cues* shown as emojis (Fig. 6b-4), and (3) *system messages* (Fig. 6a-3) indicating dynamics like hand-raising status (counts, approvals) if applicable. The testing interactions introduced below rely on the *Preview Panel*.

5.2.1 Testing with the Real-Human Mode. One straightforward testing approach for users is to role-play the human *participant* as configured in the *group dynamics*. Using the *Preview Panel*, users can input reactions when the human turns approach. Yet, DialogLab provides a more powerful testing approach to address the needs identified in **DG2**.

5.2.2 Testing with the Simulated-Human Mode. As mentioned in §4.2, to support user participation during testing and simulate spontaneous human contributions, DialogLab introduces the simulated-human testing mode, which introduces an AI-driven *Human Agent* (Fig. 7) that emulate the real human behaviors authored in the conversation. These agents serve as both testing tools and design collaborators, helping authors stress-test dialogue flow by injecting unexpected or contextually rich contributions **(DG2)**.

We support three testing modes controlling the *Human Agent* behavior: (1) *Reactive Mode* (Fig. 7e): The *Human Agent* only responds when directly prompted, mimicking conventional turn-based AI behavior. (2) *Human Control Mode* (Fig. 7a–d): The user actively monitors the conversation and can selectively introduce agent contributions from a system-curated list. (3) *Autonomous Mode*: The *Human Agent* proactively participates using pre-defined rules (random or fixed), triggering spontaneous actions such as topic shifts or emotional responses without user input. Specifically, for the *Human Control Mode*, suggestions are previewed through an *Audit Panel* (Fig. 7b), where users can: (1) preview message candidates; (2) adjust the intent/mode (*e.g.*, shift topic, offer emotional response) via a dropdown menu (Fig. 7b–c); (3) edit and send the messages manually; (4) dismiss suggestions entirely. To guide this mode, we incorporate cognitively grounded behaviors that reflect common dynamics in human dialogue [43, 79]: (1) *Drift:* Introduces informal topic shifts; (2) *Extend:* elaborates to build shared understanding; (3) *Question:* challenges, clarifies, or redirects conversation focus [5]; (4) *Emotional:* Adds affecting reactions (*e.g.*, surprise, frustration) [48]. User can select one of these four dynamics and regenerate the conversation (Fig. 7b-c), or edit the message. Then, once users click "Consider" on a message (Fig. 7c), the system inserts it into the conversation, initiating an impromptu phase (Fig. 7d) where the *Human Agent* temporarily steers the conversation for multiple turns.

Fig. 8 illustrates a comparison between a *Reactive Mode Human Agent* (left) and a *Human Control Mode Human Agent* (right), with all other configurations held unchanged (topic: DialogLab Design Review). We simulated five distinct conversation scenarios—ranging from social events with strangers, DialogLab Paper presentations with Q&A, and stakeholder meetings, to a historical debate between Confucius and Socrates, and an interactive game dialogue. For each scenario, we generated five iterations per mode (human agent vs. reactive agent), resulting in 50 total conversations (5 scenarios × 5 iterations × 2 modes), which enables systematic comparison of how different agent behaviors influence conversational dynamics



Inspector Panel ×

Paper Presentation

› General Context

› Avatars in Scene

› Scene • paper presentation

Basic Configuration Advanced Settings

› Snippet Details

⌄ Conversation Prompt Regenerate

Conversation prompt will appear here...

⌄ General Context

a social event

⌄ Scene • paper presentation

Description Regenerate

Alice presents her paper to an audience including Bob, Ivy, David, Grace, Henry, and Jane, while also moderating the ensuing discussion.

⌄ Snippet Details

Topic: paper presentation about HCI

Turns: 3

Initiator: Alice

Interaction Pattern: neutral

Basic Configuration Advanced Settings

› Snippet Details

Conversation Prompt Regenerate

Generate a natural, neutral conversational dialogue that takes place at a social event following Alice's paper presentation on HCI: Alice moderates a discussion among Bob, Ivy, David, Grace, Henry, and Jane, with the conversation focusing on aspects of the paper, questions for clarification, and related thoughts on Human-Computer Interaction topics, ensuring the tone remains professional and inquisitive rather than argumentative or overly enthusiastic. The dialogue should reflect a typical post-presentation Q&A session, including both praise and constructive criticism, and consider the social context of the event.

× Cancel Save

Basic Configuration Advanced Settings

Interruption Rules

Add new rule ⊕ Add Rule

From To Emotion

Alice → Bob with Amused ×

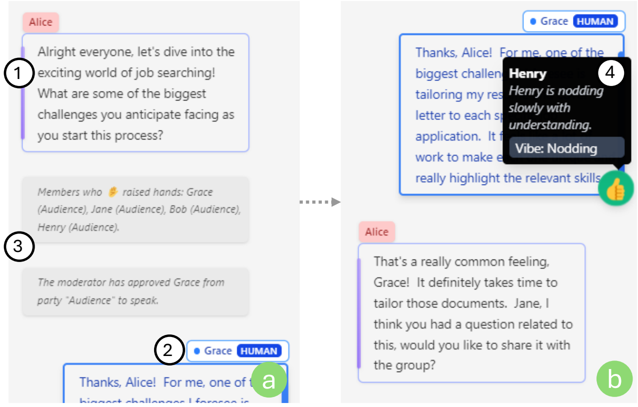
Back Channel Rules

Add new rule ⊕ Add Rule

From To Emotion

Alice → David with Skeptical ×

Figure 5: The configuration of the *snippets* in the *Inspector Panel*. It includes: (a) a general context field that defines the situational backdrop (*e.g.*, "a social event"); (b) the automatically updated *scene* description; (c) snippet details, including topic, number of turns, initiator, and interaction pattern; (d) an editable conversation prompt, auto-generated by combining the general context, scene description, and snippet details; (e) Advanced Settings, where users can configure interruption and backchannel rules between characters, along with associated emotional tones.



Alice: Alright everyone, let's dive into the exciting world of job searching! What are some of the biggest challenges you anticipate facing as you start this process?

Members who ✋ raised hands: Grace (Audience), Jane (Audience), Bob (Audience), Henry (Audience).

The moderator has approved Grace from party "Audience" to speak.

● Grace HUMAN: Thanks, Alice! For me, one of the biggest challen… tailoring my res… letter to each sp… application. It f… work to make e… really highlight the relevant skills

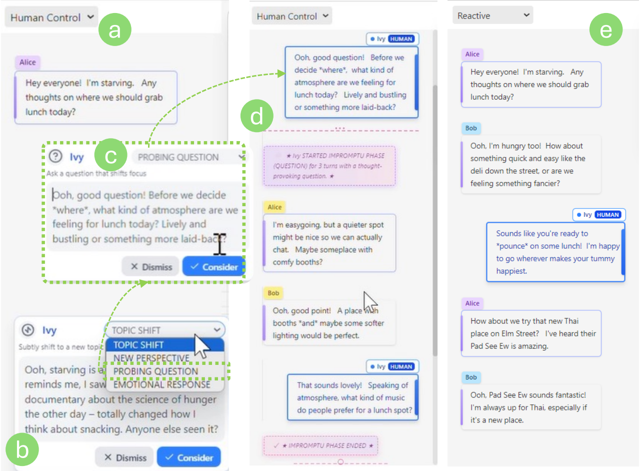
Henry: *Henry is nodding slowly with understanding.* Vibe: Nodding 👍

Alice: That's a really common feeling, Grace! It definitely takes time to tailor those documents. Jane, I think you had a question related to this, would you like to share it with the group?

Figure 6: Preview Panel. The Preview Panel visualizes different message types in real-time conversations: (1–2) Speaking turns showing contributions from participants; (3) System messages indicating moderated group dynamics for hand raising (4) Backchannel messages visualized as emojis attached to messages, providing non-verbal feedback.

across diverse contexts. The analysis is included in the Appendix Appendix B.

In summary, these simulated-human testing modes enable authors to simulate richer, more authentic dialogue by experimenting with how spontaneous turns affect flow, tone, and group dynamics. Now, we go to the final stage for users to further verify the testing outcome as a whole to gain more insights to improve the group conversation design from a higher-level perspective.



Human Control

Alice: Hey everyone! I'm starving. Any thoughts on where we should grab lunch today?

Ivy — PROBING QUESTION — Ask a question that shifts focus

Ooh, good question! Before we decide *where*, what kind of atmosphere are we feeling for lunch today? Lively and bustling or something more laid-back?

× Dismiss ✓ Consider

Ivy — TOPIC SHIFT — Subtly shift to a new topic

TOPIC SHIFT / NEW PERSPECTIVE / PROBING QUESTION / EMOTIONAL RESPONSE

Ooh, starving is a… reminds me, I saw… documentary about the science of hunger the other day – totally changed how I think about snacking. Anyone else seen it?

× Dismiss ✓ Consider

Human Control

● Ivy HUMAN: Ooh, good question! Before we decide *where*, what kind of atmosphere are we feeling for lunch today? Lively and bustling or something more laid-back?

★ *Ivy STARTED IMPROMPTU PHASE (QUESTION) for 3 turns with a thought-provoking question.* ★

Alice: I'm easygoing, but a quieter spot might be nice so we can actually chat. Maybe someplace with comfy booths?

Bob: Ooh, good point! A place with booths *and* maybe some softer lighting would be perfect.

● Ivy HUMAN: That sounds lovely! Speaking of atmosphere, what kind of music do people prefer for a lunch spot?

✓ ★ *IMPROMPTU PHASE ENDED* ★

Reactive

Alice: Hey everyone! I'm starving. Any thoughts on where we should grab lunch today?

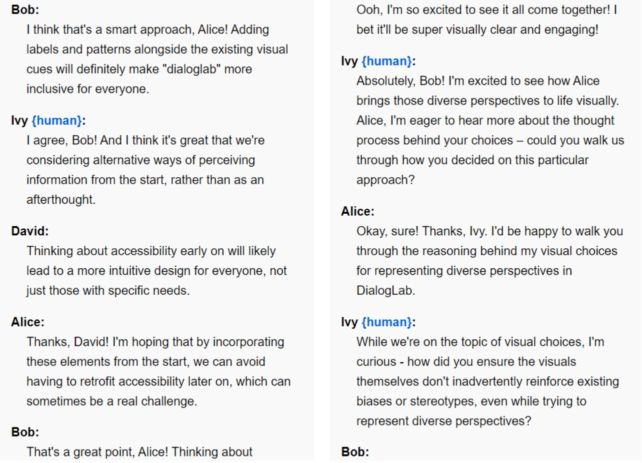
Bob: Ooh, I'm hungry too! How about something quick and easy like the deli down the street, or are we feeling something fancier?

● Ivy HUMAN: Sounds like you're ready to *pounce* on some lunch! I'm happy to go wherever makes your tummy happiest.

Alice: How about we try that new Thai place on Elm Street? I've heard their Pad See Ew is amazing.

Bob: Ooh, Pad See Ew sounds fantastic! I'm always up for Thai, especially if it's a new place.

Figure 7: Two modes for the *Human Agent* behaviors in simulated-human testing. (a) In *Human Control Mode*, users can configure the behavior of the Human Agent. In specific, users can review, modify, or regenerate the turn in the (b) *Audit Panel* by selecting different intent types such as topic shifts or emotional responses. (c) Upon selecting "Consider," the chosen message is inserted into the conversation. (d) This triggers an impromptu phase, where the Human Agent engages in a multi-turn contribution. On the other hand, in the (e) *Reactive Mode*, the Human Agent only responds when directly addressed by other non-human characters, simulating typical AI turn-taking behavior.



Bob:
I think that's a smart approach, Alice! Adding labels and patterns alongside the existing visual cues will definitely make "dialoglab" more inclusive for everyone.

Ivy {human}:
I agree, Bob! And I think it's great that we're considering alternative ways of perceiving information from the start, rather than as an afterthought.

David:
Thinking about accessibility early on will likely lead to a more intuitive design for everyone, not just those with specific needs.

Alice:
Thanks, David! I'm hoping that by incorporating these elements from the start, we can avoid having to retrofit accessibility later on, which can sometimes be a real challenge.

Bob:
That's a great point, Alice! Thinking about

Ooh, I'm so excited to see it all come together! I bet it'll be super visually clear and engaging!

Ivy {human}:
Absolutely, Bob! I'm excited to see how Alice brings those diverse perspectives to life visually. Alice, I'm eager to hear more about the thought process behind your choices – could you walk us through how you decided on this particular approach?

Alice:
Okay, sure! Thanks, Ivy. I'd be happy to walk you through the reasoning behind my visual choices for representing diverse perspectives in DialogLab.

Ivy {human}:
While we're on the topic of visual choices, I'm curious - how did you ensure the visuals themselves don't inadvertently reinforce existing biases or stereotypes, even while trying to represent diverse perspectives?

Bob:

Figure 8: Difference Between Human Agent and Reactive Agent

5.3 Verification Stage

The final stage of the workflow introduces the interactive **Verification Dashboard** (Fig. 9) for analyzing conversation outputs.

The interface is composed of four core panels: The *Conversation Selector* (Fig. 9a) allows users to browse and load multiple conversation runs for comparative analysis. The *Interactive Timeline* (Fig. 9b) visualizes the flow of the conversation, marking speaker turns and highlighting key events such as *Interruptions*, *Questions*, *Impromptu* moments, and *System* messages—helping users locate significant interaction points. The *Synced Transcript Panel* (Fig. 9d) displays the full dialogue with color-coded speaker turns and bidirectional linking with the timeline for seamless navigation. The *Analytics Panel* (Fig. 9c) provides visual summaries of computed metrics: a *pie chart* for *Speaking Time Distribution*, *bar charts* for *Engagement Levels*, a *line chart* for *Cumulative Speaking Time*, and *progress bars* for *Participation Balance*, *Topic Coherence*, and *Sentiment*. A *radar chart* supports normalized *Participant Comparison* across multiple interaction dimensions, inspired by prior work (*e.g.*, [81]) Finally, the *Q&A Interface* (Fig. 9e) allows users to pose natural language questions about the conversation and receive AI-generated responses that summarize patterns and insights, enhancing interpretability for both experts and non-experts. Importantly, these metrics extend beyond a single conversation instance and can be aggregated across multiple iterations, enabling broader insight during the verification process **(DG3)**.

5.4 Implementation

Deployed as a web application, the system's frontend communicates with a Node.js/Express server. The server handles requests, manages the multi-agent interaction logic (§A.2), and interfaces with external APIs (Gemini-Flash for text generation; Google TTS/ElevenLabs for speech synthesis). To support embodied interaction, generated speech audio is processed server-side to extract visemes for lip synchronization (output as JSON). This audio and lip-sync data is then served to the client. This enables realistic mouth movements during avatar-driven conversation playback.

The multi-agent mechanism is implemented as a rule-based system composed of three core modules: agent management, interaction handling, and conversation generation. These modules work in coordination to produce dynamic, context-aware dialogue among multiple agents, which is detailed in §A.2.

On the frontend, three.js [9] and React Three Fiber [16] render the 3D avatars. The system utilizes 10 pre-configured Ready Player Me[4] avatars (5 male, 5 female) with various animations, enabling realistic playback synchronized with the generated audio and viseme data.

6 APPLICATIONS

Fig. 10 shows different types of potential applications of DialogLab.

Conversation Practice, Skill Development, and Immersive Learning. DialogLab supports structured conversation practice for diverse user groups. For educational practitioners, it enables the design of learning experiences (Fig. 10b) where users build communication skills by interacting with AI agents simulating diverse personalities, historical events, or spiritual dialogues. Simultaneously, regular users—without technical backgrounds—can configure agents to rehearse specific real-world scenarios, such as meetings, interviews, and presentations (*e.g.*, Fig. 10a). While educators emphasize general skill-building, individual users focus on targeted preparation. In both cases, DialogLab offers a safe environment for authoring, practicing, and receiving feedback, tailored to different levels of specificity and conversational goals. For instance, users can simulate social parties (Fig. 10c) or conference social events to refine communication skills of meeting with strangers, or create historical scenes to explore, such as an immersive debate between Socrates and Confucius (Fig. 10d).

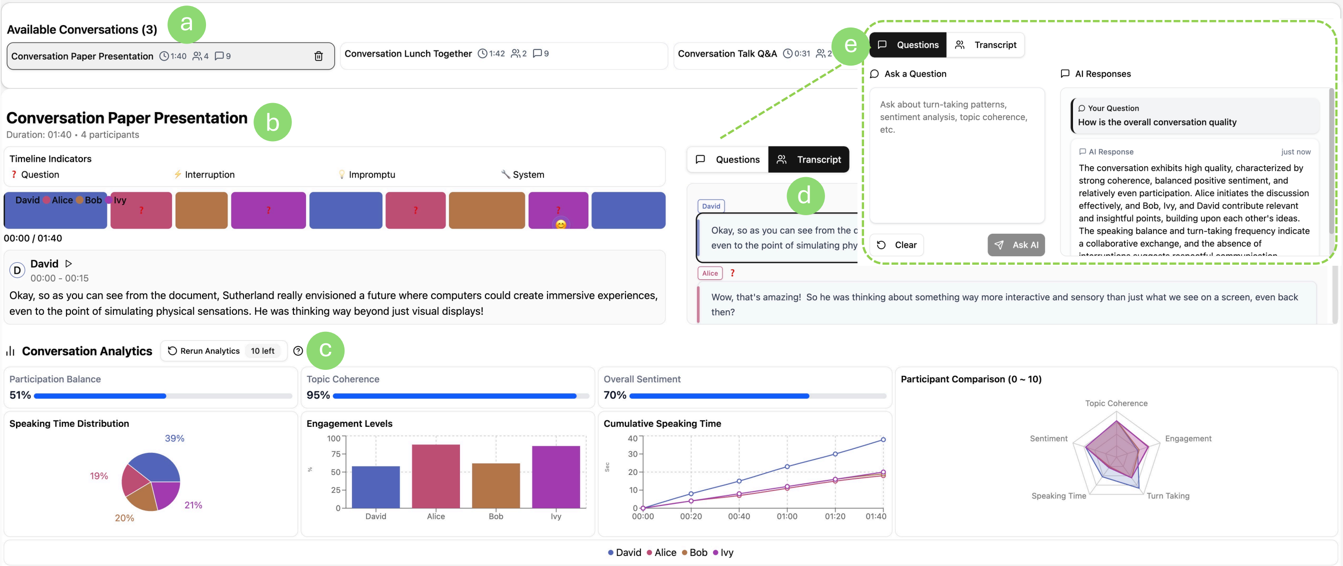
Inspiration of Character and Game Design. Writers, directors, and game designers/developers can use DialogLab to explore character dynamics and dialogue flow [73, 95] (Fig. 10e). It allows simulation of natural, multi-party interactions between AI-driven personas to refine tone, test reactions, and explore conflict. Screenwriters can analyze ensemble scenes for authenticity, while game developers can prototype branching NPC dialogue. Persona uploads and behavior editing further support creative iteration and storytelling.

Research and Testing of Group Dynamics. DialogLab also serves as a platform for social science research, allowing researchers to design and test conversations with agents exhibiting specific behaviors. This supports controlled studies of group dynamics—such as dominance, role hierarchy, or conflict—and helps explore how patterns shift with different compositions. Researchers can simulate experimental conditions, validate hypotheses before involving human participants, and gather detailed interaction analytics by engaging human participants immersed in the simulation.

7 EVALUATION

To evaluate the usability and understand how different user groups would utilize DialogLab for creating dynamic multi-party conversations, we conducted an ethics-approved user study with participants representing each of our target application areas (§5.4).

[4] Ready Player Me: https://readyplayer.me



Okay, so as you can see from the document, Sutherland really envisioned a future where computers could create immersive experiences, even to the point of simulating physical sensations. He was thinking way beyond just visual displays!

Wow, that's amazing! So he was thinking about something way more interactive and sensory than just what we see on a screen, even back then?

Ask about turn-taking patterns, sentiment analysis, topic coherence, etc.

How is the overall conversation quality

The conversation exhibits high quality, characterized by strong coherence, balanced positive sentiment, and relatively even participation. Alice initiates the discussion effectively, and Bob, Ivy, and David contribute relevant and insightful points, building upon each other's ideas. The speaking balance and turn-taking frequency indicate a collaborative exchange, and the absence of

Figure 9: The Verification Dashboard supports structured review and reflection through four integrated panels: (a) conversation selector, (b) interactive timeline of speaker turns and key events, (c) analytics panel with visual metrics, (d) synced transcript viewer, and (e) natural language Q&A interface for querying conversation dynamics.

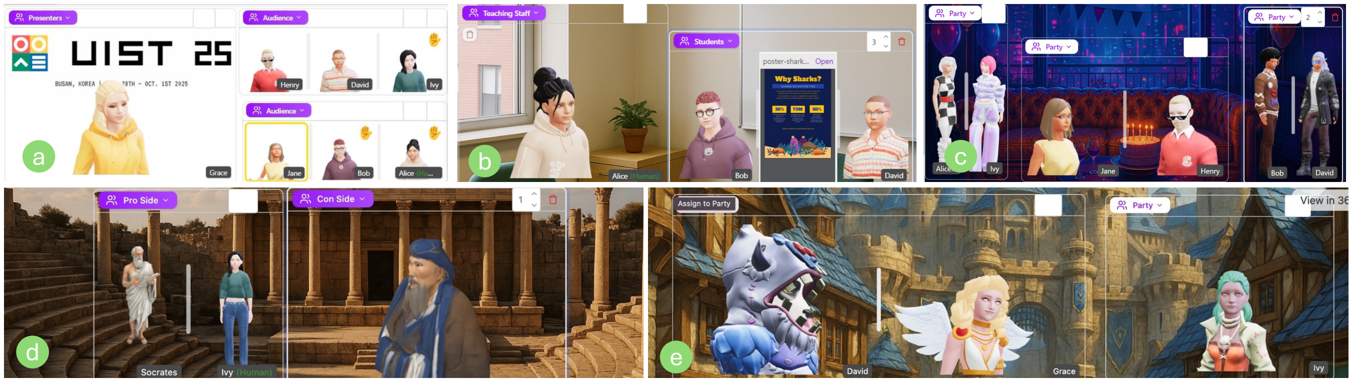


Figure 10: Applications: (a) Conference Q&A, (b) Office Hour, (c) Party Event (multiple groups), (d) Spiritual Historical Debate between Socrates and Confucius, (e) Inspiring Game Dialog Design.

7.1 Participants

We recruited 14 participants (7 female, and 7 male, aged 23-41, $\bar{x}$ = 29.7) across two primary categories: domain experts (technical versus non-tech), and regular users based on the application. Their demographics are attached in Appendix (Table 2). The nine domain experts included five game and avatar designers/engineers/researchers (E1–E5) from large tech companies, all with expertise in LLM-driven interactive character and narrative design. They evaluated the system’s design and interaction fidelity. We also included four non-technical domain experts: two educational practitioners (E6–E7) with training and instruction experience, and two social science researchers (E8–E9) focused on group dynamics.

The five regular users included four students (R1–R4), two of whom had backgrounds in human-AI research, and one working professional (R5) from a U.S.-based company, representing everyday users who frequently engage in multi-party workplace meetings.

7.2 Tasks and Procedure

Participants completed two distinct tasks designed to evaluate different aspects of the system. Before the two tasks, we instruct users how to configure scenes, parties, and snippets, how to export the conversations to verification with a tutorial (15min).

Walkthrough and Authoring of Group Dynamics Transition (20min). To explore how participants would envision using DialogLab for creating different group dynamics with different element settings and scenes, we asked them to compose a series of scenes for a social event: paper discussions (share a paper, party assigned), social events (no party assigned), and a formal Q&A session (hand raising). After authoring these scenes, participants loaded the conversation into the verification module to review group dynamics metrics and reflect aloud on their usefulness and interpretability.

Testing and Simulation of Human Agents (15min). First, we explore the system's testing human agent mode for a group discussion scenario that decide for a lunch. Participants were asked to create a simple scene with one snippet and load it in the conversation. They then create three conversations (each with 15 speaking turns) with three different control configurations: the human-controlled agents against two baselines: autonomous human agents and reactive human-controlled agents). The order of conditions was counterbalanced across participants.

7.3 Findings

Most participants agreed that DialogLab was easy to use (Median = 4, IQR = 1 , on a scale from 1 (Disagree) to 5 (Agree)) and intuitive (Median = 4, IQR = 1). During the conversation creation process, participants reported smooth experience in setting up scenes such as dragging avatars and assigning party roles (Median = 4.5, IQR = 1). After setting up scenes, they felt customizing conversation snippets such as turns and subtopics was also smooth (Median = 4.5, IQR = 1). Their sense of control and involvement were also high (Median = 4, IQR = 1). Participants mentioned feeling confident (Median = 4.5, IQR = 1) during the creation process and somehow agreed on the result was expected (Median = 4, IQR = 2).

7.3.1 *Flexible and Controllable Authoring Process.*

Participants generally praised DialogLab 's visual and modular authoring interface. The direct manipulation aspects, such as drag-and-drop scene setup and avatar-based role assignment, were highlighted by many (10/14) as intuitive and effective. R1 found that "*Dragging the avatars and forming different roles makes the setup process more fun and efficient.*" The integrated visualization was also valued; E1 appreciated seeing "*avatars and their settings all in one place,*" and E5 found the node-graph representation made the conversation flow "*visually logical.*" R2 echoed this, finding the process "*very engaging, where we can assign roles and control the snippet direction dynamically.*"

More than half of the participants (8/14) also appreciated the balance between configurability and auto-generation of prompts and agent behaviors through the Inspector and Node Graph views. E2 explained - "*...make the conversation align more with our expectation by changing the sub-description and regenerating the conversation prompt.*"

Furthermore, most participants highlighted the flexibility offered by party-based configuration of conversation attributes and moderation strategies that enabled the diversity and structure of conversation created. E6 described the setup as "very intuitive and systematic," suggesting value in supporting various moderation tactics such as "*asking opposing opinions or giving priority to quieter participants.*"

The real-time audio and avatar modes were highlighted for enhancing naturalness and aiding the identification of flow and tone issues, which is essential for most applications where real-time-ness is essential. E5 highlighted listening was preferable to text review, "*I could listen to 10 minutes of talk without any problem... But if I look at those texts for 10 min, I would get a little bit of a headache.*"

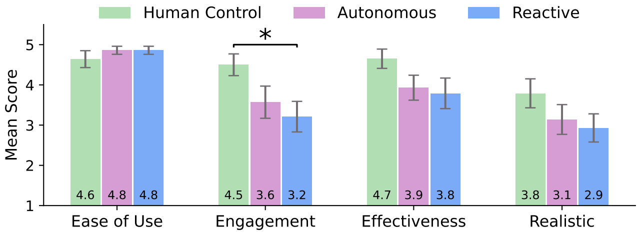


Figure 11: Ratings for human control, autonomous, and reactive agents in Task 2.

Overall, participants found DialogLab 's visual design and real-time interactivity not only intuitive but also critical to supporting creativity, engagement, and expressive control over complex multi-party scenarios. This is aligned with our **DG1** in creating a structured and flexible way of creating multi-party conversations.

7.3.2 *Iteration, testing, and verification.*

Fig. 11 shows the ease of use, engagement, effectiveness and realism of the three testing conditions. We analyzed data with Kruskal-Wallis test and found that the Human-Control condition has statistically higher engagement score ($p < .05$). Though other ratings were not statistically significant, Human-Control condition were perceived as more effective and realistic to simulate human behavior in the conversation.

When rating their preferences, most participants favored the *Human-Control* mode for its balance of agency, immersion, and flexibility in guiding the conversation. Many described it as the only mode that felt genuinely human-driven. As E3 explained, *"...To me, Human mode means I get to be the one interacting in place of the character."* Others appreciated the ability to influence the dialogue, noting it *"feels like involving [myself] into the conversation"* (E4) and *"gives me more options to respond...."* Several emphasized its support for co-authorship: E9 stated, *"You can freely edit and intervene in the conversation,"* and E6 added, *"I would like to have control and feel more immersed by taking part in the conversation myself instead of just reading the messages."* R2 also highlighted the design potential: *"This could help improve the design of real-world human-AI conversations since the designer can test their AI agent using this tool and gain an immersive experience..."*

Additionally, several participants (5/14) valued the ability to simulate human agents with different characteristics. R2 remarked, *"It is interesting to see/simulate how these human participants with different characteristic interact… This can also be a good metric for measuring the human-AI conversation system."* E3 noted, *"They can help people understand a (perhaps non-binary, but more on a spectrum) separation of AI and Human participants."* E7 added, *"It could help… because there are different modes of communication like human-human communication. So, distinguishing between these modes kind of reflect that."*

The *Reactive* mode was consistently rated lowest. It was described as *"passive"* and *"only respond[ing] when addressed."* One participant noted it offered *"no real involvement,"* and another commented, *"I like the ability to control the conversation a little more."*

E3 remarked, *"...I found the responses boring, so I didn't even attempt to follow along most of them."*

The *Autonomous* mode, though powered by the same mechanism as Human-Control, was seen as less engaging due to lack of control. Still, two participants preferred it for its generative efficiency. E5 stated, *"Autonomous mode helps us generate massive usable data,"* and R2 appreciated the effortless part, *"I enjoy seeing how it generate impromptu speech, which is more surprising... without my effort."*

Some participants suggested making **Human-Control** more expressive. E5 suggested, *"...For human-control, there are much more control required for real human interaction, like eye contact, subtle voice changes, which is hardly modeled by this approach."*

Overall, participants found the Human-Control mode to be the most effective and realistic for simulating real-world conversations—aligning closely with our design goal of supporting the testing more effectively with human agents (DG2).

7.3.3 Verification as a Diagnostic and Analysis Tool. Participants valued the **verification panel** as a diagnostic tool for efficiently exploring conversation dynamics, allowing quick review of turn-taking patterns without manually scanning full transcripts. In addition to the visualized metrics, the Q&A panel was seen as helpful for open-ended analysis. For example, E9 shared, *"I would like to try the word cloud,"* then typed the request into the input box and received a list of top frequent words from the conversation.

Social science researchers particularly saw value beyond testing, viewing it as useful not only for testing but also for data collection and analysis. Several participants suggested additional metrics for analyzing longer or more complex conversations. For instance, E8 proposed integrating network analysis: *"I'd like to see network analysis—for example, to check Alice and Bob's relationship, like who is closer with whom based on some metrics."*

Beyond visualization, participants also saw the verification stage as a way to evaluate system behavior across different LLM configurations. E5 explained, *"A very useful use case would be to test if they're consistent when I upgrade to a new backend."* He elaborated, *"If we want to let the agent understand what is 'concise', we used to use a small model and had to prompt really hard... When we upgraded to [a more advanced LLM], and if we use the original prompt, the agent became too concise—just 2 or 3 words."* In this context, DialogLab can help surface turn-taking metrics and behavioral differences, enabling designers to verify prompt effectiveness and agent response consistency across models.

7.3.4 In-depth Feedback on Application Scenarios. Participants envisioned DialogLab being useful for education, game design, and AI training.

Education and Instructional Use. Participants saw potential for structuring classroom discussions, facilitating engagement, and simulating challenging social scenarios for practice, *e.g.*, E1 shared, "*I would like to use a system like this to simulate debates about historical events so I could learn more as I listen.*"

Game Design and Interactive Storytelling. In creative domains, DialogLab was viewed as a promising tool for scripting NPC interactions and inspiring the dialog design. E3, drawing on game design experience, noted: "*This tool could generate naturally flowing NPC dialogue as an alternative to boring, repeating monologues most background characters perform.*" He also proposed blending AI with human players: "*Imagine a group of 8 characters with 6 human players and 2 AI NPCs that actually participate in conversations and express personalities—masking the fact that they aren't real players.*" Furthermore, E4 shared that this can be a effective tool for novice to be immersed in the character design, "*I've been in the field for 10 years, but for novice, this tool, especially the audio and avatar-based interaction can inspire them more for designing scripts and dialogs.*"

AI Testing, Prototyping, and Research. DialogLab was also seen as valuable for testing AI agents and studying human-AI interactions. For example, E4 imagined integration into real-world platforms: "*It could be used as a bot in Zoom by simulating various meeting scenarios before deployment.*" Researchers also appreciated the potential for qualitative testing and validation. E9 shared her social science research perspective that DialogLab can be used as experimental settings for "Qualitative data collection", and can be used to test some "*sensitive topics such as stereotypes and bias*", which can be otherwise hard to collect via survey.

7.3.5 Desire for Realism and Dynamic Flow. While realism is not the primary focus of DialogLab, participants generally felt that the system effectively captured key aspects of multi-party dynamics, such as natural turn-taking and shifting group roles. E6 noted - "*The system effectively created conversation scenarios based on context cues... I would expect it to simulate presentations, interviews, or difficult conversations where I could learn to anticipate emotional responses.*" Similarly, R2 observed, "*DialogLab has a great representation of real-world multi-person interactions,*" and envisioned scenarios involving group-to-classroom transitions, "*where the TA first breaks people into smaller groups, then has each present their ideas to the whole class.*" E9 also noted its potential for research use, saying, "*I think it is more effective than I expected... I would apply this to academic data collection such as focus groups and interviews.*"

However, a few participants (2/14) highlighted limits in realism. E3 pointed out that "*in the real world... turn-taking collapses easily, because people are non-robotic, talk over each other,*" and critiqued overly polished agent behavior: "*They're too enthusiastic—thanking and complimenting each other constantly... I don't think the actual implementation here is very realistic.*" This may be partially due to the avatar configurations, which were primarily set to be friendly, thus leading more positive tones of the conversation.

8 DISCUSSION

Our evaluation demonstrated that users could effectively create and manage complex conversational scenarios, leveraging both scripted and generative elements.

8.1 Reflection on the Workflows Supporting Multi-Party Group Dynamics

8.1.1 Support Diverse Group and Conversational Dynamics. DialogLab 's explicit modeling of group roles (parties) and configurable turn-taking mechanisms allowed users to simulate complex conversation structures **(DG1)**, different from prior work that enables authoring avatars for one-on-one conversations [21, 23, 76].

Participants appreciated the ability to represent diverse conversational dynamics, such as collaborative voice, hierarchical turn-taking, and role-based interruptions. These perspectives are essential for simulating interactions among multiple agents, a capability that is not emphasized in other multi-agent frameworks like AutoGen [102] that focuses more on task completion.

In the evolving landscape of conversational AI, DialogLab distinguishes itself by addressing the important challenges in designing and testing dynamic, multi-party, hybrid human-AI conversations.

8.1.2 Hybrid Control Should Be First-Class. The introduction of configurable human agents opens new opportunities for stress-testing conversational systems in multi-party contexts, especially in scenarios involving conflict, confusion, or emotional shifts. Participants preferred the Human-Control mode not just for agency, but because it enabled deeper engagement and a feeling of co-authorship. This indicates the need for conversation tools to support mixed-initiative interaction [37], where users can fluidly intervene, revise, or redirect conversation flow. While many tools [102] supports customizable agent behaviors, their emphasis is more on collaborative workflows rather than the nuanced control of interaction modes within a single conversation.

In summary, DialogLab fills a critical gap by providing tools specifically designed for the detailed authoring, simulation, and verification of multi-party hybrid human-AI conversations. Its focus on scenario-specific testing, complex group dynamics, hybrid interaction control, and dedicated verification aligns with the nuanced requirements of designers and researchers in this domain.

8.2 Design Implications

Our findings suggest several broader implications for designing tools that support hybrid human-AI conversation development:

8.2.1 Treat Conversation as a Design Material. DialogLab's snippet-based design, coupled with verification and playback, positions conversation not as static text but as a *dynamic artifact*—to be authored, explored, and iteratively refined [82, 85]. This resolves the challenges identified in the formative study where people use some random text-based tests or rely or leader's examples to test the conversation, which is not systematic. Tools should support both structured scripting and exploratory authoring, allowing designers to define flows while also discovering new emergent interactions. This aligns with broader goals in prototyping interactive systems: supporting what-if experimentation, alternative trajectories, and edge case exploration.

8.2.2 Incorporate Multimodal and Paralinguistic Cues. Some participants expressed a desire for richer human-control mechanisms that go beyond text, including eye contact, gesture, tone, or silence. This highlights a broader opportunity for tools to support multimodal conversation prototyping—either via avatar controls, speech input, or contextual timing adjustments [71]. Even lightweight modeling of nonverbal cues could help simulate more realistic interactions and test how AI agents interpret subtle signals.

8.2.3 Context-Aware Verification Tools. While the generic metrics on the verification panel helped participants quickly scan conversation dynamics, some also suggested additional analysis (*e.g.*, network diagrams or word clouds) to better interpret interpersonal flow, specifically in the social science experiment context. This suggests that verification tool to adapt to the specific scenario it applies, with different focus based on the conversation creator's need [1]. Building verification tools that balance clarity and depth—offering high-level summaries alongside exploratory tools—can better support both design iteration and research insight [85].

9 LIMITATIONS AND FUTURE WORK

Lack of Controlled Baseline Comparison. Given that DialogLab introduces a framework for authoring and simulating hybrid human-AI group conversations, our evaluation focused on exploring its design utility and potential through scenario-based tasks with users. While we compared DialogLab's capabilities to existing tools (*e.g.*, single-agent chat platforms, avatar-based simulation tools) in related work section, a controlled baseline comparison against traditional scripting or manual testing workflows was beyond the scope of this study. Future work could involve comparative studies that benchmark DialogLab against existing authoring pipelines to better understand the trade-offs between manual scripting, fully generative systems, and hybrid authoring approaches.

As conversations grow in the number of agents, roles, and snippet transitions, managing and visualizing scene logic can become cognitively demanding for users, especially those without prior experience in conversation authoring.

Limitations of LLM-Driven Simulation. LLMs may generate fabricated or inaccurate information (*i.e.*, hallucinations) [42], such as inventing plausible but false historical events for well-known figures. Second, LLMs may refuse to generate responses deemed sensitive or inappropriate [8]. This poses a challenge when simulating characters meant to behave provocatively or unethically, which can be important in scenarios involving conflict, bias, or power dynamics. Additionally, the prompt templates and tuning strategies in DialogLab were designed for the Gemini model used in our deployment. These prompts may not generalize well across other LLMs without adaptation. Furthermore, this work did not formally evaluate the output dialogue quality, as our primary contribution is the authoring framework itself, not the underlying LLM's generative quality, which evolves rapidly. Future studies may consider conducting rigorous, metric-based evaluations for further validating the findings, which are beyond our current scope.

Integrating Richer Multi-modal Embodied Behaviors, Audio, and Spatial Environment. The current version of DialogLab enables limited support for avatar embodiment and animation. Notably, it lacks mechanisms for pointing and referencing gestures, which are crucial for natural communication to enhance clarity [14, 46, 64].

Furthermore, DialogLab renders conversational agents in a 2D layout, which limits the spatial configuration and depth that can be achieved in simulations. The current 2D setups can also be used as an input prompt for guiding the generative video creation [54]. Enabling a 3D environment could enhance the realism and immersion of interactions, allowing for more complex avatar behaviors and emergent spatial formations. Furthermore, incorporate text-to-image or text-to-3D models may support more flexible environment and object creation experience [41, 74], which is out of the scope of this work.

Future versions of DialogLab could benefit from integrating photorealistic avatars [2, 72], richer nonverbal cues—such as eye contact [36], micro-expressions [67], and posture mirroring [35]—to increase simulation realism. Moreover, incorporating audio-to-audio models (*e.g.*, expressive voice cloning, emotional prosody control, or real-time speech-to-speech agents) would allow more fluid, conversational experiences without relying solely on TTS output.

Authoring Scalability and Simulation Controls. Our approach provides a foundational structure for managing large-scale conversations through the core concepts of Parties and Speaking/Turn-Taking Modes. For instance, a 100-person meeting can be modeled as a single "Audience" party with a "Moderated Hand-Raising" mode to ensure orderly turn-taking. The current Verification Dashboard supports multi-run comparative analysis, offering a natural foundation for more advanced, scalable testing. Future system should extend this to support programmatic batch testing, allowing researchers to systematically vary agent personas or conversational snippets to automatically generate and evaluate a large design space of conversational interactions. Integrating agents with mirrored world [18] and more comprehensive automated logging would further enhance the system's ability to robustly evaluate emergent conversational behaviors in real-world, open-ended contexts.

10 CONCLUSION

We present DialogLab, a prototyping toolkit for creation and iterative design of multi-party conversations between humans and AI agents. DialogLab allows users to configure various conversation attributes, including agent roles, interaction patterns, and turn-taking dynamics, to simulate realistic and dynamic dialogues. The toolkit supports both scripted and adaptive conversational elements, offering a hybrid approach that combines the structure of predetermined dialogues with the flexibility of spontaneous interactions. DialogLab aims to streamline the development process by providing an intuitive interface for conversation setup and real-time testing and validation, addressing key challenges such as the complexity of configuration, uncertainty of AI responses, and lack of systematic refinement methods. Our evaluation with 14 participants demonstrates its effectiveness in enhancing user control, reducing development time, and facilitating realistic multi-party conversation simulations across various domains. We hope that DialogLab's approach and contributions will help advance and inspire the emerging field of Human-AI group conversation dynamics.

ACKNOWLEDGMENTS

We thank the participants of our formative study and the experts from our final evaluation for their time and valuable contributions, which were instrumental in shaping this research. We would also like to thank our reviewers for their insightful feedback.

REFERENCES

[1] Riku Arakawa, Kiyosu Maeda, and Hiromu Yakura. 2025. ConverSearch: Supporting Experts in Human Behavior Analysis of Conversational Videos With a Multimodal Scene Search Tool. *ACM Trans. Interact. Intell. Syst* 15, 1, Article 6 (February 2025), 31 pages. https://doi.org/10.1145/3709012

[2] Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, Rohit Pandey, Ping Tan, Thabo Beeler, SeanRyan Francesco Fanello, and Yinda Zhang. 2023. Learning Personalized High Quality Volumetric Head Avatars From Monocular RGB Videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 16890–16900. https://doi.org/10.1109/CVPR52729.2023.01620

[3] Adrian Bangerter and Herbert H Clark. 2003. Navigating Joint Projects with Dialogue. *Cognitive Science* 27, 2 (2003), 195–225. https://doi.org/10.1016/S0364-0213(02)00118-0

[4] Amy L Baylor. 2002. Agent-Based Learning Environments As a Research Tool for Investigating Teaching and Learning. *Journal of Educational Computing Research* 26, 3 (2002), 227–248. https://doi.org/10.2190/PH2K-6P09-K8EC-KRDK

[5] Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 396–403.

[6] Timothy Bickmore and Justine Cassell. 2005. Social Dialongue With Embodied Conversational Agents. *Advances in Natural Multimodal Dialogue Systems* (2005), 23–54. https://doi.org/10.1007/1-4020-3933-6_2

[7] Ian Bogost. 2010. *Persuasive Games: The Expressive Power of Videogames*. mit Press.

[8] Ali Borji and Mehrdad Mohammadian. 2023. Battle of the Wordsmiths: Comparing ChatGPT, Gpt-4, Claude, and Bard. *GPT-4, Claude, and Bard (June 12, 2023)* (2023).

[9] Ricardo Cabello et al. 2010. Three.js: A JavaScript 3D Library. https://threejs.org/.

[10] Simon Chauvin, Guillaume Levieux, Jean-Yves Donnart, and Stphane Natkin. 2015. Making Sense of Emergent Narratives: An Architecture Supporting Player-Triggered Narrative Processes. In *2015 IEEE Conference on Computational Intelligence and Games (CIG)*. 91–98. https://doi.org/10.1109/CIG.2015.7317936

[11] Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024. Socialbench: Sociality Evaluation of Role-Playing Conversational Agents. In *Findings of the Association for Computational Linguistics ACL 2024*. 2108–2126. https://doi.org/10.48550/arXiv.2403.13679

[12] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. *ArXiv Preprint ArXiv:2404.18231* (2024). https://doi.org/10.48550/arXiv.2407.11484

[13] HH Clark. 1991. Grounding in Communication. *Perspectives on Socially Shared Cognition/APA* (1991). https://doi.org/10.48550/arXiv.2507.07284

[14] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

[15] Frederick G Conrad, Jessica S Broome, José R Benkí, Frauke Kreuter, Robert M Groves, David Vannette, and Colleen McClain. 2013. Interviewer Speech and the Success of Survey Invitations. *Journal of the Royal Statistical Society Series A: Statistics in Society* 176, 1 (2013), 191–210. https://doi.org/10.48550/arXiv.2502.20140

[16] React Three Fiber Contributors. 2024. React-Three-Fiber. https://github.com/pmndrs/react-three-fiber.

[17] Wen Dong, Bruno Lepri, Taemie Kim, Fabio Pianesi, and Alex Sandy Pentland. 2012. Modeling Conversational Dynamics and Performance in a Social Dilemma Task. In *2012 5th International Symposium on Communications, Control and Signal Processing*. 1–4. https://doi.org/10.1109/ISCCSP.2012.6217775

[18] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI, 685)*. ACM, 13. https://doi.org/10.1145/3290605.3300915

[19] Starkey Duncan. 1974. On the Structure of Speaker-Auditor Interaction During Speaking Turns. *Language in Society* 3, 2 (1974), 161–180. http://www.jstor.org/stable/4166761

[20] Yrjö Engeström. 2015. *Learning by Expanding*. Cambridge University Press.

[21] M. Fink. 2024. GPTAvatar Authoring Tool. Retrieved from https://avatar-research.com.

[22] M. C. Fink, S. A. Robinson, and B. Ertl. 2024. AI-Based Avatars Are Changing the Way We Learn and Teach: Benefits and Challenges. *Frontiers in Education* (2024). https://doi.org/10.3389/feduc.2024.1416307

[23] Maximilian C Fink, Lars Walter, Bettina Eska, and Bernhard Ertl. 2025. An Authoring Tool for Individual and Collaborative Learning Scenarios with AI-Based Avatars [avatar-research.com]. (2025). https://doi.org/10.35542/osf.io/ckd9e_v1

[24] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots*. Now Foundations and Trends. https://doi.org/10.48550/arXiv.1809.08267

[25] James Garvey. 1978. Characterization in Narrative. *Poetics* 7, 1 (1978), 63–78. https://doi.org/10.48550/arXiv.2507.03214

[26] Melanie C Green and Timothy C Brock. 2000. The Role of Transportation in the Persuasiveness of Public Narratives. *Journal of Personality and Social Psychology* 79, 5 (2000), 701.

[27] Jens Emil Sloth Grønbæk, Marcel Borowski, Eve Hoggan, Wendy E Mackay, Michel Beaudouin-Lafon, and Clemens Nylandsted Klokmose. 2023. Mirrorverse: Live Tailoring of Video Conferencing Interfaces. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14. https://doi.org/10.1145/3586183.3606767

[28] Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, Yao Hu, Hongwei Feng, and Yanghua Xiao. 2024. Agent Group Chat: An Interactive Group Chat Simulacra For Better Eliciting Collective Emergent Behavior. *CoRR* abs/2403.13433 (2024). https://doi.org/10.48550/arXiv.2403.13433

[29] J Richard Hackman. 1968. Effects of Task Characteristics on Group Products. *Journal of Experimental Social Psychology* 4, 2 (1968), 162–187. https://doi.org/10.48550/arXiv.2207.09655

[30] J Richard Hackman. 2002. *Leading Teams: Setting the Stage for Great Performances*. Harvard Business Press.

[31] J Richard Hackman and Nancy Katz. 2010. Group Behavior and Performance. *Handbook of Social Psychology* 2 (2010), 1208–1251. https://doi.org/10.48550/arXiv.2507.07767

[32] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects. *Authorea Preprints* 1 (2023), 1–26.

[33] Jessica Hammer. 2007. Agency and Authority in Role-Playing "texts". *A New Literacies Sampler* 29 (2007), 67–94.

[34] D Fox Harrell and Jichen Zhu. 2009. Agency Play: Dimensions of Agency for Interactive Narrative Design. In *AAAI Spring Symposium: Intelligent Narrative Technologies II*. 44–52.

[35] Zhenyi He, Ruofei Du, and KenH. Perlin. 2020. CollaboVR: A Reconfigurable Framework for Multi-user to Communicate in Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 542–554. https://doi.org/10.1109/ISMAR50242.2020.00082

[36] Zhenyi He, Keru Wang, BrandonY. Feng, Ruofei Du, and KenH. Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-aware 3D Photos. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 769—-782. https://doi.org/10.1145/3472749.3474785

[37] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 159–166. https://doi.org/10.48550/arXiv.2502.04029

[38] Erzhen Hu, Md Aashikur Rahman Azim, and Seongkook Heo. 2022. FluidMeet: Enabling Frictionless Transitions Between In-Group, Between-Group, and Private Conversations During Virtual Breakout Meetings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, Article 511, 17 pages. https://doi.org/10.1145/3491102.3517558

[39] Erzhen Hu, Jens Emil Sloth Grønbæk, Austin Houck, and Seongkook Heo. 2023. OpenMic: Utilizing Proxemic Metaphors for Conversational Floor Transitions in Multiparty Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 793, 17 pages. https://doi.org/10.1145/3544548.3581013

[40] Erzhen Hu, Jens Emil Sloth Grønbæk, Wen Ying, Ruofei Du, and Seongkook Heo. 2023. ThingShare: Ad-Hoc Digital Copies of Physical Objects for Sharing Things in Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 365, 22 pages. https://doi.org/10.1145/3544548.3581148

[41] Erzhen Hu, Mingyi Li, Andrew Hong, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2025. Thing2Reality: Enabling Spontaneous Creation of 3D Objects from 2D Content using Generative AI in XR Meetings. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (Busan, Republic of Korea). Association for Computing Machinery. https://doi.org/10.1145/3746059.3747621

[42] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55. https://doi.org/10.48550/arXiv.2311.05232

[43] Gail Jefferson. 1984. On stepwise transition from talk about a trouble to inappropriately next-positioned matters. *Structures of social action: Studies in conversation analysis* 191 (1984), 222.

[44] W Lewis Johnson and James C Lester. 2018. Pedagogical Agents: Back to the Future. *AI Magazine* 39, 2 (2018), 33–44.

[45] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic Detection of Discourse Structure for Speech Recognition and Understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 88–95.

[46] Adam Kendon. 1994. Do Gestures Communicate? a Review. *Research on Language and Social Interaction - RES LANG SOC INTERACT* 27 (07 1994), 175–200. https://doi.org/10.1207/s15327973rlsi2703_2

[47] Patrick Kenny, Arno Hartholt, Jonathan Gratch, William Swartout, David Traum, Stacy Marsella, and Diane Piepol. 2007. Building Interactive Virtual Humans for Training Environments. In *Proceedings of I/itsec*, Vol. 174. 911–916. https://doi.org/10.48550/arXiv.2410.23535

[48] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social cognition* 26, 2 (2008), 169–181.

[49] Benjamin Klieger, Charis Charitsis, Miroslav Suzara, Sierra Wang, Nick Haber, and John C Mitchell. 2024. ChatCollab: Exploring Collaboration Between Humans and AI Agents in Software Teams. *arXiv preprint arXiv:2412.01992* (2024).

[50] Dominique Knutsen, Gilles Col, and Ludovic Le Bigot. 2018. An Investigation of the Determinants of Dialogue Navigation in Joint Activities. *Applied Psycholinguistics* 39, 6 (2018), 1345–1371.

[51] Robert E Kraut, Robert S Fish, Robert W Root, Barbara L Chalfonte, et al. 1990. Informal Communication in Organizations: Form, Function, and Technology. In *Human Reactions to Technology: Claremont Symposium on Applied Social Psychology*, Vol. 145. 199. https://doi.org/10.48550/arXiv.2407.11939

[52] Michael Kriegel, Ruth Aylett, Joao Dias, and Ana Paiva. 2007. An Authoring Tool for an Emergent Narrative Storytelling System. In *AAAI Fall Symposium: Intelligent Narrative Technologies*. 55–62.

[53] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social Interaction with Agents and Avatars in Immersive Virtual Environments: A Survey. *Frontiers in Virtual Reality* 2 (2022), 786665. https://doi.org/10.3389/frvir.2021.786665

[54] David Ledo. 2025. Generative Rotoscoping: A First-Person Autobiographical Exploration on Generative Video-to-Video Practices. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. ACM, 931–948. https://doi.org/10.1145/3698061.3726926

[55] Geonsun Lee, Dae Yeol Lee, Guan-Ming Su, and Dinesh Manocha. 2024. "May I Speak?": Multi-Modal Attention Guidance in Social VR Group Conversations. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[56] Lennard A. Leighton, Gary E. Stollak, and Lucy R. Ferguson. 1971. Patterns of Communication in Normal and Clinic Families. *Journal of Consulting and Clinical Psychology* 36, 2 (1971), 252–256. https://doi.org/10.1037/h0030759

[57] Kurt Lewin. 1947. Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change. *Human Relations* 1, 1 (1947), 5–41.

[58] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008. https://doi.org/10.48550/arXiv.2303.17760

[59] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. *ArXiv Preprint ArXiv:2205.14217* (2022). https://doi.org/10.48550/arXiv.2205.14217

[60] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. ACM, Article 659, 17 pages. https://doi.org/10.1145/3613904.3642947

[61] Zhuoran Lu, Qian Zhou, and Yi Wang. 2025. WhatELSE: Shaping Narrative Spaces at Configurable Level of Abstraction for AI-Bridged Interactive Storytelling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, Article 333, 18 pages. https://doi.org/10.1145/3706598.3713363

[62] Guido Makransky, Philip Wismer, and Richard E Mayer. 2019. A Gender Matching Effect in Learning With Pedagogical Agents in an Immersive Virtual Reality Science Simulation. *Journal of Computer Assisted Learning* 35, 3 (2019), 349–358.

[63] Jennifer Marlow, Scott Carter, Nathaniel Good, and Jung-Wei Chen. 2016. Beyond Talking Heads: Multimedia Artifact Creation, Use, and Sharing in Distributed Meetings. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1703–1715.

[64] David Mcneill. 1994. Hand and Mind: What Gestures Reveal About Thought. *Bibliovault OAI Repository, the University of Chicago Press* 27 (06 1994). https://doi.org/10.2307/1576015

[65] Aditi Mishra, Frederik Brudy, Qian Zhou, George Fitzmaurice, and Fraser Anderson. 2025. WhatIF: Branched Narrative Fiction Visualization for Authoring Emergent Narratives Using Large Language Models. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. ACM, 590–605. https://doi.org/10.1145/3698061.3726933

[66] P. Mukherjee. 2023. Introducing Convai. Retrieved from https://convai.com/blog/introducing-convai.

[67] Catherine Oh Kruzic, David Kruzic, Fernanda Herrera, and Jeremy Bailenson. 2020. Facial Expressions Contribute More Than Body Movements to Conversational Outcomes in Avatar-Mediated Virtual Environments. *Scientific Reports* 10, 1 (2020), 20626. https://doi.org/10.1038/s41598-020-76539-8

[68] Gary M Olson and Judith S Olson. 2000. Distance Matters. *Human-Computer Interaction* 15, 2-3 (2000), 139–178. https://doi.org/10.48550/arXiv.2507.07482

[69] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. ACM, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[70] Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living Memories: AI-Generated Characters As Digital Mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 889–901.

[71] Anna Penzkofer, Philipp Müller, Felix Bühler, Sven Mayer, and Andreas Bulling. 2021. Conan: A Usable Tool for Multimodal Conversation Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 341–351.

[72] Xun Qian, Feitong Tan, Yinda Zhang, BrianMoreno Collins, Alex Olwal, David Kim, Karthik Ramani, and Ruofei Du. 2024. ChatDirector: Enhancing Video Conferencing with Space-Aware Scene Rendering and Speech-Driven Layout Transition. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 16. https://doi.org/10.1145/3613904.3642110

[73] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation With LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. ACM, Article 1051, 19 pages. https://doi.org/10.1145/3613904.3642105

[74] Shwetha Rajaram, Nels Numan, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D Wilson. 2024. BlendScape: Enabling End-User Customization of Video-Conferencing Environments Through Generative AI. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. ACM, Article 40, 19 pages. https://doi.org/10.1145/3654777.3676326

[75] Albert Rizzo and Thomas Talbot. 2016. Virtual Reality Standardized Patients for Clinical Training. *The Digital Patient: Advancing Healthcare, Research, and Education* (2016), 255–272. https://doi.org/10.48550/arXiv.2504.09955

[76] S. A. Robinson. 2023. GPTAvatar. Retrieved from https://github.com/SethRobinson/GPTAvatar.

[77] John Rudnik, Sharadhi Raghuraj, Mingyi Li, and Robin N. Brewer. 2024. Care-Journal: A Voice-Based Conversational Agent for Supporting Care Communications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. ACM, Article 526, 22 pages. https://doi.org/10.1145/3613904.3642163

[78] Marie-Laure Ryan. 2004. Will New Media Produce New Narratives? Marie-Laure Ryan. *Narrative Across Media: The Languages of Storytelling* 337 (2004).

[79] Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A Simple Systematic for the Organisation of Turn Taking in Conversation. *Language* 50 (12 1974), 696–735. https://doi.org/10.2307/412243

[80] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (1974), 696–735.

[81] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, Article 252, 13 pages. https://doi.org/10.1145/3411764.3445615

[82] R Keith Sawyer. 2021. The Iterative and Improvisational Nature of the Creative Process. *Journal of Creativity* 31 (2021), 100002. https://doi.org/10.1080/10749039.2021.1893337

[83] Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Lawrence Erlbaum Associates.

[84] Emanuel A Schegloff. 1987. Analyzing Single Episodes of Interaction: An Exercise in Conversation Analysis. *Social Psychology Quarterly* (1987), 101–114.

[85] Donald A Schön. 1992. Designing As Reflective Conversation With the Materials of a Design Situation. *Knowledge-Based Systems* 5, 1 (1992), 3–14. https://doi.org/10.48550/arXiv.2306.09716

[86] Abigail Sellen. 1995. Remote Conversations: The Effects of Mediating Talk With Technology. *Human-Computer Interaction* 10 (12 1995), 401–444. https://doi.org/10.1207/s15327051hci1004_2

[87] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role Play With Large Language Models. *Nature* 623, 7987 (2023), 493–498. https://doi.org/10.48550/arXiv.2507.07247

[88] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-Llm: A Trainable Agent for Role-Playing. *ArXiv Preprint ArXiv:2310.10158* (2023). https://doi.org/10.48550/arXiv.2409.11726

[89] Ryan Shea and Zhou Yu. 2023. Building Persona Consistent Dialogue Agents With Offline Reinforcement Learning. *ArXiv Preprint ArXiv:2310.10735* (2023). https://doi.org/10.48550/arXiv.2310.10735

[90] Susan G Straus and Joseph E McGrath. 1994. Does the Medium Matter? the Interaction of Task Type and Technology on Group Performance and Member Reactions. *Journal of Applied Psychology* 79, 1 (1994), 87.

[91] Lucille Alice Suchman. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge university press.

[92] William R Swartout, Jonathan Gratch, Randall W Hill Jr, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. Toward Virtual Humans. *AI Magazine* 27, 2 (2006), 96–96. https://doi.org/10.48550/arXiv.2507.02634

[93] Maria Tomprou, Young Ji Kim, Prerna Chikersal, Anita Williams Woolley, and Laura A Dabbish. 2021. Speaking Out of Turn: How Video Conferencing Reduces Vocal Synchrony and Collective Intelligence. *PLoS One* 16, 3 (2021), e0247655.

[94] Bruce W Tuckman. 1965. Developmental Sequence in Small Groups. *Psychological Bulletin* 63, 6 (1965), 384. https://doi.org/10.48550/arXiv.2112.13811

[95] Lina Varotsi. 2019. *Conceptualisation and Exposition: A Theory of Character Construction*. Routledge.

[96] L. Walter. 2024. GPTAvatar—Multiplayer. Retrieved from https://github.com/walterlars/Bachelorthesis-GPTAvatar.

[97] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023. Rolellm: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. *ArXiv Preprint ArXiv:2310.00746* (2023). https://doi.org/10.48550/arXiv.2310.00746

[98] Sheida White. 1989. Backchannels across cultures: A study of Americans and Japanese. *Language in Society* 18, 1 (1989), 59–76.

[99] Steve Whittaker. 2003. Things to Talk About When Talking About Things. *Human-Computer Interaction* 18, 1-2 (2003), 149–170. https://doi.org/10.48550/arXiv.2404.16839

[100] John M Wiemann and Mark L Knapp. 2017. Turn-Taking in Conversations. *Communication Theory* (2017), 226–245. https://doi.org/10.48550/arXiv.2507.07518

[101] Jason D Williams and Steve Young. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech & Language* 21, 2 (2007), 393–422.

[102] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling Next-Gen Llm Applications via Multi-Agent Conversation Framework. *ArXiv Preprint ArXiv:2308.08155* (2023). https://doi.org/10.48550/arXiv.2308.08155

[103] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J. Cai. 2022. PromptChainer: Chaining Large Language Model Prompts Through Visual Programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, 1–10. https://doi.org/10.1145/3491101.3519729

[104] Zhen Wu, Serkan Kumyol, Shing Yin Wong, Xiaozhu Hu, Xin Tong, and Tristan Braud. 2025. Orchid: A Creative Approach for Authoring LLM-Driven Interactive Narratives. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. ACM, 774–791. https://doi.org/10.1145/3698061.3726906

[105] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The Rise and Potential of Large Language Model Based Agents: A Survey. *Science China Information Sciences* 68, 2 (2025), 121101. https://doi.org/10.48550/arXiv.2503.21422

[106] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-Training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, Online, 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30

[107] Z. Zhao, Z. Yin, J. Sun, and P. Hui. 2024. Embodied AI-Guided Interactive Digital Teachers for Education. In *Proceedings of the SIGGRAPH Asia 2024 Educator's Forum*. 1–8. https://doi.org/10.1145/3680533.3697070

[108] Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, Online, 142–149. https://doi.org/10.18653/v1/2020.acl-demos.19

A APPENDIX

A.1 Demographics

Table 1 and Table 2 shows the Demographics of participant recruitment.

Participant ID	Age	Gender	Job Title	Area of Designing Agent-Involved Conversations	Types of Conversations Designed	AI Use in Human-Agent Conversations
P1	29	Female	PhD Student	Embodied or Voice-based Agent, Multi-modal interaction	Multi-party	3
P2	27	Male	PhD Student	Embodied or Voice-based Agent, Natural Language Generation, UX design	Multi-party	4
P3	31	Female	UX Engineer	Embodied or Voice-based Agent, Multi-modal interaction, UX design	One-on-one	4
P4	27	Female	Software Engineer	Dialogue system design (Turn-taking mechanisms involved), Natural Language Generation, Multi-modal interaction	One-on-one	5
P5	40	Male	ML Tech Lead / Software Engineer	Dialogue system design (Turn-taking mechanisms involved), Natural Language Generation, Multi-modal interaction	One-on-one	5
P6	31	Female	Applied Vision Scientist / Research Scientist	Embodied or Voice-based Agent, Multi-modal interaction	One-on-one	1
P7	29	Male	PhD Student	Dialogue system design (Turn-taking mechanisms involved), Natural Language Generation, Specific Contexts (e.g., education, health-care)	Multi-party	2

Table 1: Formative Study Participant Demographics (1 being not familiar, 5 being familiar)

Participant ID	Age	Gender	Job Title	Experience With LLM	Experience With Programming	Experience With Human-AI / Human-Agent Conversations
E1	37	Male	Technical Art Director, Avatar UX	3	4	3
E2	33	Male	Game SWE	3	4	3
E3	35	Male	Game Designer/Developer	3	5	3
E4	41	Female	Game Designer	5	4	4
E5	29	Male	SWE/ Researcher Working on Agents	4	4	4
E6	32	Female	Education	4	2	1
E7	29	Male	Education	3	4	2
E8	30	Female	Social Science	4	2	1
E9	28	Female	Social Science	4	3	1
R1	23	Female	Student	5	4	4
R2	23	Female	Student	4	4	2
R3	24	Male	Student	4	5	1
R4	24	Female	Student	5	5	4
R5	28	Male	Working Professional	1	1	1

Table 2: User Evaluation Participant Demographics (1 being not familiar, 5 being familiar)

A.2 Algorithm

A.2.1 Agent Management. Agents are managed by the *ConversationManager* class in "Generative Multi-Party Conversation" (Fig. 12), which orchestrates complex dialogue flows, party-based interactions, and dynamic interruption mechanisms. Each agent is instantiated with specific personality traits, custom attributes, and roles, either as AI-driven entities or human proxies. The system supports multiple turn-taking modes including round-robin, direct selection, and sophisticated party-based allocation strategies.

Given N agents, interruption rules $\mathbf{R}$ are defined as a set of probabilistic interruption relationships:

$$\mathbf{R} = \{(a_i, a_j, p_{ij}, v_{ij}) | i, j \in [1, N], i \neq j\}$$

where p_{ij} represents the interruption probability from agent a_i to agent a_j, and v_{ij} denotes the interruption vibe (e.g., *critical*, *supportive*). Additionally, the system incorporates specialized *derailer agents* $\mathcal{D} \subset \mathcal{A}$ that can spontaneously intervene based on conversation context and configurable thresholds.

Algorithm 1 Generative Multi-party Conversation

```
Procedure main():
    $A$, config, , speakerQueues ← initializeEnvironment()
    $H \leftarrow \emptyset, t \leftarrow 0$
    memory ← ConversationMemory()
    $s_0$ ← config.initiator
    $C$ ← generateStartingContext($s_0$, config.topic, config.conversationPrompt)
    $\rho_0$ ← generateStartingMessage($s_0$, $C$)
    streamToClient($\rho_0$)
    $H \leftarrow H \cup \{(s_0, \text{“All”}, \rho_0)\}$
    $t \leftarrow 1$
    while $t <$ config.maxTurns do
        if isWaitingForApproval ∨ conversationPaused then break
        // Checks for moderated party transitions.
        if config.partyMode ∧ config.partyTurnMode = “moderated” then
            handleModeratedTransition(lastSpeaker)
        end if
        // Checks for derailer interventions.
        if ¬impromptuPhaseActive ∧ derailingEnabled then
            derailRequest ← checkForDerailInterventions(lastSpeaker, lastMessage)
            if derailRequest ≠ null then
                if autoApproveImpromptu then
                    startImpromptuPhase(derailRequest.derailerAgent, derailRequest.mode)
                    streamToClient(derailRequest.message)
                    lastSpeaker ← derailRequest.sender
                    continue
                else
                    pauseForApproval(derailRequest)
                    return
                end if
            end if
        end if
        $s_{t+1}$ ← selectNextSpeaker(lastSpeaker, participants)
        $r_{t+1}$ ← determineRecipient($s_{t+1}$, config)
        $C$ ← generateContext($s_{t+1}$, $H$, config.topic, lastSpeaker)
        $\rho_t$ ← generateReplyMessage($s_{t+1}$, $C$, $r_{t+1}$)
        if $\rho_t$.requiresHumanInput then
            onHumanInputRequired($s_{t+1}$)
            return
        end if
        message ← createMessage($s_{t+1}$, $\rho_t$, $r_{t+1}$)
        streamToClient(message)
        backchannels ← processBackchannels(participants, $s_{t+1}$, message)
        for each backchannel ∈ backchannels do
            streamToClient(backchannel)
        end for
        $H \leftarrow H \cup \{(s_{t+1}, r_{t+1}, \rho_t)\}$
        memory.addMessage(message)
        lastSpeaker ← $s_{t+1}$
        $t \leftarrow t + 1$
    end while
    onConversationComplete()
```

Key Functions

```
Function generateContext($s_t$, $H$, $\tau$, lastSpeaker):
    $C$ ← buildBaseContext($s_t$, $H$, $\tau$, memory)
    if partyMode then
        partyContext←getPartySpecificContext($s_t$, lastSpeaker)
        $C \leftarrow C \cup$ partyContext
    end if
    return $C$

Function selectNextSpeaker($s_t$, participants):
    if partyMode then
        return selectPartyBasedSpeaker($s_t$,participants)
    else
        return getNextRoundRobinSpeaker(participants, $s_t$)
    end if

Function selectNextParty(currentParty):
    if partyTurnMode = “free” then
        otherParties ← filter($p \neq$ currentParty)
        return otherParties[random(0, |otherParties|)]
    else if partyTurnMode = “round-robin” then
        currentIndex ← parties.indexOf(currentParty)
        return parties[(currentIndex + 1) mod |parties|]
    else if partyTurnMode = "moderated" then
        if currentParty = moderatorParty then
            return approvedSpeakers[0].party
        else
            return moderatorParty
        end if
    end if

Function generateReplyMessage($s_t$, $C$, $r_t$):
    if agents[$s_t$].isHumanProxy then
        return {requiresHumanInput : true, speaker : $s_t$}
    end if
    prompt ← buildAgentPrompt($s_t$, $C$, $r_t$)
    response ← llmProvider.generate(prompt)
    return processInterruption(response)

Function checkForDerailInterventions($s_t$, lastMessage):
    derailerAgents ← agents.filter(∀$a$ s.t. $a$.isDerailer ∧ $a$.name ≠ $s_t$)
    for each derailer ∈ derailerAgents do
        shouldIntervene ← random() < derailThreshold
        if shouldIntervene then
            derailResponse ← generateDerailResponse(lastMessage)
            if isDerailing then
                turnCount ← random(minTurns, maxTurns)
                return {sender : name, message : derailResponse, turnCount}
            end if
        end if
    end for
    return null
```

Figure 12: Generative Multi-party Conversation with Tailorable Conversation Flow

AI-driven agents leverage large language models with rich contextual prompts that integrate conversation history, party affiliations, dynamic attributes, and interaction patterns to generate contextually appropriate responses.

A.2.2 Interaction Handling Mechanism.

Conversation Flow. The conversation flow is managed by the “Generative Multi-Party Conversation” algorithm (Fig. 12), which handles multi-modal dialogue structures including standard turn-taking, party-based discussions, and impromptu intervention phases. The conversation state S_t at turn t encompasses not only the current speaker and message but also party configurations, speaker queues, and intervention states:

$$S_{t+1} = F(S_t, a_i, m_t, \mathcal{P}_t, Q_t, \mathcal{I}_t)$$

where a_i is the responding agent, m_t is the message at turn t, $\mathcal{P}_t$ represents the current party state, Q_t denotes speaker queues, and $\mathcal{I}_t$ indicates any active intervention phases.

Interaction Patterns and Context Generation. The system defines interaction patterns $\mathcal{P} \in \{\text{disagree, agree, neutral}\}$ for conversation-level configuration, with additional patterns available for individual agent customization. Context generation C integrates multiple components including interaction patterns, conversation history, party dynamics, and thematic analysis:

$$
\begin{aligned}
C_{\text{base}} &= \text{“Topic: ”} + \tau + \text{“ Recent discussion: ”} + \mathcal{H}_{\text{recent}} \\
C_{\text{pattern}} &= \begin{cases} \text{“Try to respectfully disagree with or challenge the last statement.”} \\ \hfill , \text{if } \mathcal{P} = \text{‘disagree’} \\ \text{“Agree with the last statement and expand on it.”} \\ \hfill , \text{if } \mathcal{P} = \text{‘agree’} \\ \text{“Contribute to the conversation with your own perspective.”} \\ \hfill , \text{if } \mathcal{P} = \text{‘neutral’} \end{cases} \\
C_{\text{party}} &= \text{getPartySpecificContext}(\text{agent}_i, \text{lastSpeaker}) \\
C &= C_{\text{base}} \oplus C_{\text{pattern}} \oplus C_{\text{party}} \oplus \mathcal{A}_i
\end{aligned}
$$

where τ represents the conversation topic, $\mathcal{H}_{\text{recent}}$ denotes recent conversation history, $\mathcal{A}_i$ represents agent i's custom attributes, and $\oplus$ indicates context concatenation.

Party-Based Interaction Management. The system supports sophisticated party-based conversations with three operational modes: *free* (random inter-party selection), *round-robin* (sequential party transitions), and *moderated* (approval-based speaking with hand-raising mechanisms). In moderated mode, agents must be approved from a speaker queue Q_{raised} before participating, enabling structured debate scenarios.

Interruption and Intervention Mechanisms. Beyond standard interruption rules, the system implements a dynamic intervention system through derailer agents that can initiate *impromptu phases* based on conversation context. These interventions support multiple derailment modes including *drift*, *extend*, *question*, and *emotional*, allowing for natural conversation flow disruptions and topic shifts that enhance conversational realism.

A.3 Prompts for the LLM

A.3.1 *Persona Definition.*

```
You are Ivy, a friendly traveling warrior
born in the wild forest.
```

A.3.2 *Conversation Context.*

```
Now, you are in a conversation where you
should follow this context: Alice, Bob, Ivy,
David, and Grace are gathered around a worn
table in the back room of the "Rusty Mug"
bar, the remnants of a poker game scattered
before them, as they soberly discuss the
implications of their recent losses and the
possibility of "taking the money away" from
the house.
The last person who spoke was Bob.
Recent discussion summary: Alice is suggesting
some kind of retaliation or "evening the
score" regarding an unspecified incident that
happened at a house last night. Bob is cautiously
receptive to the idea but urging careful
consideration before acting.
```

A.3.3 *Primary Task.*

```
Contribute to the conversation with your own
perspective.
Reply to Alice's message.
```

A.3.4 *Group Affiliation/Role.*

```
You are part of the "Travel Warrior" group.
Party role: Warrior should take trying to
take bounties from the NPCs
```

A.3.5 *Reference: Last Message.*

```
Last message: Evening the score, huh? I'm
listening, Alice. But let's think this through
carefully before we jump into anything rash.
```

A.3.6 *Response Constraints.*

```
Respond briefly (1-2 sentences), building on
previous points without repeating them.
Keep your response conversational and natural.
```

A.3.7 *Interaction Flow Information.*

```
After you speak, David will respond.
```

A.3.8 *Moderator Role and Task (special role).*

```
You are the moderator for this discussion.
The previous speaker was Alice. She just
finished speaking and may have asked a question
or made a point requiring acknowledgment.
Several members have raised their hands.
Your task is to: 1. Acknowledge Grace's contribution
thoughtfully. If she asked a question, briefly
address or incorporate it into your transition.
Do not just quote her. 2. Choose one member
who raised their hand and invite them to
speak next, addressing them clearly by name.
3. Combine your acknowledgment and invitation
into a single, smooth, and natural-sounding
message to keep the conversation flowing.
Constraints: - Show you understood people's
point in your reply. - Do not mention any
party names (like "The Noble House"). - Do
not make assumptions or predictions about
what the next speaker will say.
```

A.3.9 Example LLM Generation (Ivy's Response). These examples show how Ivy might respond to Bob's message ("*Evening the score, huh? I'm listening, Alice. But let's think this through carefully before we jump into anything rash.*"), categorized by interaction type:

Reactive (Follows rules strictly)

```
Bob's right, charging in blind is foolish.
But if there's a debt owed, the forest teaches
you to collect.
```

Rationale: Directly addresses Bob's caution, aligns with Ivy's background, and fulfills the prompt precisely.

Autonomous (Human interfered + auto-approval)

```
Thinking first is wise, Bob, saves arrows
for the real targets. What exactly are we
walking into here? Knowing the terrain is
half the hunt.
```

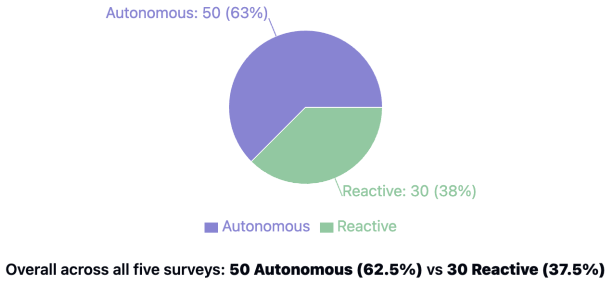


Figure 13: Overall preference distribution between Autonomous and Reactive modes across all surveys.

Rationale: Agrees with caution but autonomously adds a practical, information-gathering step consistent with the persona, feeling like a natural extension.

Human-Controlled (Topic shifting, etc.)

```
Aye, thinking's good, Bob, less blood that
way. But this stings, doesn't it? Feels like
more than just bad luck at cards - are we
sure we weren't cheated?
```

Rationale: Acknowledges Bob but introduces emotion ("stings") and a probing question ("weren't cheated?"), potentially shifting the conversation's focus beyond the direct prompt.

B SMALL-SCALE SIMULATION STUDY

Although not part of the system interface or primary evaluation, we conducted a small-scale study comparing conversations authored with *proactive (autonomous) human agents* and *reactive scripted agents*. As noted in the paper, Autonomous agents behave like Human-Control agents (will create cut-ins during the conversation), with the key difference being the absence of a human-in-the-loop. For five application scenarios that mentioned in the paper (*World of Warcraft, Design Review, Ice Breaker, Philosophical Debate, and Presentation*), we generated five iterations per scenario in each mode, resulting in a total of 50 conversations (5 scenarios × 5 iterations × 2 modes). The model we use is the Gemini-2.0-flash. Ten independent raters, blind to the experimental conditions, consistently rated the conversations with human agents as more natural and engaging than those with reactive agents. Here we report the results of the small simulation study. Each one of the participants rate 2 matching snippets of reactive and autonomous side by side (order randomized), followed by 3 single sample snippets, including all five categories of application topics.

Overall, participants preferred Autonomous mode over Reactive ones, with 62.5% (50 vs. 30 responses) favoring the Autonomous mode (see Fig. 13).

Among conversation attributes, perceived *Naturalness* showed the greatest difference favoring the Autonomous mode (mean difference of 0.40), while *Transitions* received the highest average ratings for both Autonomous and Reactive conversations (Fig. 14).

There was no significant difference perceived per topic. The largest average rating advantage for the Autonomous mode over reactive mode was observed in the *Ice Breaker* topic, with a mean difference of 0.35. Ratings for the *World of Warcraft* topic were the most consistent across surveys, while the *Design Review* topic received notably high ratings.

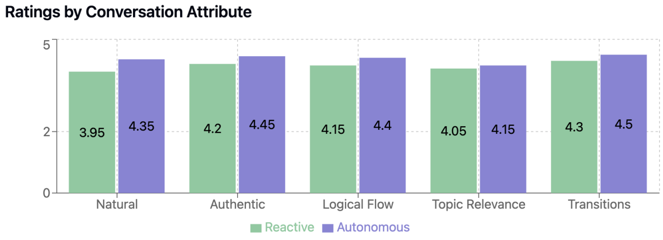


Figure 14: Average ratings comparison between Reactive and Autonomous modes across different conversation attributes.

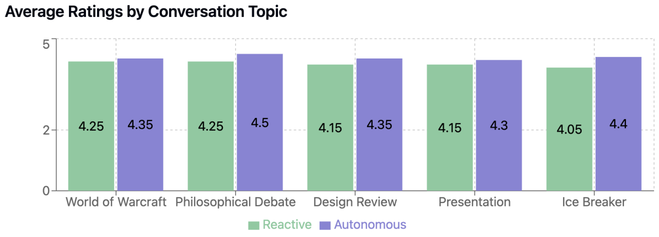


Figure 15: Average ratings comparison between Reactive and Autonomous modes across different conversation topics.

Based on open-ended responses comparing conversations, participants frequently preferred the generated conversations from autonomous mode, describing them as more *natural*, *human-like*, and marked by *subtlety*, *humor*, and *messiness*. In contrast, conversations generated from reactive mode were seen as more *structured*, *formal*, and *constructed for clarity*. Participants consistently attributed greater *creativity* and *dynamism* to the Autonomous displays, which indicates a preference for more organic and expressive interaction styles.